ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447511

CONTEXTUAL HUMAN OBJECT INTERACTION UNDERSTANDING FROM PRE-TRAINED LARGE LANGUAGE MODEL

Jianjun Gao[1], *Kim-Hui Yap*[1], *Kejun Wu*[1], *Duc Tri Phan*[1], *Kratika Garg*[2], *Boon Siew Han*[2]

[1] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.
[2] Schaeffler Hub for Advanced Research at NTU, Singapore

ABSTRACT

Existing human object interaction (HOI) detection methods have introduced zero-shot learning techniques to recognize unseen interactions, but they still have limitations in understanding context information and comprehensive reasoning. To overcome these limitations, we propose a novel HOI learning framework, ContextHOI, which serves as an effective contextual HOI detector to enhance contextual understanding and zero-shot reasoning ability. The main contributions of the proposed ContextHOI are a novel context-mining decoder and a powerful interaction reasoning large language model (LLM). The context-mining decoder aims to extract linguistic contextual information from a pre-trained vision-language model. Based on the extracted context information, the proposed interaction reasoning LLM further enhances the zero-shot reasoning ability by leveraging rich linguistic knowledge. Extensive evaluation demonstrates that our proposed framework outperforms existing zero-shot methods on the HICO-DET and SWIG-HOI datasets, as high as 19.34% mAP on unseen interaction can be achieved.

***Index Terms*—** Human Object Interaction, Zero-shot Learning, Vision-Language Model, Context Learning, Interaction Reasoning

1. INTRODUCTION

Human object interaction (HOI) detection aims to detect humans and objects and recognize their interactions. The detection results are given in the form of a triplet <human, interaction, object>. HOI detection takes important roles in various tasks, such as human-robot interaction, image and video understanding, and image captioning. Existing HOI methods can be categorized into one- or two-stage methods depending on whether object detection and association are performed in one or two steps. Two-stage methods [1, 2, 3, 4] are generally developed from the off-the-shelf object detectors like Faster RCNN [5] and DETR [6] with separate association algorithms. Most one-stage methods [7, 8, 9, 10] are developed from one-stage object detectors by introducing unified object detection and interaction association frameworks. Although existing solutions have been developed with advances to detect interactions, they still suffer from detecting unseen interaction categories and limited zero-shot learning ability. To mitigate the gap, several methods [7, 8] have proposed different zero-shot learning strategies in HOI detection. However, existing zero-shot methods face challenges when attempting to identify the unseen interaction depicted in Fig. 1 (a) These challenges arise from a lack of contextual information, such as the relative positions of objects (e.g., "beside") and tools (e.g., "hose"), as well as the need for a substantial knowledge base to facilitate zero-shot reasoning about such interactions.

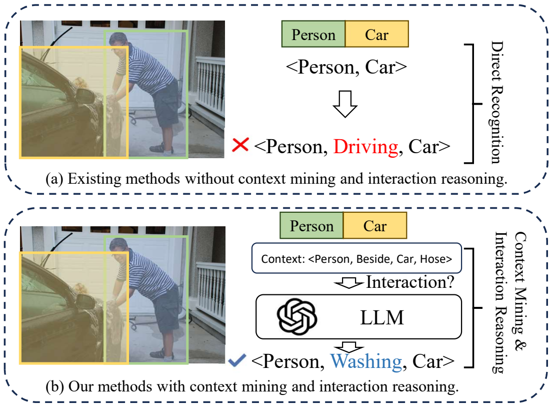
Person
Car
<Person, Car>
✗ <Person, Driving, Car>
Direct Recognition
(a) Existing methods without context mining and interaction reasoning.
Person
Car
Context: <Person, Beside, Car, Hose>
Interaction?
LLM
✓ <Person, Washing, Car>
Context Mining & Interaction Reasoning
(b) Our methods with context mining and interaction reasoning.

Fig. 1: The motivation of our proposed contextual HOI detector. Compared with existing methods (a) that recognize interactions directly, our proposed method (b), with context mining and interaction reasoning, can perform better on interaction recognition.

