Bilateral Adaptation for Human-Object Interaction Detection with Occlusion-Robustness

Guangzhi Wang[1] Yangyang Guo[2] Ziwei Xu[2] Mohan Kankanhalli[2]
[1]Institute of Data Science, National University of Singapore
[2]School of Computing, National University of Singapore
guangzhi.wang@u.nus.edu, guoyang.eric@gmail.com, {ziwei-xu,mohan}@comp.nus.edu.sg

Abstract

Human-Object Interaction (HOI) Detection constitutes an important aspect of human-centric scene understanding, which requires precise object detection and interaction recognition. Despite increasing advancement in detection, recognizing subtle and intricate interactions remains challenging. Recent methods have endeavored to leverage the rich semantic representation from pre-trained CLIP, yet fail to efficiently capture finer-grained spatial features that are highly informative for interaction discrimination. In this work, instead of solely using representations from CLIP, we fill the gap by proposing a spatial adapter that efficiently utilizes the multi-scale spatial information in the pre-trained detector. This leads to a bilateral adaptation that mutually produces complementary features. To further improve interaction recognition under occlusion, which is common in crowded scenarios, we propose an Occluded Part Extrapolation module that guides the model to recover the spatial details from manually occluded feature maps. Moreover, we design a Conditional Contextual Mining module that further mines informative contextual clues from the spatial features via a tailored cross-attention mechanism. Extensive experiments on V-COCO and HICO-DET benchmarks demonstrate that our method significantly outperforms prior art on both standard and zero-shot settings, resulting in new state-of-the-art performance. Additional ablation studies further validate the effectiveness of each component in our method.

1. Introduction

The rapid progress of modern detection systems [5, 11, 46] has spawned a growing interest in the research of Human-Object Interaction (HOI) detection. There are generally two dispensable objectives in this emerging task: (1) detecting humans and objects in a given image and (2) recognizing human-object interactions, which are articulated as verb-object phrases. HOI Detection is important to human-centric scene understanding and underpins a variety of high-level tasks such as image captioning [53] and visual question answering [1].

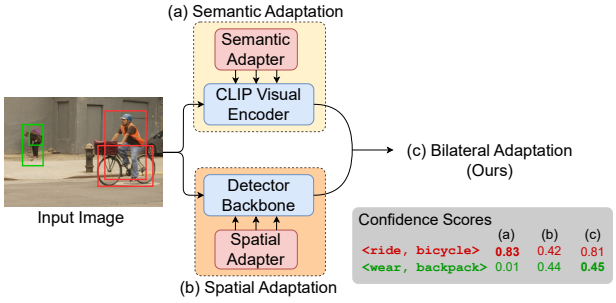

Figure 1. (a) Representations from CLIP focus on high-level semantics but ignore fine-grained details, resulting in a high score (0.83) on salient interaction ⟨ride, bicycle⟩, but a low score (0.01) on small-scale interaction ⟨wear, backpack⟩. (c) Our bilateral adaptation complements the semantic branch with spatial adaptation (b), resulting in comprehensive interaction understanding (high scores on both interactions).

As indicated by the two objectives, existing methods improve HOI detection from either better object detection or improved interaction recognition. Methods in the former group introduce various detector architectures [5, 11, 46] with proper adaptation for precise human-object detection [27, 33, 34, 48, 50, 63, 68, 69]. Pertaining to the methods in the latter group, early work designs graph-based architectures [14, 44, 51, 64] to enable information propagation between humans and objects, subsequent methods with the transformer-based architectures [27, 34, 48, 56, 63] utilize the attention mechanism [52] for better interaction understanding. To improve the representation discriminability, various human cues extracted from external knowledge sources have been utilized, such as pose [13, 17, 42, 54, 59], intentions [61], 3D representations [32], and motions [38]. Until recently, the CLIP [45] model has delivered im-

pressive results with its rich semantic representations [24]. As a result, advanced methods have leveraged its visual and textual representations to enhance interaction recognition [29, 34, 41].

Albeit the strong performance brought up by CLIP, we observe that it falls short of exploiting fine-grained spatial information for interaction recognition. For example, in Figure 1, CLIP in the semantic branch focuses on the salient interaction, *e.g.*, ⟨ride, bicycle⟩, with a high confidence score but neglects the fine-grained details of insignificant areas. In fact, CLIP is pre-trained with low-resolution images with coarse textual descriptions, making it powerful at learning high-level semantics but less effective at understanding fine-grained details [30, 62]. As a result, the detailed spatial information that is important for recognizing interactions can be easily overlooked, leading to suboptimal performance. To overcome this, we propose to incorporate finer-grained spatial representations readily available from pre-trained detectors, resulting in a bilateral adaptation of prior knowledge. Specifically, we introduce a set of adapter modules to the frozen object detector, so as to efficiently adapt its innate spatial knowledge of multiple granularity for interaction recognition. As such, the spatial representation is efficiently adapted to complement CLIP's semantic representation, offering a comprehensive interaction understanding.

Moreover, in real-world scenarios, humans and objects are often occluded by each other, especially when they are interacting. This occlusion phenomenon is even more severe when there are multiple humans and objects in the image. Such occlusion severely hinders the understanding of human and object details, resulting in challenges for precise interaction recognition. In light of this, we propose an Occluded Part Extrapolation (OPE) strategy to improve the model's interaction recognition capability under occlusions. In particular, we deliberately occlude the feature map of an instance, which is then reconstructed based on its context. Our OPE learns to recover fine-grained details from occlusions, resulting in occlusion-robust representations for interaction recognition. To further utilize the spatial information, we design a Conditional Contextual Mining (CCM) module to mine the most informative clues via a delicate cross-attention mechanism.

