Dual-Prior Augmented Decoding Network for Long Tail Distribution in HOI Detection

Jiayi Gao[1], Kongming Liang[1*], Tao Wei[2], Wei Chen[2], Zhanyu Ma[1], Jun Guo[1]

[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
[2] Space AI, Li Auto
{gaojiayi, liangkongming, mazhanyu, guojun}@bupt.edu.cn, {weitao, chenwei10}@lixiang.com

Abstract

Human object interaction detection aims at localizing human-object pairs and recognizing their interactions. Trapped by the long-tailed distribution of the data, existing HOI detection methods often have difficulty recognizing the tail categories. Many approaches try to improve the recognition of HOI tasks by utilizing external knowledge (e.g. pre-trained visual-language models). However, these approaches mainly utilize external knowledge at the HOI combination level and achieve limited improvement in the tail categories. In this paper, we propose a dual-prior augmented decoding network by decomposing the HOI task into two sub-tasks: human-object pair detection and interaction recognition. For each subtask, we leverage external knowledge to enhance the model's ability at a finer granularity. Specifically, we acquire the prior candidates from an external classifier and embed them to assist the subsequent decoding process. Thus, the long-tail problem is mitigated from a coarse-to-fine level with the corresponding external knowledge. Our approach outperforms existing state-of-the-art models in various settings and significantly boosts the performance on the tail HOI categories. The source code is available at https://github.com/PRIS-CV/DP-ADN.

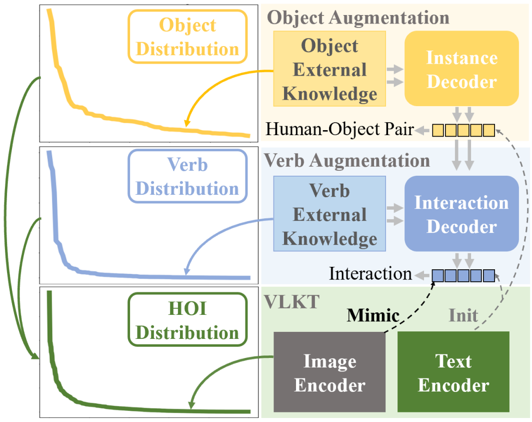

Figure 1: HICO-DET dataset (Chao et al. 2017) exhibits a severe long-tail distribution at the object, verb, and HOI combination levels. Inspired by the above observation, we utilize external knowledge to enhance the model's recognition capabilities for tail categories from the object and the verb levels, which further improves the model's ability to recognize tail categories in HOI detection.

Introduction

Human-Object Interaction (HOI) Detection aims at localizing human-object pairs and recognizing the interactions between them, which is a significant task to make the machine understand human activities in a still image. It can benefit many high-level computer vision tasks, e.g. image captioning, visual grounding, visual question answering, etc.

Current HOI detection methods can be summarized into two paradigms: one-stage methods (Xie et al. 2023), and two-stage methods (Gao, Zou, and Huang 2018; Qi et al. 2018; Kim et al. 2021). These two paradigms have made significant progress with the development of deep learning. However, in compliance with gaps in the frequency of various categories of interactions and objects in the natural world, existing HOI data presents an extremely long-tailed distribution. This poses a challenge for the recent algorithms to effectively learn the tail categories. As a common way to address the long-tailed distribution problem, external knowledge (e.g vision-language pretraining model) is utilized to enhance the model's ability on HOI detection (Radford et al. 2021; Li et al. 2021, 2022a; Bao et al. 2022). These vision-language models are trained on large-scale image text data and show great zero-shot performances in a variety of downstream tasks. Therefore, they can be regarded as a reliable external knowledge source for HOI detection (Wu et al. 2022a; Ning et al. 2023). For instance, Gen-VLKT (Liao et al. 2022) transfers the semantic representation of HOI combinations from CLIP (Radford et al. 2021) to the model via knowledge distillation, endowing the model with enhanced HOI recognition ability and zero-shot detection capability.

However, there are still issues with the way it utilizes external knowledge, among which the most questionable is that such a strategy has not led to significant improvements in model performance on the tail categories where external knowledge is most needed. How to effectively migrate external knowledge to HOI detection remains an open question. According to the explanation of VCL (Hou et al. 2020), The

[*] Corresponding author.

HOI triplet $< human, verb, object >$ can be viewed as a composition consisting of object type and interaction (verb) type since the interacting subject is identified as human. In general, human-object Interaction detection can be divided into two sub-tasks: human-object pair detection and Interaction classification. Since the decision conditions of the entire composition are more stringent than either of its elements, problems in sub-tasks also affect the entire task. As shown in Figure 1, the data corresponding to the object and verb levels have serious long-tailed distributions, which exacerbates the difficulty of the model in recognizing the tail categories at the HOI composition level. Existing methods mainly focus on the long-tail effect at the combination level, and ignore the impact at the object and verb levels. Therefore, their ability to recognize the tail HOI categories is limited.

According to the above analysis, we propose a dual-prior augmented decoding process to reduce the impact of long-tail distribution from both the object level and interaction(verb) level. Specifically, we leverage the knowledge of the external object and verb classifiers to augment the decoding process. For each decoder, we acquire the semantic prior from the obtained candidate categories. For the object prior, we hope to select the objects that participate in the interaction. For the interaction(verb) prior, we want to reduce the impact of misclassification caused by the interaction classifier. The selected prior knowledge is then embedded in the encoded image feature by the prior semantic embedding module to assist the subsequent decoding process. Finally, we introduce a Conditional DeTR decoder (Meng et al. 2021) to better extract the category semantics.

We evaluate our model on two HOI detection datasets, HICO-DET (Chao et al. 2017) and V-COCO (Gupta and Malik 2015), and conduct experiments in the fully-supervised setting and the zero-shot setting. The experimental results demonstrate our method can achieve competitive performances compared with the SOTA methods in various settings. The main contributions of our paper can be summarized as follows:

- We decompose the long-tail problem at the HOI combination level into the object and verb levels. The external knowledge is utilized to improve the learning of tailed categories at a finer granularity.
- For the two-branch decoding, we design dual-prior acquisition, dual-prior embedding, and conditional decoder to effectively utilize the subtask knowledge.
- Our method achieves state-of-the-art in various settings and significantly boosts the performance on the rare HOI combinations.