To overcome these challenges in existing zero-shot HOI methods, we propose ContextHOI, a contextual HOI detector to enhance both contextual and zero-shot learning ability for detecting novel interactions. We first propose a context-mining decoder to extract the contextual information from a text instruction and images. The context-mining decoder consists of a pre-trained vision-language model and a transformer decoder. The pre-trained vision-language model is responsible for generating the linguistic context embeddings, while the transformer decoder is proposed to obtain context-enhanced HOI embeddings. With the context-enhanced HOI embeddings, we develop an interaction reasoning LLM to

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its IAF-ICP Programme I2001E0067 and the Schaeffler Hub for Advanced Research at NTU.

enhance unseen interaction reasoning empowered by the zero-shot learning ability of LLMs. In detail, we utilize an encoder-decoder LLM model to reason interactions from rich linguistic knowledge as a visual question-answering task. The insights of the proposed ContextHOI are as shown in Fig. 1 (b). It can be observed that our method can recognize ambiguous interactions while typical solutions struggle to distinguish.

Our contributions can be summarized as follows. (i) We develop a context-mining decoder to discover more informative context cues from a visual-language model. (ii) We propose an interaction reasoning LLM to transfer the interaction recognition task as a visual question-answering task to enhance the zero-shot reasoning ability. (iii) We evaluated our methods on two popular HOI datasets, HICO-DET and SWIG-HOI, and the evaluation results show our advances over existing zero-shot HOI detection methods.

2. RELATED WORK

2.1. Generic HOI Detectors

Generic HOI detectors fall into one-stage and two-stage methods. They share the same goal of detecting and associating objects in a triplet of <human, interaction, object>. Two-stage methods typically build on existing object detectors, like Faster-RCNN [5] or DETR [6], with separate algorithms for interaction association. Some methods [1, 2, 4, 11] rely solely on detected objects to infer interactions, while others [12, 13, 14, 15] consider both the detected objects and the entire image or union regions as well as additional cues like as human poses [16] and distances [17, 14]. In contrast, one-stage methods combine object detection and interaction association in a unified manner. These models [7, 8, 9, 18, 10] often take the form of a two-branch network, where one branch handles object detection, and the other focuses on interaction recognition.

2.2. Zero-shot HOI Detectors

Zero-shot HOI Detectors have the goal of recognizing interaction categories that were not seen during the training phase. This involves transferring knowledge gained from seen interactions to identify unseen ones. In this context, a method presented in [19, 20, 21] decomposes HOI detections into separate verb and object networks. Other approaches, such as those discussed in [22, 23, 24, 25, 26], employ techniques like functionalities, analogies, and semantics to transfer visual phrase embeddings from known training triplets to unseen test triplets. More recently, linguistic knowledge [7, 8] has been leveraged to enhance the recognition of unseen interactions by integrating vision-language models like CLIP [27].

3. METHODS

3.1. Overview

The overview structure of our proposed HOI detector is illustrated in Fig. 2. Given an image $I \in \mathbb{R}^{224\times224\times3}$, we first employ a HOI vision encoder to encode the image in conjunction with the initial HOI tokens into image and HOI embeddings. These embeddings are then subjected to two processing branches. The first branch, known as the bounding box head, focuses on regressing and classifying detected humans and objects, yielding detection results as (b_p, b_o, c), where b_p, b_o, and c are bounding boxes for persons and objects and categories for objects. Meanwhile, the second branch comprises our proposed context-mining decoder and an interaction reasoning LLM. Encoded image embeddings and HOI embeddings are first fed into the context-mining decoder to extract linguistic context information for HOI embeddings. Once context features are obtained, we use our proposed interaction reasoning LLM to recognize interactions from extensive linguistic knowledge. The inputs in the text encoder side follow the settings in our baseline model [8]. The inputs can be formulated as [PREFIX], [ACT], [CONJUN], and [OBJ] as the input tokens for the text encoder, where [PREFIX] and [CONJUN] are learnable tokens to automatically define prefix and conjunctions, and [ACT] and [OBJ] are true action and object categories. The encoded text embeddings are used as supervision for final HOI embeddings for HOI learning.