We conducted extensive experiments with our proposed method on two benchmark datasets, namely, HICO-DET [6] and V-COCO [16]. The experimental results demonstrate that our method outperforms previous approaches by a large margin in both standard and zero-shot settings, resulting in new state-of-the-art performance. We also perform detailed ablation studies to justify the effectiveness of each component in our method. Furthermore, we provide in-depth analyses of the proposed method, which demonstrates our method indeed improves the performance under occluded scenarios.

2. Related Work

2.1. One-stage HOI Detection

One-stage methods perform both human-object detection and interaction recognition in an end-to-end manner. In addition to detecting humans and objects, early work detects interaction points [33, 57, 68] or human-object union [26] regions as interaction clues. With the recent success of Transformer [52] in the computer vision community, Transformer Detector (DETR) [5]-based architectures have gained increasing popularity. To empower DETR with the ability of pairwise detection and interaction recognition, some methods adopt extra classification heads [48, 70], encoder [69] or decoder [8, 34, 63] for effective interaction classification. Moreover, instead of learning the queries from scratch, semantically meaningful queries, which can be extracted from image-level [23] or instance-level [10, 34] information, are utilized to speed up convergence.

2.2. Two-stage HOI Detection

Two-stage methods first perform object detection with an off-the-shelf detector and recognize the interactions of each human-object pair in the second stage. Early work [6, 47] utilize Faster-RCNN [46] for object detection, followed by ROIAlign [18] to extract human-object appearance features. Recent two-stage methods employ DETR [5] as the detector due to its efficiency [65] and superior performance.

Given the detected instances, methods based on graphs [14, 44, 51, 55, 64] and attention mechanism [56, 65] propagate information between humans and objects, enabling contextual understanding for interaction recognition. Some approaches also propose to reject non-interactive human-object pairs before interaction recognition to improve the performance [31, 37, 59, 67]. Moreover, Some HOI detection methods take advantage of external knowledge for recognizing interactions. The knowledge typically includes human pose and body parts [13, 17, 54, 59], human intentions [61] and motions [38]. Additionally, recent literature has resorted to the pre-trained CLIP [45], by utilizing its textual [34, 41] and visual representations [29, 42]. Despite its promising results, we argue that these methods show limitations in effectively utilizing fine-grained spatial features. In this work, we complement the semantic branch with spatial information, enabling improved interaction recognition with a better understanding of fine-grained details.

2.3. Mask and Modeling in Computer Vision

Mask modeling has been extensively studied in the natural language processing domain [9, 25]. Recent efforts have adapted this idea for self-supervised computer vision tasks

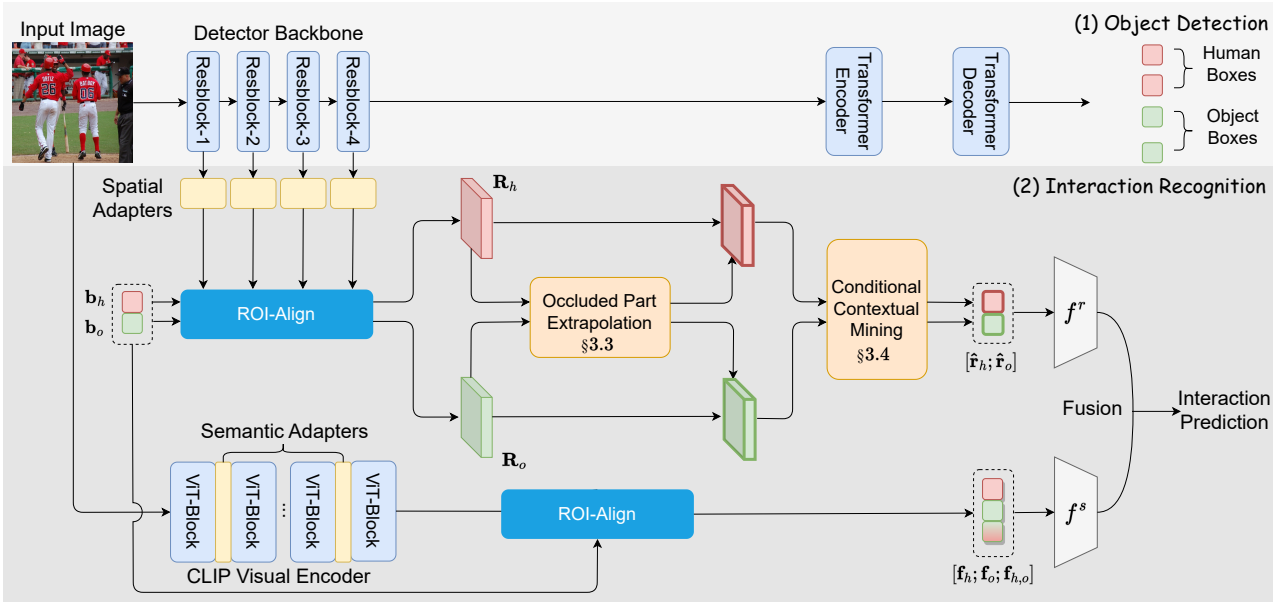

Figure 2. Overview of the BCOM model. (1) We perform object detection in the first stage. (2) In the second stage, we recognize interactions for each human-object pair. Besides utilizing the CLIP visual encoder for the semantic branch, we also leverage ROI-Align [18] to crop multi-scale spatial features, *i.e.*, $\mathbf{R}_h$ and $\mathbf{R}_o$ from the adapted detector backbone (Sec. 3.2). They are first refined with a learned Occluded Part Extrapolation (OPE) module (Sec. 3.3) and then fed into the Conditional Contextual Mining (CCM) module to extract informative contexts (Sec. 3.4). The interaction recognition results from f^r and f^s are fused to obtain the final results.

and witnessed pervasive success. Specifically, during pre-training, some image patches are deliberately masked and a neural network is then trained to reconstruct the original image. This mask-and-reconstruction paradigm has led to the success of a series of visual pre-training methods such as BEiT [3], BEiT-v2 [43], and MAE [19]. Based on this, following studies have contributed to improving this paradigm via different pre-training objectives [58, 60], masking strategies [7], and data modalities [2, 12, 40, 49]. In this work, we adopt a similar philosophy by manually occluding human/object features and reconstructing them during training. This approach enables the model to better understand human/object representations with partial occlusion, leading to occlusion-robust HOI detection.