Related Work

HOI Detection. Previous HOI detection methods can be categorized into two-stage and one-stage paradigms. The two-stage methods (Gao et al. 2020; Li et al. 2020a; Ulutan, Iftekhar, and Manjunath 2020; Wan et al. 2023; Zhong et al. 2020; Yang and Zou 2020; Zhang, Campbell, and Gould 2021; Park, Park, and Lee 2023) use an independent detector to obtain object locations and categories, followed by specific modules for human-object association and interaction recognition. In contrast, the one-stage paradigm (Zhong et al. 2022; Zhou and Chi 2019; Wang et al. 2020; Zhong et al. 2021; Yuan et al. 2022b) directly detects human-object pairs with interactions, without the need for stage-wise processing. Recently, several HOI methods, inspired by DeTR-based Detectors (Zhu et al. 2020), have achieved promising performance. QPIC is the first to introduce the DeTR-based detector into HOI detection, effectively aggregating image-wide contextual information and expediting the process of HOI learning. CDN (Zhang et al. 2021) based on a cascade decoder model to mine the benefits of the two-stage and one-stage HOI detectors. We follow this structure and optimize each of the two branches, thus using the enhancements to the model's ability to recognize objects and interactions to improve the model's detection of HOI compositions.

HOI Detection with External Knowledge. There have been a number of approaches that have utilized a wide variety of external knowledge to enhance the model's ability to detect HOIs, CATN (Dong et al. 2022) introduces the categories of external detectors as a prior, and RLIP (Yuan et al. 2022a) leverages the VG dataset (Xu et al. 2017) containing relational labels for training. Recently, the Vision-Language Models (VLM) (Radford et al. 2021; Li et al. 2021, 2022a; Gao et al. 2021; Devlin et al. 2018) has demonstrated remarkable generalization capabilities across various downstream tasks (Du et al. 2022; Feng et al. 2022; Gu et al. 2021; Li, Savarese, and Hoi 2022), thus were also transferred into the HOI detection task by previous methods. GEN-VLTK (Liao et al. 2022) employs image feature distillation and initializes classifiers with HOI prompts. HOICLIP (Ning et al. 2023) uses the features obtained by the VLM (Radford et al. 2021) visual encoder. However, these methods only utilize external knowledge at a coarse level. We utilize external knowledge at different levels to better identify the tail categories.

Methodology

Model Architecture

Figure 2 depicts the overall architecture of our method. The proposed model consists of three main parts: visual encoder, dual-prior augmented decoders, and visual-linguistic knowledge transfer module (VLKT) (Liao et al. 2022). Given an input image I, a convolutional neural network (He et al. 2015) is first utilized to extract the visual features. Then the visual features are supplemented with the positional embedding P_V and further fed into the transformer encoder to obtain the feature map $V_d \in \mathbb{R}^{HW \times C}$. The dual-prior augmented decoders take I and V_d as input, and process them with object-prior and verb-prior augmented decoders in a cascaded way. These two decoders acquire object-level and verb-level external knowledge to alleviate class imbalance in human-object pair detection and interaction classification respectively. Finally, we introduce the VLKT to utilize external knowledge at the HOI combination level and improve it with a more powerful VLM named BLIP (Li et al. 2022a).

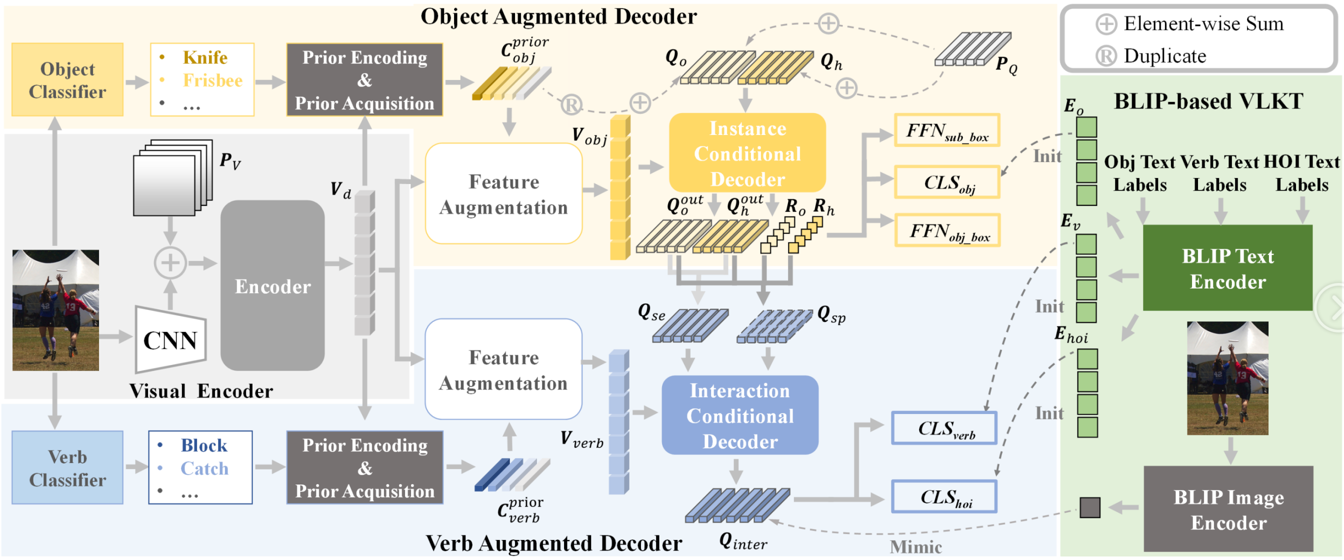

Figure 2: The Architecture of our Method. The entire model consists of the visual encoder, dual-prior augmented decoder, and Visual-Linguistic Knowledge Transfer Module. Given an image, the visual encoder extracts its visual features, and the external classifiers in dual-prior augmented decoders recognize the object and verb categories as prior knowledge. Both dual-prior augmented decoders are composed of Prior Acquisition, Prior Embedding, and Conditional Decoder modules. Based on these three modules, they use external knowledge and visual features in a similar manner to accomplish human-object pair detection and interaction recognition respectively. In addition, we utilize Visual-Linguistic Knowledge Transfer to enhance the performance of the model from the HOI combination level.