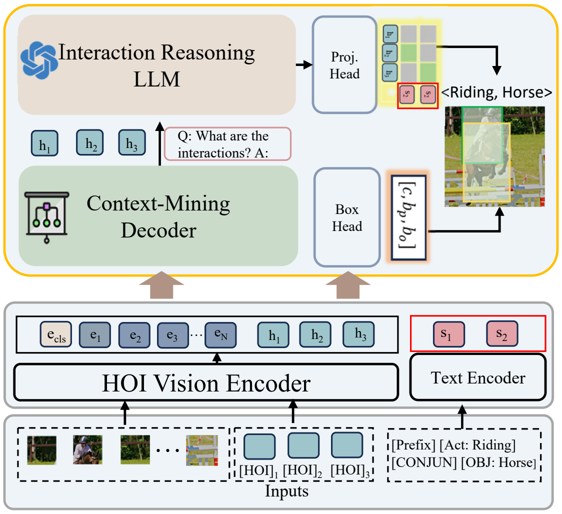

Fig. 2: The overview of our proposed ContextHOI. The HOI vision encoder and text encoder are first used to encode image patches with initial HOI tokens and text ground truth into the image, HOI, and text embeddings. After that, the context-mining decoder is proposed to extract context information for HOI embeddings. Subsequently, the interaction reasoning LLM is proposed to enhance zero-shot reasoning ability in HOI embeddings. The similarities between final HOI embeddings after projection head and text embeddings are calculated for interaction learning. Apart from that, a box head is proposed to localize humans and objects from image embeddings and HOI embeddings from the HOI vision encoder.

3.2. Context-mining Decoder

To enhance the context-understanding ability, we proposed a context-mining decoder incorporating a large vision-language model. Given the HOI embeddings $H = [h_1, h_2, ..h_M] \in \mathbb{R}^{M\times D}$ and image embeddings $E = [e_{cls}, e_1, e_2, ..., e_N] \in \mathbb{R}^{N\times D}$ encoded from the HOI vision encoder, the context-mining decoder learn the context information for HOI em-

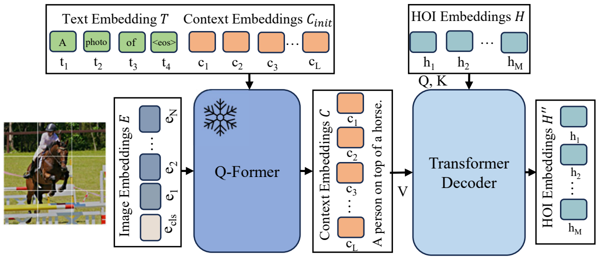

Fig. 3: The structure of our proposed context-mining decoder. It employs a Q-Former to generate context embeddings that describe the input image. Then, a transformer decoder is proposed to learn the context information for HOI embeddings.

bedding from transferred linguistic context embeddings $C = [c_1, c_2, ..., c_L] \in \mathbb{R}^{L \times D}$, where M, N, and L are the number of HOI, image, and context embeddings and D is the channel dimension. This process is illustrated in Fig. 3.

The Q-former in Fig. 3 aims to generate linguistic context embedding from visual features. The image embeddings are first passed through the Q-former $QF(\cdot)$, which is defined by:

$$C = QF(C_{init}, T, E), \tag{1}$$

where the initialized context embeddings C_{init}, text embeddings T, and image embeddings E are the inputs of the Q-Former. Text description "a photo of <eos>" is encoded into text embeddings $T = [t_1, t_2, t_3, t_4] \in \mathbb{R}^{4 \times C}$, where <eos> is the end-of-sentence token. Since the Q-Former is trained for image-grounded text generation, it enables the context embeddings to learn comprehensive linguistic context information as a description of a given image. With the text instructions T and image content E, the initialized context embeddings are updated and generated to describe the image and obtain information like "a person on top of a horse".

The transformer decoder in Fig. 3 aims to transfer the context information to HOI embeddings. The HOI embeddings only assess the visual and spatial information without context and semantic information before going through the decoder. We apply a transformer decoder on the HOI embeddings and context embeddings to learn the detailed context information for individual HOI embeddings. The learning process can be formulated as

$$H' = MHA(H, C, C) + H, \tag{2}$$

$$H'' = MLP(LN(H')) + H', \tag{3}$$

where $MHA(\cdot)$ is the multi-head attention, LN is the layer norm, and MLP is the multilayer perception. In the multi-head attention, we use interaction embeddings H as the query and context embeddings C as the key and value to extract context information for each HOI embedding.