3. Methodology

3.1. Overview

HOI detection involves detecting and predicting a set of ⟨human, verb, object⟩ triplets in an image, where interactions are defined as verb-object phrases. In this work, we propose a two-stage framework named **B**ilateral adaptation Network with occlusion-aware **CO**ntextual **M**ining (**BCOM**) for HOI detection. The framework is illustrated in Figure 2.

In the first stage, we use DETR [5] to detect all instances (*i.e.*, humans and objects) in the input image. We filter out the instances with low confidence and keep the number of detected instances within a certain range, resulting in the detection results $\{\mathbf{b}_i, c_i, s_i\}_{i=1}^{n}$, where $\mathbf{b}_i \in \mathbb{R}^4$ is the bounding box, c_i and s_i are scalars representing the class and confidence score of the detected instance, respectively.

In the second stage, we perform interaction recognition for each human-object pair. Based on the detection results, we permute over the detected instances to obtain a set of human-object pairs $\{\langle h, o\rangle \mid h, o \in \{1..n\} \wedge c_h = \text{human} \wedge h \neq o\}$[1]. To extract discriminative representation for each human-object pair, we propose to efficiently utilize the transferrable knowledge from CLIP visual encoder and detector backbone using a set of semantic adapters and spatial adapters respectively (§ 3.2). For the purpose of interaction understanding in occluded scenarios, we propose an Occluded Part Extrapolation (OPE) module, which is incorporated to teach our model to recognize interactions under occlusions (§ 3.3). To fully utilize the spatial representations, we design a Conditional Contextual Mining (CCM) module to mine the most informative contextual clues conditioned on the involved human and object (§ 3.4).

3.2. Bilateral Representation Adaptation

Accurate interaction recognition requires a discriminative visual representation that captures informative visual clues

[1]We allow "human" to be a type of "object" to enable the recognition of human-human interactions.

from multiple aspects. Early work mainly adopted the *spatial* representations from detectors [56, 65, 66], while recent work [29, 34, 41] has extensively employed the *semantic* representations from CLIP [45]. In this work, we find that they are both crucial for comprehensive interaction understanding and are in fact complementary to each other. In particular, the limitation of CLIP lies in its utilization of low-resolution images and coarse-grained textual descriptions during pre-training. This makes CLIP effective at capturing high-level semantics but struggles in learning finer-grained interaction clues [9, 30]. A direct solution is to fine-tune CLIP with high-resolution images [42], but this inevitably introduces excessive computational overhead. As such, we propose to adapt the spatial representations from the detector to complement the semantic representation from CLIP [45] for interaction recognition, resulting in an efficient bilateral adaptation structure.

Semantic Representation Adaptation. We adapt the rich semantic knowledge from the CLIP visual encoder to represent each human-object pair. To efficiently utilize the knowledge in the pretraining stage, we follow [29] to insert a set of learnable semantic adapter layers in each block of the visual encoder while keeping other parameters frozen. Then, we use ROI-Align [18] followed by mean pooling to obtain the human representation $\mathbf{f}_h$, object representation $\mathbf{f}_o$ as well as their union region representation $\mathbf{f}_{h,o}$. These features are concatenated as the semantic interaction representation for the human-object pair:

$$\mathbf{F}_{h,o} = [\mathbf{f}_h; \mathbf{f}_o; \mathbf{f}_{h,o}]. \tag{1}$$

Spatial Representation Adaptation. To complement the spatial details that CLIP struggles with, we complement the semantic representation with rich spatial representations. Instead of learning the spatial clues from scratch, we propose to utilize the knowledge from the pre-trained detector, which contains informative representations about instance locations and identity [66]. In particular, for each block in the detector backbone, we propose a set of learnable spatial adapter modules that efficiently transfer the spatial knowledge from the detector, obtaining multi-level spatial representations. Different from semantic adapters which are sequentially inserted into the visual encoder, the spatial adapters are appended to each residual block in parallel. This design does not interfere with the detector inference process, making it more efficient with a multi-scale structure. We then utilize multi-scale ROI-Align to crop the human region $\mathbf{R}_h$ and object region $\mathbf{R}_o$.

3.3. Occluded Part Extrapolation

Accurately recognizing human-object interactions requires a detailed understanding of human-objects. However, real-world scenarios can be crowded, resulting in occlusions between humans and objects. To address this problem, we propose an Occluded Part Extrapolation (OPE) module to facilitate interaction understanding in the presence of occlusions.