Dual-prior Augmented Decoders

The decoding process consists of object-prior augmented decoder and verb-prior augmented decoder which are connected in a cascading manner. To efficiently utilize both the object and verb prior knowledge, we designed three main modules: dual-prior acquisition, dual-prior embedding, and conditional decoding.

Dual-prior Acquisition. For an input image I, we use MLDecoder (Ridnik et al. 2023) as the external classifier to recognize the objects in it. Assume that there are E objects predicted by the external object classifier, the output object-prior categories can be denoted as $O = \{o_1, o_2, ...o_E\}$. For verb-prior acquisition, we propose prompts like "Is the person [verb]ing something?" and ask BLIP (Li et al. 2022a) VQA to give a "Yes" or "No" answer, the category corresponding to a positive answer will serve as a prior for prediction. Assume that there are F verbs predicted by the VQA model, the output verb-prior categories can be denoted as $V = \{v_1, v_2, ..., v_F\}$. Both priors can be acquired offline.

To encode the prior categories while keeping consistency between the encoding process and the final classification, we encode the priors based on the weights of the corresponding categories of the classifier. We define the weights of the object classifier CLS_{obj} in object augmented decoder and verb classifier CLS_{verb} in verb augmented decoder as $W_o = [w_1^o, w_2^o, ..., w_{N_o}^o]$ and $W_v = [w_1^v, w_2^v, ..., w_{N_v}^v]$ respectively, where $w \in \mathbb{R}^C$, N_o, N_v are the number of object and verb categories. Detailed information about CLS_{obj} and CLS_{verb} will be introduced in the subsequent sections. The predicted dual-prior can be encoded as:

$$\begin{aligned} W_o^{prior} &= \text{MLP}([w_{o_1}^o, w_{o_2}^o, ..., w_{o_E}^o]) \\ W_v^{prior} &= \text{MLP}([w_{v_1}^v, w_{v_2}^v, ..., w_{v_F}^v]) \end{aligned} \quad (1)$$

Then, we incorporate the visual feature V_d to filter the object and verb candidates in the dual-prior. For the object prior, we aim to keep the object candidates that may participate in the interaction. For the verb prior, we aim to reduce the misclassification impact caused by the interaction classifier. We get the score for each prior through an MLP and a cross-attention module (Vaswani et al. 2017) as:

$$\begin{aligned} S_o &= \text{Sigmoid}(\text{CrossAttn}(W_o^{prior}, V_d)) \\ S_v &= \text{Sigmoid}(\text{CrossAttn}(W_v^{prior}, V_d)) \end{aligned} \quad (2)$$

The input query is the predicted dual-prior W_o^{prior} and W_v^{prior}, while the key and value of the attention module are based on the image feature map V_d. Then, we get the dual-prior sets $C_o^{prior} = \{c_o | TopK_o(S_o)\}$, $C_v^{prior} = \{c_v | TopK_v(S_v)\}$, where K_o and K_v represent the prior quantities of reserved objects and verbs. Next, we introduce how to set up the supervision for S_o and S_v. Specifically, we extract the objects and interactions present in the image I and remove the categories that do not exist in the dual-prior categories O and V. The remaining categories are converted into ground-truth labels denoted as S_o^{GT} and S_v^{GT}. We use focal loss (Lin et al. 2017b) to estimate the discrepancy for S_o and S_v:

$$\begin{aligned}\mathcal{L}_o^{cls} &= \text{FocalLoss}(S_o, S_o^{GT}) \\ \mathcal{L}_v^{cls} &= \text{FocalLoss}(S_v, S_v^{GT})\end{aligned} \tag{3}$$

Dual-prior Embedding. The dual-prior embedding module aims to explicitly incorporate the filtered prior semantics into the subsequent decoding process. For the object-prior augmented decoder, we concatenate C_o^{prior} with a background embedding to obtain the final object semantic embedding. In the same way, we can acquire the final verb semantic embedding. Similar to CATN (Dong et al. 2022), we extend C_o^{prior} to the same dimension as the input query Q_o by repeating, and use it to initialize Q_o :

$$\hat{Q}_o = Q_o + \text{Repeat}(C_o^{prior}, N_q) \tag{4}$$

In the process of feature augmentation, $C_o^{prior} \in \mathbb{R}^{K_o \times C}$, $C_v^{prior} \in \mathbb{R}^{K_v \times C}$ are embedded into the encoded visual features V_d to get the object augmented visual features V_o and the verb augmented visual features V_v respectively:

$$\begin{aligned}W_o &= \text{Softmax}(\text{MLP}(V_d) \times {C_o^{prior}}^\top) \odot C_o^{prior} \\ W_v &= \text{Sigmoid}(\text{MLP}(V_d) \times {C_v^{prior}}^\top) \odot C_v^{prior}\end{aligned} \tag{5}$$

$$V_o = V_d + W_o, V_v = V_d + W_v \tag{6}$$

Since multiple verbs can occur simultaneously while only one object can occur at the same location, we set the activation function in the embedding processes of verb and object as Sigmoid and Softmax respectively.

Conditional Decoding. In order to better detect human-object pairs and recognize interactions based on the prior semantic information embedded in visual features, we improve the decoder structure.