3.3. Interacction Reasoning LLM

Inspired by the demonstrated zero-shot learning abilities of LLMs [28] in various vision tasks [29], we propose an interaction reasoning LLM by leveraging Flan-T5 [28] as shown

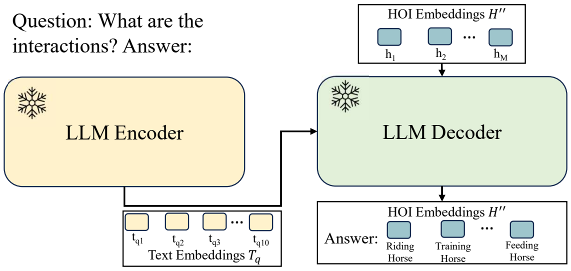

Fig. 4: The proposed interaction reasoning LLM. In the encoder, we encode the question as the instruction for the following reasoning in the decoder. In the decoder, we reason the interaction from the HOI embeddings with linguistic context information.

in Fig. 4. Flan-T5 has an encoder-decoder structure and is capable of zero-shot question answering. Thus, we take advantage of Flan-T5 and transfer our interaction recognition task into a visual question-answering task.

In detail, we provide a question for LLM and let it answer the question according to HOI embeddings. On the encoder side, we give the text input question as "Question: What are the interactions? Answer: <eos>". And we use the encoder from Flan-T5 to encode the question to text embeddings $T_q = [t_{q1}, t_{q2}, ..., t_{q10}] \in \mathbb{R}^{10 \times C}$. On the decoder side, we task the HOI embeddings from the previous context-mining decoder as inputs. The pre-trained LLM decoder then reasons the interactions for interaction embeddings according to the encoded question and the rich linguistic knowledge:

$$H''' = H'' + Decoder(H'', T_q) \tag{4}$$

where H'' is the HOI embeddings output from the context-mining decoder with linguistic context knowledge, and H'' is the output for the LLM decoder. With the output embeddings, we follow the training setting in our baseline THID [8], where the supervision is from the CLIP text encoder.

4. EXPERIMENTS

4.1. Experiment Settings

Datasets. Following our baseline, We conduct experiments based on two HOI detection datasets, HICO-DET and SWIG-HOI. HICO-DET differs from SWIG-HOI, and we must stimulate the zero-shot learning setting, while SWIG-HOI originally has the zero-shot learning setting. For HICO-DET, it includes 600 combinations of 117 human interactions and 80 objects. During the training phase, we only use 480 interactions and 120 interactions as unseen interactions. For SWIG-HOI, we use the natural dataset settings, and it includes about 5.5k interactions with around 1.2k unseen interactions.

Implementaion Details. For the training setting, we strictly follow the settings in THID. We use the pre-trained CLIP for the image and text encoders and freeze them during training. We also adopted the Q-former from BLIP2 [29] and the encoder-decoder Flan-T5 [28] as the LLM for interaction reasoning. We also freeze them during the training phase. The model is trained on 3 GPUs with a batch size of 8. We train our model for 100 epochs and set the initial learning rate as 0.0001, and use the Adam optimizer and decoupled weight decay regularization.

Table 1: The performance comparisons on the HICO-DET dataset under stimulated zero-shot learning settings.

Method	One-stage	Unseen	Seen	Full
Shen (WACV18) [19]	×	5.62	-	6.26
FG (AAAI20) [22]	×	10.93	12.60	12.26
VCL (ECCV20) [20]	×	10.06	24.28	21.43
ATL (CVPR21) [23]	×	9.18	24.67	21.57
FCL (CVPR21) [21]	×	13.16	24.23	22.01
THID (CVPR22) [8]	✓	15.53	24.32	22.96
ContextHOI (Ours)	✓	**19.34**	**25.33**	**24.01**

Table 2: The performance comparisons on SWIG-HOI dataset under zero-shot learning settings.

Method	Non-rare	Rare	Unseen	Full
GSR (ECCV20) [25]	10.01	6.10	2.23	6.08
CHOID (ICCV21) [26]	10.93	6.63	2.64	6.64
QPIC (CVPR21) [15]	16.95	10.84	6.21	11.12
THID (CVPR22) [8]	17.67	12.82	10.04	13.26
ContextHOI (Ours)	**18.66**	**13.32**	**10.69**	**13.92**

Evaluation Metrics. We adopt mean Average Precision to evaluate HOI detection performance. To detect the true positives, the model needs to detect the human and object bounding boxes correctly with an intersection-over-union (IoU) of 0.5 to the ground truth, and the categories of interaction should be predicted correctly. To evaluate the zero-shot learning ability, the interactions are split into non-rare, rare, and unseen cases according to the occurrences in the training set.