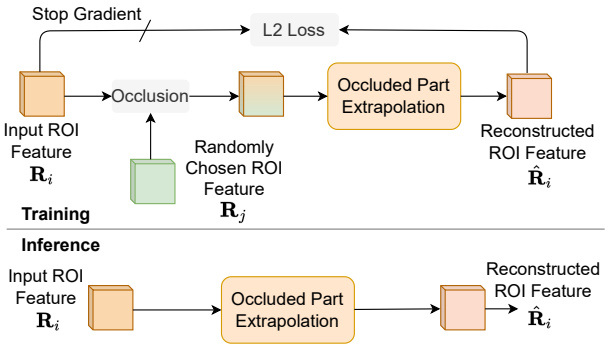

Figure 3. The mechanism of the Occluded Part Extrapolation (OPE) module. During training, the ROI feature $\mathbf{R}_i$ is occluded by a randomly chosen ROI feature $\mathbf{R}_j$. The OPE module is trained to reconstruct $\mathbf{R}_i$ from the occluded feature. During inference, we employ OPE to extrapolate the full details of $\mathbf{R}_i$.

The pipeline of our proposed OPE is depicted in Figure 3. The main objective of OPE is to train a module capable of extrapolating the original details of an instance from its partially occluded counterpart. For example, given a human spatial feature $\mathbf{R}_i \in \mathbb{R}^{d\times h\times w}$, we randomly select another spatial feature $\mathbf{R}_j, i \neq j$ from the same image to occlude $\mathbf{R}_i$ with the following operations:

$$\begin{aligned} \bar{r}_j &= mean(\mathbf{R}_j) \\ \mathbf{M}^{h\times w} &\sim I[\text{Uniform}(0,1) > \rho] \\ \tilde{\mathbf{R}}_i &= \mathbf{R}_i * \mathbf{M} + \bar{r}_j * (1 - \mathbf{M}), \end{aligned} \tag{2}$$

where ρ is the masking ratio and I is the indicator function. $\bar{r}_j$ and $\mathbf{M}$ are broadcasted to the dimension of $\mathbf{R}_i$ in the thrid equation. With this masking strategy, some elements in $\mathbf{R}_i$ are randomly replaced by the average elements of $\mathbf{R}_j$. This operation is performed to simulate real-world occlusions at the feature level.

Once we obtain the occluded feature $\tilde{\mathbf{R}}_i$, we feed it to the Occluded Part Extrapolation (OPE) module, which is a stack of multiple transformer encoder layers. The OPE module aims to reconstruct the feature before occlusion, denoted as $\hat{\mathbf{R}}_i = \tilde{\mathbf{R}}_i + \texttt{OPE}(\tilde{\mathbf{R}}_i)$. To ensure that the reconstructed feature recovers the original details, we use L_2 loss between the reconstructed feature and the original one [19]:

$$\mathcal{L}_{recon} = ||\hat{\mathbf{R}}_i - sg(\mathbf{R}_i)||_2, \tag{3}$$

Here, $sg()$ denotes the stop gradient operation. In this way, we encourage the OPE module to learn to extrapolate the original details from the occluded feature, improving its representation robustness under occlusions. During Infer-

ence, we augment the ROI feature with the trained OPE:

$$\mathbf{R}_i \leftarrow \mathbf{R}_i + \texttt{OPE}(\mathbf{R}_i). \tag{4}$$

The residual connection allows the utilization of both reconstructed and original features, leading to more robust representations.

3.4. Conditional Contextual Mining

People interact with different objects using different parts. For example, we often `kick` a `football` with *feet* but `hold` a `baseball pat` with *hands*. Inspired by this, we propose to mine the fine-grained spatial features that are most informative for interaction recognition via a Conditional Contextual Mining (CCM) module. CCM is a stack of transformer decoder layers, with a tailored cross-attention mechanism at its core. Specifically, given the spatial feature of a human-object pair $\langle \mathbf{R}_h, \mathbf{R}_o \rangle$, we obtain the refined feature with $\hat{\mathbf{r}}_h = \texttt{CCM}(\mathbf{R}_h, \mathbf{R}_o)$ as follows:

$$\begin{aligned} \mathbf{Q}^{1\times d} &= \texttt{mean}(\mathbf{R}_o) \\ \mathbf{K}^{hw\times d}, \mathbf{V}^{hw\times d} &= \texttt{flatten}(\mathbf{R}_h) \\ \hat{\mathbf{r}}_h &= \texttt{attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}), \end{aligned} \tag{5}$$

where `attn` refers to the standard multi-head attention operation [52]. The mean pooling operation summarizes the information in $\mathbf{R}_o$ as the query, which is used to search the most informative human feature in $\mathbf{R}_h$ for interaction understanding. In addition, we also introduce a mirror operation to mine informative object clues:

$$\hat{\mathbf{r}}_o = \texttt{CCM}(\mathbf{R}_o, \mathbf{R}_h). \tag{6}$$

The obtained features $\mathbf{R}_{h,o} = [\hat{\mathbf{r}}_h; \hat{\mathbf{r}}_o]$ are then utilized for interaction recognition together with the semantic features.

3.5. Training and Inference

Training. Given the detection results, we formulate interaction recognition as a verb classification task. Since the semantic representation and spatial representation lie in different spaces, we maintain a classifier for each branch to obtain the corresponding logits. For the semantic branch, f^s is taken as the concept memory in [29], while for the spatial branch, the f^r is a linear layer learned from scratch. The logits for each human-object pair $\langle h, o \rangle$ are taken as the weighted combination of the two classifiers:

$$\mathbf{s}^v_{h,o} = \sigma(\alpha f^s(\mathbf{F}_{h,o}) + (1-\alpha) f^r(\mathbf{R}_{h,o})), \tag{7}$$

where α is a learnable parameter and $\sigma(\cdot)$ is the sigmoid function. Since there can be multiple interactions between a human-object pair, we optimize the model with the binary cross-entropy loss and focal loss [36]. The whole model is optimized with:

$$\mathcal{L}_{full} = \mathcal{L}_{cls} + \beta \mathcal{L}_{recon}, \tag{8}$$

where $\mathcal{L}_{cls}$ is the verb classification loss and β is a factor balancing the two objectives.