The DeTR-based (Carion et al. 2020) decoder is a stack of decoder layers, each of which is composed of a self-attention layer, a cross-attention layer and a feed-forward layer. The cross-attention layer takes three inputs: queries, keys and values. The key is formed by adding the visual feature denoted as c_k and its corresponding positional encoding denoted as p_k. and the query formed by adding output from the self-attention layer c_q and the spatial query p_q. Thus, the attention weight can be computed as :

$$\begin{aligned}&(c_q + p_q)^\top (c_k + p_k) \\ =&c_q^\top c_k + c_q^\top p_k + p_q^\top c_k + p_q^\top p_k\end{aligned} \tag{7}$$

In order to enable the decoder to better understand the semantic information embedded in the visual features, we follow the design of Conditional DeTR (Meng et al. 2021), adopting the conditional cross-attention mechanism that forms the query by concatenating c_q and p_q and the key by concatenating c_k and p_k, thus the attention weights are:

$$c_q^\top c_k + p_q^\top p_k \tag{8}$$

This mechanism allows queries to focus independently on visual features and spatial information. The semantic information embedded in the visual features can be recognized without interference from the spatial information.

In the object-prior augmented decoder, we adopt the **c**onditional **c**ross-attention mechanism (represented as C2Decoder). Since the former is to detect the human-object pair while the latter is to recognize the interaction, the operation details could have discrepancies. For the instance conditional decoder, we use the same decoder architecture as for Conditional DeTR. Here, we set c_k as V_o and p_k as P_V. The decoder predicts the bounding box based on the reference point $R \in \mathbb{R}^{N_q \times 2}$, which is the unnormalized 2D coordinate generated from the Position Embedding P_Q using an MLP. In each decoder layer, the conditional spatial query p_q is predicted from the output query Q_i^{out} of the previous decoder layer and the reference point R:

$$p_q = \text{MLP}(Q_i^{out}) \odot \text{sinusoidal}(\text{sigmoid}(R)) \tag{9}$$

and c_q would be the output query of the self-attention layer. In this way, the instance conditional decoder takes the human query $Q_h \in \mathbb{R}^{N_q \times C}$ and $\hat{Q}_o$ for predicting the position of humans and objects. Q_h and $\hat{Q}_o$ are concatenated together to serve as the input. Thus, the decoding process of the instance decoder can be represented as:

$$Q^{out}, R = \text{C2Decoder}((Q_h : \hat{Q}_o), P_Q, V_o, P_V) \tag{10}$$

Unlike the instance conditional decoder, the interaction conditional decoder aims at obtaining the verb types and the HOI combination types of the detected human-object pairs. So the reference point is not imported here. The c_k in the verb-augmented decoder is V_v and p_k is still P_V. To construct the semantic and position queries corresponding to c_k and p_k from the output of the instance decoder, we decouple the the spatial information based on $Q^{out} = [Q_h^{out} : Q_o^{out}]$ and R:

$$\begin{aligned}S_{p_h} &= \text{MLP}(Q_h^{out}) \odot sinusoidal(R_h) \\ S_{p_o} &= \text{MLP}(Q_o^{out}) \odot \text{sinusoidal}(R_o) \\ Q_{sp} &= \text{MLP}(Concat(S_{p_h}, S_{p_o}))\end{aligned} \tag{11}$$

where we set p_q as Q_{sp} in various layers since the spatial location of the human-object pair has been determined in the instance decoder. And c_q is based on semantic query $Q_{se} \in \mathbb{R}^{N_q \times C}$ computed as $(Q_h^{out} + Q_o^{out})/2$.

In this way, we decouple the positional and semantic information Q_{sp}, Q_{se} from the output of the instance decoder, and they are used as the input of the interaction decoder. Thus, the decoding process of the interaction decoder can be represented as

$$Q_{inter} = \text{C2Decoder}(Q_{se}, Q_{sp}, V_v, P_V) \tag{12}$$

Visual-Linguistic Knowledge Transfer

In order to preserve the way of utilizing external knowledge at the combination level, following Gen-VLKT (Liao et al. 2022), we leverage and improve the Visual-Linguistic Knowledge Transfer (VLKT) module by replacing CLIP (Radford et al. 2021) with BLIP (Li et al. 2022a) as the teacher network for VLKT, so as to utilize its ability to pay attention to textual details that it possesses due to the Language Modeling loss to better recognize interactions. Specifically, we exploit the knowledge of its visual encoder by using L1 loss to pull the distance between the average value of Q_{inter} and the global features of the entire image computed by the BLIP image encoder. For the text encoder, we construct the prompt from three levels: object, verb, and combination. The prompts are designed in the way of "a person and a [object]", "a person is [verb]ing something" and "a person is [verb]ing a [object]". In this way, we obtain the corresponding semantic embeddings $E_o \in \mathbb{R}^{N_o \times C}$, $E_v \in \mathbb{R}^{N_v \times C}$, $E_{hoi} \in \mathbb{R}^{N_{hoi} \times C}$, where N_{hoi} is the number of HOI compositions. These embeddings will be used to initialize the weights of the corresponding classifiers based on one linear layer CLS_{obj}, CLS_{verb}, CLS_{hoi}. The weights are trained with a lower learning rate for preserving the knowledge in the initial text features.