Table 3: Results for ablation study of our proposed components comparing with our baseline THID [8].

Method	Unseen	Seen	Full
Baseline	15.53	24.32	22.96
Baseline + CMD	18.23	25.21	23.66
Baseline + CMD + IRL	**19.34**	**25.33**	**24.01**

4.2. Evaluation Results

Benchmark Results. We perform benchmark evaluation on two benchmark datasets compared with multiple state-of-the-art methods. For evaluation on the HICO-DET dataset, we compare our method with methods in Table. 1, as they follow the same zero-shot learning settings. According to the comparison, our method outperforms all the listed methods on all the unseen, seen, and full cases. Noticeably, our model can obtain a 3.81 mAP gain on unseen interaction recognition. As for the evaluation results on the SWIG dataset, we also obtained improvements compared with our baseline THID [8] and other methods in Table. 2. Even though SWIG-HOI is a large-scale dataset including tremendous unseen interactions, our model can still outperform the state-of-the-art methods on non-rare, rare, and unseen cases.

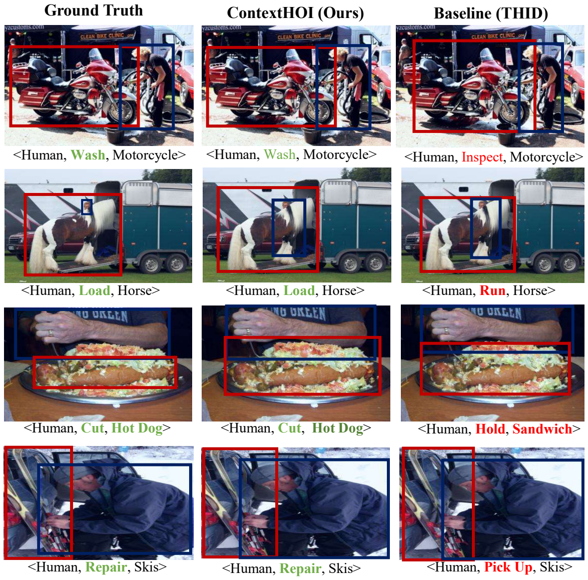

Fig. 5: Visualization of sampled Ground Truth, results from our ContextHOI and baseline.

Visualization. In this section, we show some visualized results from the HICO-DET dataset to demonstrate the advantages of our proposed method. First, we show some successful cases in Fig. 5. We can see that our method can detect the objects and the interactions correctly by obtaining more context information and zero-shot reasoning ability. Our method can understand interactions not only from human and object pairs but also from contextual information. For example, our methods can recognize interactions in the first three samples as it can understand the context like hose, knife, and truck. Besides, the powerful zero-shot reasoning ability enables our methods to recognize interactions in complex environments like in the last sample.

Ablation Study. We conduct the ablation study by adding the context-mining decoder (CMD) and interaction reasoning LLM (IRL) into our baseline step-by-step on the HICO-DET dataset. The results in Table. 3 demonstrates that our proposed method can learn better context information and show better zero-shot reasoning ability.

5. CONCLUSION

In this work, we proposed a ContextHOI detector to enhance the zero-shot HOI detectors via a context-mining decoder and an interaction reasoning LLM. The context-mining decoder integrates a vision-language model and transformer decoder to extract linguistic context information from visual features. The interaction reasoning LLM transfers the interaction recognition to a visual question-answering task, allowing the recognition of novel interaction from context information. Experiment results show that our proposed ContextHOI surpasses existing state-of-the-art methods on HICO-DET and SWIG-HOI datasets under zero-shot settings.

References

[1] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018.

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

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

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

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

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

[8] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2022.

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

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

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

[12] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6857–6866, 2018.

[13] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 843–851, 2019.

[14] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13617–13626, 2020.

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

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

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

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

[19] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1568–1576. IEEE, 2018.

[20] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 584–600. Springer, 2020.

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

[22] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10460–10469, 2020.

[23] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 495–504, 2021.

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

[25] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 314–332. Springer, 2020.

[26] Suchen Wang, Kim-Hui Yap, Henghui Ding, Jiyan Wu, Junsong Yuan, and Yap-Peng Tan. Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13475–13484, 2021.

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

[28] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.