Inference. Following [65], the confidence score of the interaction human-object $\langle h, o \rangle$ is computed as:

$$\mathbf{s}^{hoi}_{h,o} = (s_h)^\lambda * (s_o)^\lambda * \mathbf{s}^v_{h,o}, \tag{9}$$

where s_h and s_o are the confidence scores from the detector, while $\mathbf{s}^v_{h,o}$ are the confidence of the verbs for this human-object pair. λ is set to 1 during training and 2.8 for inference to suppress overconfident false positive detections.

4. Experiments

4.1. Experimental Setups

4.1.1 Datasets

We conduct experiments using two widely used HOI-detection benchmarks, namely HICO-DET [6] and V-COCO [16]. **HICO-DET** comprises a training set of 38,118 images and a test set of 9,658 images. This dataset includes 80 object categories and 117 verb classes. Their combination results in a total of 600 interaction classes. On the other hand, **V-COCO** is built upon the MS-COCO [35] dataset and includes 5,400 and 4,946 images for training and test, respectively. The dataset involves 24 action categories and 80 object classes.

4.1.2 Evaluation Protocol

Mean Average Precision (mAP) is used as the evaluation metric in our experiments. An HOI detection result was considered a true positive only if two conditions are satisfied: (1) the predicted human and object bounding boxes have Intersection over Union (IoU) greater than 0.5 with their corresponding ground-truth boxes, and (2) the predicted action/interaction class is accurate.

Standard Setting. To evaluate our model's performance on HICO-DET, we reported the results for all 600 interaction classes under the following three categories: *full*, *rare* (less than 10 training instances), and *non-rare* (10 or more training instances) classes. For V-COCO, we provide results for action classes under two evaluation settings: *Scenario 1* and *Scenario 2*. In the former setting, the detector is required to report an empty box when no object is involved in the interaction, while the object box can be ignored in the latter.

Zero-shot Setting. We also provide the evaluation of our model under zero-shot settings on HICO-DET. In particular, we followed [20] to train the model on 480 seen interactions and also evaluated the model on the other 120 unseen interactions. We provide results on Rare First and Non-rare First settings, the former preferably selects unseen categories from tail HOIs, while the latter prefers head HOIs.

Table 1. Results (mAP×100) on the HICO-DET and V-COCO datasets. The best results are highlighted in **bold** while the second best ones are underscored.

		HICO-DET						V-COCO	
		Default Setting			Known-Object Setting				
Method	Feature Extractor	Full	Rare	Non-rare	Full	Rare	Non-rare	Scenario 1	Scenario 2
InteractNet [15]	ResNet-50-FPN	9.94	7.16	10.77	-	-	-	40.0	-
HOTR [27]	ResNet-50	25.10	17.34	27.42	-	-	-	55.2	64.4
FCL [22]	ResNet-50	25.27	20.57	26.67	27.71	22.34	28.93	52.4	-
HOI-Trans [70]	ResNet-50	26.61	19.15	28.84	29.13	20.98	31.57	52.9	-
AS-Net [8]	ResNet-50	28.87	24.25	30.25	31.74	27.07	33.14	53.9	-
SCG [64]	ResNet-50-FPN	29.26	24.61	30.65	32.87	27.89	34.35	54.2	60.9
QPIC [48]	ResNet-50	29.90	23.92	31.69	32.38	26.06	34.27	58.8	61.0
CDN [63]	ResNet-50	31.44	27.39	32.64	34.09	29.63	35.42	61.7	63.8
UPT [65]	ResNet-50	31.66	25.94	33.36	35.05	29.27	36.77	59.0	64.5
SDT [56]	ResNet-50	32.45	28.09	33.75	35.95	31.30	37.34	60.3	65.7
MUREN [28]	ResNet-50	32.87	28.67	34.12	35.52	30.88	36.91	**68.8**	**71.0**
GEN-VLKT [34]	ResNet-50	33.75	29.25	35.10	36.78	32.75	37.09	62.4	64.5
HOI-CLIP [41]	ResNet-50	34.69	31.12	35.74	37.61	34.47	38.54	63.5	64.8
PViC [66]	ResNet-50	34.69	32.14	35.45	38.14	35.38	38.97	62.8	67.8
Part-Map [59]	ResNet-50	35.15	33.71	35.58	37.56	35.87	38.06	63.0	65.1
AGER [50]	ResNet-50	36.75	33.53	37.71	39.84	35.58	40.23	65.7	69.7
RmLR [4]	ResNet-50+BERT	36.93	29.03	**39.29**	38.29	31.41	40.34	63.8	69.8
ViPLO [42]	ResNet-101+CLIP	37.22	35.45	37.75	40.61	38.82	41.15	62.2	68.0
ADA-CM [29]	ResNet-50+CLIP	38.40	37.52	38.66	-	-	-	58.6	63.9
BCOM (Ours)	ResNet-50+CLIP	**39.34**	**39.90**	39.17	**42.24**	**42.86**	**42.05**	65.8	69.9

4.1.3 Implementation Details

Our object detection pipeline utilized the DETR [5] with ResNet-50 as the backbone, which is pretrained on MS-COCO [35] and fine-tuned on the corresponding dataset as described in [65]. The images in the test set of V-COCO are filtered out for detector training. We followed [29] to use CLIP released by [45] as the backbone for the semantic branch. Both semantic and spatial adapters are implemented as a block of two linear layers, with ReLU as non-linearity in the middle. After detection, we directly filtered out instances with confidence less than 0.2 and retained 3-15 instances with the highest confidence scores for each image. The CCM and OPE are composed of two transformer decoder and encoder layers, respectively, both with a hidden dimension of 256. The ROI feature size h and w are both set to 7, the masking ratio ρ was set to 0.25, and β is set to 0.1. The model is optimized with the AdamW [39] optimizer for 15 epochs. During training, the CLIP visual encoder and the detector are kept frozen, while only their adapters, CCM and OPE are optimized. We set the learning rate of the semantic adapter to 1e-3 and the remaining modules to 1e-4. Cosine decay is employed throughout the training process. All models are trained on 4 GPUs with a batch size of 6 on each.