Training and Inference

Training. In the training stage, we utilize multiple FFN heads to output the bounding boxes of human-object pairs. Specifically, due to the introduction of conditional decoder (Meng et al. 2021), given that $R = (R_o, R_h)$, the final bounding box of the human $B_h \in \mathbb{R}^{N_q \times 4}$ and the object $B_o \in \mathbb{R}^{N_q \times 4}$ is generated as:

$$B_o = R_o + \mathrm{FFN}\left(S_{p_o}\right), B_h = R_h + \mathrm{FFN}\left(S_{p_h}\right) \quad (13)$$

To make the output verb prior correspond to the output, Following (Ning et al. 2023), We employ MLP to extract verb features Q_{verb} from Q_{inter} and use verb classifier CLS_{obj} and HOI classifier CLS_{obj} to obtain verb and HOI combination level scores S_{verb}, S_{hoi} respectively:

$$Q_{verb} = \mathrm{MLP}(Q_{inter}) \quad (14)$$

$$S_{verb} = CLS_{verb}(Q_{verb}) \quad (15)$$

$$S_{hoi} = CLS_{hoi}(Q_{inter}) \quad (16)$$

Define the interaction score as S_{inter}, the final HOI combination logit is :

$$S = S_{hoi} + \alpha S_{verb} \quad (17)$$

Where α is a weighting parameter. The cost for bipartite matching shares the same strategy with previous methods, consisting of the box regression loss L_b, the intersection-over-union loss L_u and the classification loss L_c:

$$\mathcal{L}_{\text{cost}} = \lambda_b \sum \mathcal{L}_b^i + \lambda_u \sum \mathcal{L}_u^j + \sum \lambda_c^k \mathcal{L}_c^k \quad (18)$$

In the process of training backpropagation, we follow the training loss of Gen-VLKT, and introduce the supervision loss $\mathcal{L}_o^{cls}$ and $\mathcal{L}_v^{cls}$ for prior selection modules :

$$\mathcal{L} = \mathcal{L}_{cost} + \lambda_{mimic}\mathcal{L}_{glo} + \lambda_o^{cls}\mathcal{L}_o^{cls} + \lambda_v^{cls}\mathcal{L}_v^{cls} \quad (19)$$

where λ_b, λ_u, λ_c^k, λ_{glo},λ_o^{cls}, λ_v^{cls} are the hyper-parameters for adjusting the weights of each loss.

Inference. Following Gen-VLKT (Liao et al. 2022), we combine object scores S_{obj} output by CLS_{obj} with previous training logits as the final HOI triplet score:

$$score^n = S^n + S_{obj}^m \cdot S_{obj}^m \quad (20)$$

where n is the HOI category index and m is the object category index corresponding with n^{th} HOI category. Finally, triplet NMS is applied to top-K HOI triplets based on $score$.

Experiments

Experimental Setting

Datasets. Our experimental evaluation is based on two widely-adopted benchmarks, namely HICO-DET (Chao et al. 2017) and V-COCO (Gupta and Malik 2015). HICO-DET comprises a total of 47,776 images, with 38,118 images allocated for training and 9,658 for testing. The dataset includes annotations for 600 categories of Human-Object Interaction (HOI) triplets, derived from 80 object categories and 117 verb categories. Notably, 138 categories within the HOI set contain fewer than 10 training instances, and are therefore classified as "Rare", while the remaining 462 categories are classified as "Non-Rare".

Evaluation Metrics. We adopt the evaluation metric of mean Average Precision (mAP) used in previous works (Chao et al. 2017; Wan et al. 2019). A HOI triplet prediction is considered a true positive example if it satisfies the following criteria: 1) The Intersection over Union (IoU) of the human bounding box and the object bounding box with respect to the ground truth (GT) bounding box is greater than 0.5. 2) The predicted interaction category is accurate.

Zero-shot Setting. Following prior works (Liao et al. 2022; Hou et al. 2020), we conduct our zero-shot experiments in four different ways: Rare First Unseen Combination (RF-UC), Non-rare First Unseen Combination (NF-UC), Unseen Verb (UV), Unseen Object (UO). In the RF-UC setting, we select tail HOI categories as unseen categories, while in the NF-UC setting, we use head HOI categories as unseen categories. Under the UV and UO settings, some verb or object categories are not included in the training set, respectively. In the UC settings, all verb and object categories are present during training, but certain HOI combinations are omitted.

Implementation Details. For a fare comparison with previous methods (Newell, Yang, and Deng 2016; Lin et al. 2017a), we use ResNet-50 (He et al. 2015) as our backbone feature extractor. we follow the hyperparameter setting of Gen-VLKT (Liao et al. 2022), and the weight coefficients λ_o^{cls} and λ_v^{cls} are set to 1. We first train the model for 60 epochs with a learning rate of 10^{-4} that decreases by a factor of 10 for another 30 epochs. All experiments are conducted on 4 NVIDIA 3090 GPUs and the batch size is 16.

Comparison to State-of-the-Art

In order to further validate the effectiveness of our approach, we conduct a comparative analysis of our model's performance against the default HOI detection settings. Table 1 presents the performance evaluation on HICO-DET and V-COCO. Our method outperforms the existing methods in various settings. Especially for the rare categories, our model achieves mAP 35.8 which significantly outperforms our baseline Gen-VLKT (Liao et al. 2022) by a margin of mAP 6.25. As for the "Full" setting, our model also outperforms the existing methods. For the V-COCO dataset, we present the results in Table 1, we surpass the performance of Gen-VLKT, achieving a better performance with an AP of 62.68. However, given the smaller scale of the V-COCO dataset, featuring fewer and simpler interaction combinations, our method's improvement on this dataset is not