4.2. Comparison with State-of-the-arts

Standard Setting. Table 1 presents the comparison between our proposed BCOM and state-of-the-art HOI detection models on HICO-DET and V-COCO datasets. Specifically, our BCOM achieves new state-of-the-art performance on the HICO-DET dataset under regular and known-object settings, surpassing previous SOTA by a large margin (**+0.94** mAP). In particular, it significantly outperforms the second-best method ADA-CM, which utilizes only the CLIP representations. This result indicates the effectiveness of adapting spatial features for interaction recognition. On V-COCO, our method achieves the second-best result. It is only inferior to MUREN [28], which has heavier architecture and a longer training schedule.

Zero-shot Settings. Thanks to the effective adaptation of CLIP's knowledge, our method is able to perform HOI detection under zero-shot settings. The comparison with previous methods is shown in Table 2. It can be observed that our method also surpasses previous methods under both rare-first (RF) and Non-rare First (NF) settings. The results show that BCOM can effectively produce highly discriminative features, even for unseen verb-object compositions.

Table 2. Comparison with previous methods under zero-shot settings on the HICO-DET dataset. "RF" denotes Rare-First setting while NF indicates Non-rare First setting.

Method	Type	Unseen	Seen	Full
ATL [21]	RF	9.18	24.67	21.57
VCL [20]	RF	10.06	24.28	21.43
GEN-VLKT [34]	RF	21.36	32.91	30.56
ADA-CM [29]	RF	27.63	34.35	33.01
BCOM (Ours)	RF	**28.52**	**35.04**	**33.74**
VCL [20]	NF	16.22	18.52	18.06
ATL [21]	NF	18.25	18.78	18.67
GEN-VLKT [34]	NF	25.05	23.38	23.71
ADA-CM [29]	NF	32.41	31.13	31.39
BCOM (Ours)	NF	**33.12**	**31.76**	**32.03**

Table 3. Ablation study on the effect of bilateral adaptation.

Method	Full	Rare	Non-rare
Spatial Adaptation	33.02	28.44	34.39
Semantic Adaptation	36.33	38.00	35.83
Early Fusion	36.90	38.49	36.42
Late Fusion (Ours)	**39.34**	**39.90**	**39.17**

Table 4. Performance under various degrees of occlusion.

#Instances/image	2-5	6-10	>10	Full
wo OPE	54.23	27.35	16.88	38.62
w/ OPE	**54.41**	**27.66**	**18.53**	39.34
Δ	+0.33%	+1.13%	+9.78%	1.87%

Table 5. Ablation Study on the number of OPE layers.

# layer	Full	Rare	Non-rare
0	38.57	39.82	38.19
1	38.96	39.61	38.77
2	**39.34**	**39.90**	**39.17**
3	39.00	39.78	38.77

Table 6. Ablation Study on the occlusion strategy in OPE.

Occlusion	Full	Rare	Non-rare
Patch	38.99	39.54	38.82
Pixel (Ours)	**39.34**	**39.90**	**39.17**

Table 7. Ablation Study on the effect of CCM layers.

# layer	Full	Rare	Non-rare
0	37.44	38.88	37.00
1	38.27	39.73	37.84
2	**39.34**	39.90	**39.17**
3	39.18	**40.06**	38.91

4.3. Ablation Studies

In this section, we conducted a series of ablation studies on the HICO-DET dataset to evaluate the effectiveness of each component in our proposed BCOM.

Results on Bilateral Adaptation. We present the results on the importance of bilateral adaptation in BCOM in Table 3. The results show that both semantic adaptation and spatial adaptation are crucial for interaction recognition. Specifically, with only the spatial adaptation, the performance drops by 6.32 mAP. On the other hand, discarding the spatial adaptation branch leads to a decrease of 3.01mAP, indicating the spatial and semantic branches complement each other. Our work adopts a late fusion strategy by fusion the prediction of the two branches. We also compared it with the early fusion strategy, which combines the features of the two branches before being fed into a classifier. As can be seen from this table, the early fusion strategy underperforms the late fusion. One possible reason for this is that the semantic and spatial features lie in different feature spaces.

Results with Different Degrees of Occlusion. We explore how our model performs in different occlusions. Since occlusion occurs in crowded scenarios, we partition all images in the HICO-DET test set into three groups, according to the number of instances (*i.e.*, humans and objects) in the image. We compared the performance of the three groups with and without our OPE in Table 4. The results show that our OPE brings the highest improvement in crowded scenes (>10 instances/image), showing its effectiveness for occlusion-robust HOI understanding.

Results on Different OPE layers. We compare the complexity of the Occlude Part Extrapolation (OPE) module in Table 5. The results demonstrate that an OPE module with 2 layers delivers the best results. Increasing the layers to 3 does not lead to further performance improvements.

Results with Different Occlusion Strategies. Our work adopts a pixel-level occlusion strategy, which occludes the feature map at random locations. Another strategy is to perform patch-level masking as is done in previous approaches [3, 19, 43]. The comparison is shown in Table. 6. The results show that the patch masking strategy that is often adopted on images is less effective on feature maps.

Results on Different CCM Layers. We study the number of layers in Conditional Contextual Mining (CCM) in Table 7. It can be observed that, without CCM (0 layers), the overall performance drops by 1.90 mAP, showing that CCM is important for spatial representation mining. The performance on rare classes increases when CCM has 3 layers. Nevertheless, the overall performance degrades by 0.16, indicating potential overfitting issues.

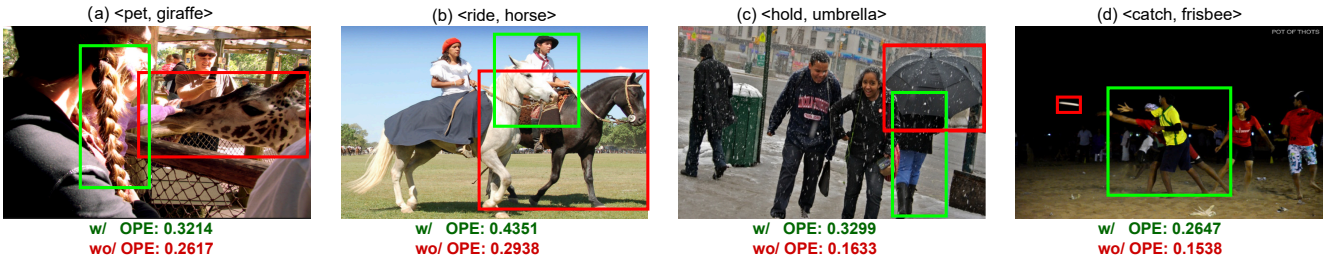

Figure 4. Qualitative results on HICO-DET test set. Our method is able to accurately recognize interactions under various occlusions and give higher confidence scores in interactions with occlusions.

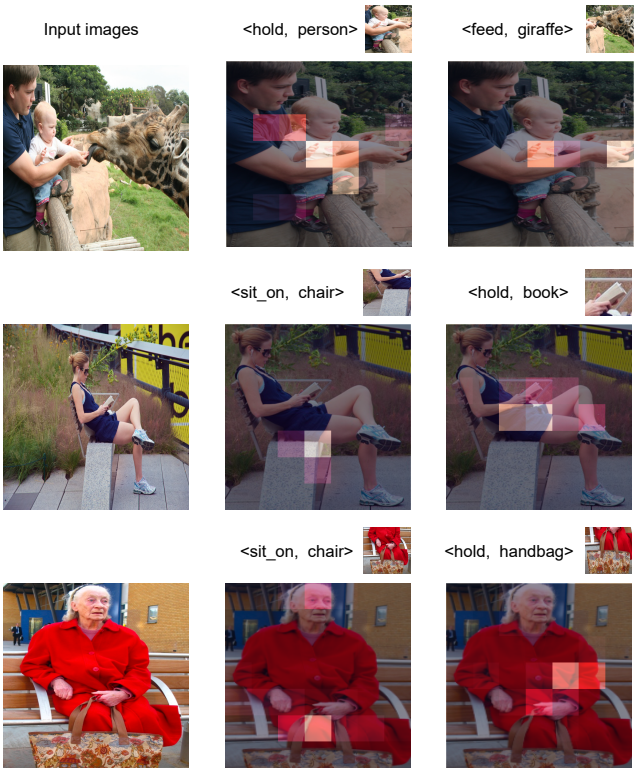

Figure 5. Attention maps generated by our Conditional Contextual Mining (CCM) mechanism. The first column displays the input image, while the small image indicates the query object *w.r.t.* the person. CCM is capable of focusing on the most important body parts of the person based on the query object.

4.4. Qualitative Results

Visualization of Occlusion Robustness. We present several qualitative results on the HICO-DET test set in Figure 4. The results demonstrate that our method is capable of accurately recognizing interactions, even under various levels of occlusion. For example, in Figure 4 (b), the person is partially occluded by a bicycle, yet our method still recognizes the interaction as `sitting on`. In Figure 4 (c), despite the heavy occlusion of the person by the involved object, our method can accurately recognize the interaction as `holding` an `umbrella` with a higher confidence score.

Visualization of CCM Attention Maps. To qualitatively understand our proposed Conditional Contextual Mining (CCM) mechanism, we randomly selected several images from the HICO-DET test set and visualized the attention maps produced by CCM in Figure 5. The results demonstrate that our CCM module allows the model to focus on the most informative contextual clues, especially human parts, for interaction recognition. For example, when using `chair` as the query in the second row of Figure 5, our CCM focuses on the bottom of the person. When `book` is used as the query, the attention highlights the area around the human hands. Similar phenomena can also be observed in the other two examples, showing that CCM is able to identify the most informative context.

5. Conclusion and Future Work

In this work, we proposed a Bilateral adaptation network with occlusion-aware COntextual Mining (BCOM) for enhancing human-object interaction recognition. Our method jointly adapts the knowledge from the pre-trained CLIP and object detector backbone to fully utilize the knowledge efficiently. Additionally, we also designed an Occluded Part Extrapolation (OPE) strategy to improve the robustness of HOI recognition under occluded scenarios. Then, we proposed a Conditional Contextual Mining (CCM) module that mines the most informative visual context in each human-object pair. Through extensive experiments on benchmark datasets, our method demonstrates superior performance over existing state-of-the-art methods. Ablation studies validate the importance of each component in our approach. For future work, we plan to explore masked-image-modeling pre-trained models to further facilitate occlusion-robust HOI understanding, which potentially empowers HOI detection with stronger capabilities.