		mAP Deault			Known Object			VCOCO	
Method	**Backbone**	Full	Rare	Non-Rare	Full	Rare	Non-Rare	AP#1 role	AP#1 role
PPDM (Liao et al. 2019)	Hourglass-104	21.73	13.78	24.10	24.58	16.65	26.84	-	-
IDN (Li et al. 2020b)	ResNet-50	23.36	22.47	23.63	26.43	25.01	26.85	53.3	60.3
HOI-Trans (Zou et al. 2021)	ResNet-50	23.46	16.91	25.41	26.15	19.24	28.22	52.9	-
ATL (Hou et al. 2021a)	ResNet-50	28.52	21.64	30.59	31.18	24.15	33.29	-	-
AS-Net (Chen et al. 2021)	ResNet-50	28.87	25.25	30.25	31.74	27.07	33.14	53.9	-
QPIC (Tamura, Ohashi, and Yoshinaga 2021)	ResNet-50	29.07	21.85	31.23	31.68	24.14	33.93	58.8	61.0
FCL (Hou et al. 2021b)	ResNet-50	29.12	23.67	30.75	31.31	25.62	33.02	52.35	-
PhraseHOI (Li et al. 2022b)	ResNet-50	29.29	22.03	31.46	31.97	23.99	34.36	57.4	-
CATN (Dong et al. 2022)	ResNet-50	31.86	25.15	33.84	34.44	27.69	36.45	60.1	-
CDN (Zhang et al. 2021)	ResNet-50	31.78	27.55	33.05	34.53	29.73	35.96	61.8	63.8
Gen-VLKT (Liao et al. 2022)	ResNet-50	33.75	29.25	35.01	36.78	32.75	37.99	62.4	64.4
HOICLIP (Ning et al. 2023)	ResNet-50	34.69	31.21	35.74	37.61	34.47	38.54	**63.5**	64.8
ViPLO (Park, Park, and Lee 2023)	ViT-B/32	34.95	33.83	35.28	38.15	36.77	38.56	60.9	**66.6**
PartMap (Wu et al. 2022b)	ResNet-50	35.15	33.71	35.58	37.56	35.87	38.06	63.0	65.1
Ours	ResNet-50	**35.91**	**35.82**	**35.94**	**38.99**	**39.61**	**38.80**	62.62	64.8

Table 1: Comparison with state-of-the-art methods on HICO-DET and V-COCO.

Method	Type	Unseen	Seen	Full
VCL (Hou et al. 2020)	RF-UC	10.06	24.28	21.43
ATL (Hou et al. 2021a)	RF-UC	9.18	24.67	21.57
FCL (Hou et al. 2021b)	RF-UC	13.16	24.23	22.01
Gen-VLKT (Liao et al. 2022)	RF-UC	21.36	32.91	30.56
HOICLIP (Ning et al. 2023)	RF-UC	25.53	34.85	32.99
Ours	RF-UC	**31.83**	**34.95**	**34.32**
VCL (Hou et al. 2020)	NF-UC	16.22	18.52	18.06
ATL (Hou et al. 2021a)	NF-UC	18.25	18.78	18.67
FCL (Hou et al. 2021b)	NF-UC	18.66	19.55	19.37
Gen-VLKT (Liao et al. 2022)	NF-UC	25.05	23.38	23.71
HOICLIP (Ning et al. 2023)	NF-UC	**26.39**	**28.10**	**27.75**
Ours	NF-UC	26.37	25.50	25.67
ATL (Hou et al. 2021a)	UO	5.05	14.69	13.08
FCL (Hou et al. 2021b)	UO	0.00	13.71	11.43
Gen-VLKT (Liao et al. 2022)	UO	10.51	28.92	25.63
HOICLIP (Ning et al. 2023)	UO	16.20	30.99	28.53
Ours	UO	**16.42**	**31.75**	**29.20**
Gen-VLKT (Liao et al. 2022)	UV	20.96	30.23	28.74
HOICLIP (Ning et al. 2023)	UV	24.30	**32.19**	31.09
Ours	UV	**27.45**	31.99	**31.35**

Table 2: Comparison with state-of-the-art methods under zero-shot settings. In the table, RF is short for rare first, NF is short for non-rare first, and UO, UV indicate unseen object and unseen verb settings, respectively.

Method	Full	Rare	Non-Rare
Random	34.77	33.48	35.15
BLIP Text Embedding	35.35	34.48	35.61
Ours	35.91	35.82	35.94

Table 3: Comparison with different prior encoding strategies.

DA(o)	DE(o)	DE(v)	DA(v)	C2D	Full	Rare	Non-Rare
baseline(CLIP-based VLKT)					33.75	29.25	35.01
baseline(BLIP-based VLKT)					33.85	30.49	34.86
✓					33.93	31.47	34.66
✓	✓				34.34	31.67	35.13
✓	✓	✓			34.54	33.73	34.78
✓	✓	✓	✓		35.14	35.10	35.16
✓	✓	✓	✓	✓	35.91	35.82	35.94

Table 4: Ablation study. The results about Dual-prior Acquisition (DA), Dual-prior Embedding (DE)), and Conditional Decoder (C2D). (o) signifies the effect of the module in the object-prior augmented decoder and (v) represents its effect in the verb-augmented decoder.

as significant as observed in HICO-DET. Moreover, we did not use the conditional decoder module for the V-COCO dataset, because many examples do not have interactive objects, which makes it difficult for the model to detect based on the reference point. We also perform experiments in zero-shot settings, including RF-UC, NF-UC, UV, and UO. The results are presented in Table 2, Our method achieves state-of-the-art performance across multiple settings. Compared with Gen-VLKT, we achieve a remarkable +10.47 mAP increase under RF-UC settings across unseen categories and a noteworthy +6.49 mAP improvement for unseen categories under the UV setting.

Model Analysis

Mitigation for the long tail effect. We analyze the model's ability to address the impact caused by long-tail distribution in three settings: HOI combinations, verbs, and objects. For each object or verb category, we take the average AP of its corresponding HOI combinations and consider it as

K_o	Rare	Non-Rare	K_v	Rare	Non-Rare
3	34.94	35.26	4	34.23	35.55
4	35.50	35.80	5	35.50	35.80
5	34.59	35.96	6	35.46	35.49

Table 5: Comparison with different K_o and K_v

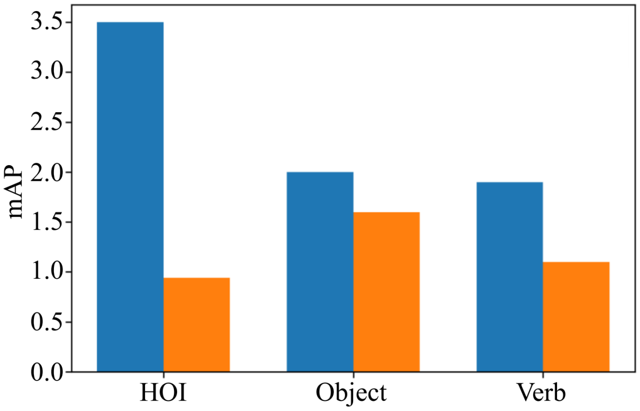

Figure 3: Mitigation for the long tail effect at the HOI, verb, and object levels.

the AP of the category. The top one-third of categories are designated by sample size as the head categories, while the remaining two-thirds are categorized as tail categories. We calculate the average difference between our method and the baseline (Liao et al. 2022) for each head category and tail category based on mAP. As shown in Figure 3, the average mAP difference for the tail categories (blue) is stronger than that for the head categories (orange), both at the object level and at the verb level, indicating that the external knowledge improves the model's ability to deal with the long-tail effect at a finer granularity, thus has a very obvious improvement on the overall combination level.

BLIP-Based VLKT evaluation. We compare the effects of Gen-VLKT (Liao et al. 2022) based on CLIP (Radford et al. 2021) and BLIP (Li et al. 2022a) as shown in Table 4, and BLIP-based Gen-VLKT improves the model's performance in the tail categories. Compared with CLIP, BLIP adopted LM loss in the training process, thus able to pay more attention to the details of the labels, and better recognize the fine-grained interaction information.

Prior Encoding. We try different strategies for encoding the prior, see Table 3, and the existing classifier weight-based encoding achieves optimal results, which shows that encoding the prior knowledge based on classifier weights can maintain the consistency of prior semantics, allowing the model to better leverage external knowledge.

Dual Prior Selection Module. We test the effect of the priority selection module on the object branch and the interaction branch respectively in Table 4. It plays an important role in the introduction of both the object and the verb priors and it improves the AP of the Rare categories by 0.98 and 1.5 respectively. For the choice of hyperparameter, we train the model with different K_o and K_v, it achieves the best performance when $K_o = 4$ and $K_v = 5$ as shown in Table 5.

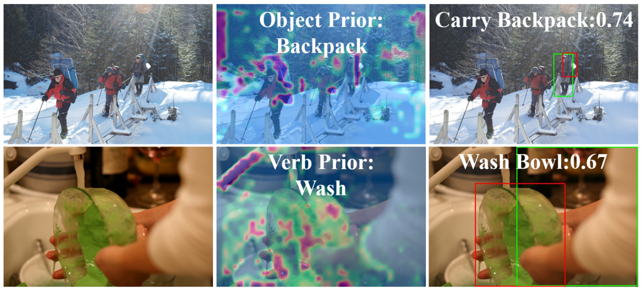

Figure 4: Visualization of predictions. The columns indicate the input image (left), the attention map for the prior category in the Dual-prior Embedding module (middle), and the final prediction result of our method (right).

Dual Prior Embedding. As shown in Table 4, the Prior Semantic Embedding module brings significant improvement, we analyze its principle by visualization. Figure 4 shows two images misrecognized by the previous method. The Prior Selection module selects the prior categories of 'backpack' (up) and 'wash' (down) for the two images respectively. Based on the weight maps, we illustrate the attended regions of the prior semantics during the computation of W_o and W_v. It can be observed that the areas of interest for the object prior include all locations in the prior images where backpacks are present. Meanwhile, the interaction branch concentrates on the main areas related to the verb 'wash', such as the bowl, hand, and faucet.

Conditional Decoder. According to Table 4, the conditional decoder improves the performance of the model on both head and tail categories. This suggests that conditional cross-attention can better assist the model in understanding visual features augmented by prior knowledge.

Conclusion

In this paper, we propose a novel dual-prior augmented network to deal with the long-tail problem in the human object interaction detection task. We decouple the HOI detection task into human-object pair detection and interaction recognition tasks and introduce external knowledge separately to alleviate the impact of the long-tail effect on these sub-tasks. We design dual-prior acquisition, dual-prior embedding, and conditional decoder to effectively utilize external knowledge. Our method achieves state-of-the-art performance across diverse configurations and demonstrates promising capability in detecting rare HOI categories.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (NSFC) No. 62106022, U19B2036, 62225601, U23B2052, in part by Beijing Natural Science Foundation Project No. Z200002 and in part by Youth Innovative Research Team of BUPT No. 2023QNTD02.

References

Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O. K.; Aggarwal, K.; Som, S.; Piao, S.; and Wei, F. 2022. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 32897–32912. Curran Associates, Inc.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2017. Learning to Detect Human-Object Interactions. *workshop on applications of computer vision*.

Chen, M.; Liao, Y.; Liu, S.; Chen, Z.; Wang, F.; and Qian, C. 2021. Reformulating HOI Detection as Adaptive Set Prediction. *computer vision and pattern recognition*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, L.; Li, Z.; Xu, K.; Zhang, Z.; Yan, L.; Zhong, S.; and Zou, X. 2022. Category-aware transformer network for better human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19538–19547.

Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; and Li, G. 2022. Learning to Prompt for Open-Vocabulary Object Detection with Vision-Language Model.

Feng, C.; Zhong, Y.; Jie, Z.; Chu, X.; Ren, H.; Wei, X.; Xie, W.; and Ma, L. 2022. PromptDet: Towards Open-vocabulary Detection using Uncurated Images.

Gao, C.; Xu, J.; Zou, Y.; and Huang, J.-B. 2020. Drg: Dual relation graph for human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 696–712. Springer.

Gao, C.; Zou, Y.; and Huang, J.-B. 2018. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. *Learning*.

Gupta, S.; and Malik, J. 2015. Visual Semantic Role Labeling. *arXiv: Computer Vision and Pattern Recognition*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *arXiv: Computer Vision and Pattern Recognition*.

Hou, Z.; Peng, X.; Qiao, Y.; and Tao, D. 2020. Visual Compositional Learning for Human-Object Interaction Detection. *european conference on computer vision*.

Hou, Z.; Yu, B.; Qiao, Y.; Peng, X.; and Tao, D. 2021a. Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 495–504.