Acknowledgement Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015. 1

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *ECCV*, 2022. 3

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*. 3, 7

[4] 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 *ICCV*, 2023. 6

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1, 2, 3, 6

[6] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 2, 5

[7] Jun Chen, Ming Hu, Boyang Li, and Mohamed Elhoseiny. Efficient self-supervised vision pretraining with local masked reconstruction. *arXiv preprint arXiv:2206.00790*, 2022. 3

[8] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, 2021. 2, 6

[9] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *NeurIPS*, 2019. 2, 4

[10] 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 *CVPR*, 2022. 2

[11] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *ICCV*, 2019. 1

[12] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022. 3

[13] Wei Feng, Wentao Liu, Tong Li, Jing Peng, Chen Qian, and Xiaolin Hu. Turbo learning framework for human-object interactions recognition and human pose estimation. In *AAAI*, 2019. 1, 2

[14] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. DRG: Dual relation graph for human-object interaction detection. In *ECCV*, 2020. 1, 2

[15] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018. 6

[16] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2, 5

[17] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *ICCV*, 2019. 1, 2

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2, 3, 4

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 3, 4, 7

[20] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020. 5, 7

[21] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *CVPR*, 2021. 7

[22] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *CVPR*, 2021. 6

[23] ASM Iftekhar, Hao Chen, Kaustav Kundu, Xinyu Li, Joseph Tighe, and Davide Modolo. What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions. In *CVPR*, 2022. 2

[24] Ying Jin, Yinpeng Chen, Lijuan Wang, Jianfeng Wang, Pei Yu, Lin Liang, Jenq-Neng Hwang, and Zicheng Liu. The overlooked classifier in human-object interaction recognition. *arXiv preprint arXiv:2203.05676*, 2022. 2

[25] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 2

[26] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. UnionDet: Union-level detector towards real-time human-object interaction detection. In *ECCV*, 2020. 2

[27] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. HOTR: End-to-end human-object interaction detection with transformers. In *CVPR*, 2021. 1, 6

[28] Sanghyun Kim, Deunsol Jung, and Minsu Cho. Relational context learning for human-object interaction detection. In *CVPR*, 2023. 6

[29] Ting Lei, Fabian Caba, Qingchao Chen, Hailin Ji, Yuxin Peng, and Yang Liu. Efficient adaptive human-object interaction detection with concept-guided memory. 2023. 2, 4, 5, 6, 7

[30] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *NeurIPS*, 2022. 2, 4

[31] 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 *CVPR*, 2019. 2

[32] 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 *CVPR*, 2020. 1

[33] 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 *CVPR*, 2020. 1, 2

[34] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. 2021. 1, 2, 4, 6, 7

[35] 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 *ECCV*, 2014. 5, 6

[36] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. 2017. 5

[37] Xinpeng Liu, Yong-Lu Li, Xiaoqian Wu, Yu-Wing Tai, Cewu Lu, and Chi-Keung Tang. Interactiveness field in human-object interactions. In *CVPR*, 2022. 2

[38] Yang Liu, Qingchao Chen, and Andrew Zisserman. Amplifying key cues for human-object-interaction detection. In *ECCV*, 2020. 1, 2

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 6

[40] Chen Min, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. Voxel-mae: Masked autoencoders for pre-training large-scale point clouds. *arXiv preprint arXiv:2206.09900*, 2022. 3

[41] Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In *CVPR*, 2023. 2, 4, 6

[42] Jeeseung Park, Jin-Woo Park, and Jong-Seok Lee. Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In *CVPR*, 2023. 1, 2, 4, 6

[43] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 3, 7

[44] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018. 1, 2

[45] 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 *ICML*, 2021. 1, 2, 4, 6

[46] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1, 2

[47] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *WACV*, 2018. 2

[48] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021. 1, 2, 6

[49] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 3

[50] Danyang Tu, Wei Sun, Guangtao Zhai, and Wei Shen. Agglomerative transformer for human-object interaction detection. In *ICCV*, 2023. 1, 6

[51] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. VSGNet: Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*, 2020. 1, 2

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1, 2, 5

[53] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE TPAMI*, 2016. 1

[54] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, 2019. 1, 2

[55] Guangzhi Wang, Yangyang Guo, Yongkang Wong, and Mohan Kankanhalli. Chairs can be stood on: Overcoming object bias in human-object interaction detection. In *ECCV*, 2022. 2

[56] Guangzhi Wang, Yangyang Guo, Yongkang Wong, and Mohan Kankanhalli. Distance matters in human-object interaction detection. In *ACM MM*, 2022. 1, 2, 4, 6

[57] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *CVPR*, 2020. 2

[58] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. In *ECCV*. Springer, 2022. 3

[59] Xiaoqian Wu, Yong-Lu Li, Xinpeng Liu, Junyi Zhang, Yuzhe Wu, and Cewu Lu. Mining cross-person cues for body-part interactiveness learning in hoi detection. In *ECCV*, pages 121–136, 2022. 1, 2, 6

[60] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022. 3

[61] Bingjie Xu, Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Interact as you intend: Intention-driven human-object interaction detection. *IEEE TMM*, 2019. 1, 2

[62] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2021. 2

[63] 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 *NeurIPS*, 2021. 1, 2, 6

[64] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *ICCV*, 2021. 1, 2, 6

[65] Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *CVPR*, 2022. 2, 4, 5, 6

[66] Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual con-

text in detecting of human-object interactions. In *ICCV*, 2023. 4, 6

[67] 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 *CVPR*, 2022. 2

[68] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *CVPR*, 2021. 1, 2

[69] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. In *CVPR*, 2022. 1, 2

[70] 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 *CVPR*, 2021. 2, 6