Hou, Z.; Yu, B.; Qiao, Y.; Peng, X.; and Tao, D. 2021b. Detecting Human-Object Interaction via Fabricated Compositional Learning. *computer vision and pattern recognition*.

Kim, D.-J.; Sun, X.; Choi, J.; Lin, S.; and Kweon, I. S. 2021. Acp++: Action co-occurrence priors for human-object interaction detection. *IEEE Transactions on Image Processing*, 30: 9150–9163.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, J.; Savarese, S.; and Hoi, S. C. H. 2022. Masked Unsupervised Self-training for Zero-shot Image Classification.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, Y.-L.; Liu, X.; Lu, H.; Wang, S.; Liu, J.; Li, J.; and Lu, C. 2020a. Detailed 2d-3d joint representation for human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10166–10175.

Li, Y.-L.; Liu, X.; Wu, X.; Li, Y.; and Lu, C. 2020b. Hoi analysis: Integrating and decomposing human-object interaction. *Advances in Neural Information Processing Systems*, 33: 5011–5022.

Li, Z.; Zou, C.; Zhao, Y.; Li, B.; and Zhong, S. 2022b. Improving human-object interaction detection via phrase learning and label composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1509–1517.

Liao, Y.; Liu, S.; Wang, F.; Yanjie, C.; Qian, C.; and Feng, J. 2019. PPDM: Parallel Point Detection and Matching for Real-time Human-Object Interaction Detection. *computer vision and pattern recognition*.

Liao, Y.; Zhang, A.; Lu, M.; Wang, Y.; Li, X.; and Liu, S. 2022. GEN-VLKT: Simplify Association and Enhance Interaction Understanding for HOI Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20123–20132.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017a. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017b. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; and Wang, J. 2021. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3651–3660.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, 483–499. Springer.

Ning, S.; Qiu, L.; Liu, Y.; and He, X. 2023. HOICLIP: Efficient Knowledge Transfer for HOI Detection with Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23507–23517.

Park, J.; Park, J.-W.; and Lee, J.-S. 2023. 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*, 17152–17162.

Qi, S.; Wang, W.; Jia, B.; Shen, J.; and Zhu, S.-C. 2018. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, 401–417.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. *international conference on machine learning*.

Ridnik, T.; Sharir, G.; Ben-Cohen, A.; Ben-Baruch, E.; and Noy, A. 2023. Ml-decoder: Scalable and versatile classification head. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 32–41.

Tamura, M.; Ohashi, H.; and Yoshinaga, T. 2021. QPIC: Query-Based Pairwise Human-Object Interaction Detection with Image-Wide Contextual Information. *computer vision and pattern recognition*.

Ulutan, O.; Iftekhar, A. S. M.; and Manjunath, B. 2020. VSGNet: Spatial Attention Network for Detecting Human Object Interactions Using Graph Convolutions. *computer vision and pattern recognition*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wan, B.; Liu, Y.; Zhou, D.; Tuytelaars, T.; and He, X. 2023. Weakly-supervised HOI Detection via Prior-guided Bi-level Representation Learning. *arXiv preprint arXiv:2303.01313*.

Wan, B.; Zhou, D.; Liu, Y.; Li, R.; and He, X. 2019. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9469–9478.

Wang, T.; Yang, T.; Danelljan, M.; Khan, F. S.; Zhang, X.; and Sun, J. 2020. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4116–4125.

Wu, M.; Gu, J.; Shen, Y.; Lin, M.; Chen, C.; Sun, X.; and Ji, R. 2022a. End-to-End Zero-Shot HOI Detection via Vision and Language Knowledge Distillation. *arXiv preprint arXiv:2204.03541*.

Wu, X.; Li, Y.-L.; Liu, X.; Zhang, J.; Wu, Y.; and Lu, C. 2022b. Mining cross-person cues for body-part interactiveness learning in hoi detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, 121–136. Springer.

Xie, C.; Zeng, F.; Hu, Y.; Liang, S.; and Wei, Y. 2023. Category Query Learning for Human-Object Interaction Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15275–15284.

Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5410–5419.

Yang, D.; and Zou, Y. 2020. A Graph-based Interactive Reasoning for Human-Object Interaction Detection. *international joint conference on artificial intelligence*.

Yuan, H.; Jiang, J.; Albanie, S.; Feng, T.; Huang, Z.; Ni, D.; and Tang, M. 2022a. Rlip: Relational language-image pre-training for human-object interaction detection. *Advances in Neural Information Processing Systems*, 35: 37416–37431.

Yuan, H.; Wang, M.; Ni, D.; and Xu, L. 2022b. Detecting human-object interactions with object-guided cross-modal calibrated semantics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3206–3214.

Zhang, A.; Liao, Y.; Liu, S.; Lu, M.; Wang, Y.; Gao, C.; and Li, X. 2021. Mining the benefits of two-stage and one-stage hoi detection. *Advances in Neural Information Processing Systems*, 34: 17209–17220.

Zhang, F. Z.; Campbell, D.; and Gould, S. 2021. Spatially Conditioned Graphs for Detecting Human-Object Interactions. *international conference on computer vision*.

Zhong, X.; Ding, C.; Li, Z.; and Huang, S. 2022. Towards Hard-Positive Query Mining for DETR-Based Human-Object Interaction Detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, 444–460. Springer.

Zhong, X.; Ding, C.; Qu, X.; and Tao, D. 2020. Polysemy deciphering network for human-object interaction detection. In *European Conference on Computer Vision*, 69–85. Springer.

Zhong, X.; Qu, X.; Ding, C.; and Tao, D. 2021. 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*, 13234–13243.

Zhou, P.; and Chi, M. 2019. Relation Parsing Neural Network for Human-Object Interaction Detection. *international conference on computer vision*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

Zou, C.; Wang, B.; Hu, Y.; Liu, J.; Wu, Q.; Zhao, Y.; Li, B.; Zhang, C.; Zhang, C.; Wei, Y.; and Sun, J. 2021. End-to-End Human Object Interaction Detection with HOI Transformer. *computer vision and pattern recognition*.