A Review of Human-Object Interaction Detection

1st Yuxiao Wang
South China University of Technology
ftwangyuxiao@mail.scut.edu.cn

2nd Yu Lei
Southwest Jiaotong University
leiyu1117@my.swjtu.edu.cn

3rd Li Cui
Langfang Open University
463485938@qq.com

4th Weiying Xue
South China University of Technology
202320163283@mail.scut.edu.cn

5th Qi Liu
South China University of Technology
drliuqi@scut.edu.cn

6th Zhenao Wei*
South China University of Technology
wza@scut.edu.cn

2024 2nd International Conference on Computer, Vision and Intelligent Technology (ICCVIT) | 979-8-3315-4004-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCVIT63928.2024.10872548

***Abstract*—Human-object interaction (HOI) detection plays a key role in high-level visual understanding, facilitating a deep comprehension of human activities. Specifically, HOI detection aims to locate the humans and objects involved in interactions within images or videos and classify their specific interactions. The success of this task is influenced by several key factors, including the accurate localization of human and object instances and the correct classification of object categories and interaction relationships. This paper systematically summarizes and discusses the recent work in image-based HOI detection. First, the mainstream datasets involved in HOI relationship detection are introduced. Furthermore, starting with two-stage methods and end-to-end one-stage detection approaches, this paper comprehensively discusses the current developments in image-based HOI detection, analyzing the strengths and weaknesses of these two methods. Additionally, the advancements of zero-shot learning, weakly supervised learning, and the application of large-scale language models in HOI detection are discussed. Finally, the current challenges in HOI detection are outlined, and potential research directions and future trends are explored.**

***Index Terms*—human-object interaction, object detection, action recognition, deep learning**

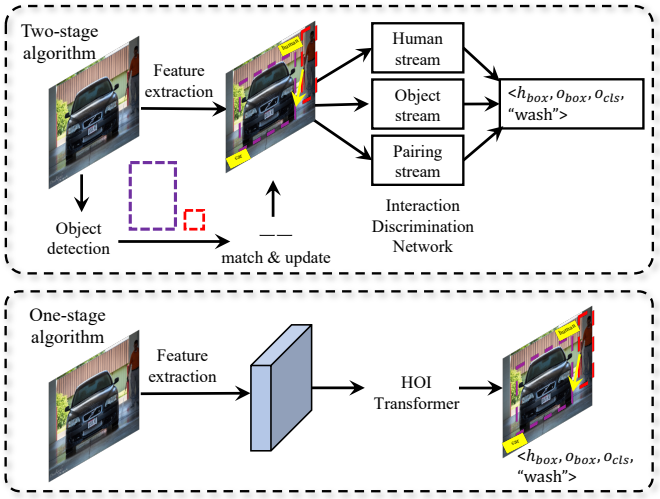

Fig. 1. The flowchart of the HOI detection algorithm.

I. Introduction

With the explosive growth of image data, understanding and analyzing the content within these images has become a crucial challenge. Relying solely on human vision for recognition is far from sufficient to meet the needs of modern society. Consequently, human-object interaction (HOI) detection has emerged as a key technology in the field of computer vision. HOI detection aims to accurately locate humans and objects in images or videos and recognize the corresponding interaction categories to better understand human activities. Specifically, HOI detection takes an image or video as input and outputs a series of triplets (< human, object, interaction >). It is widely used in autonomous driving, action recognition, human-computer interaction, social network analysis, emotion recognition, security monitoring, and video surveillance. This paper primarily reviews the research achievements in image-based HOI detection.

Existing HOI detection algorithms can be roughly divided into two categories: two-stage [1]–[4] and one-stage (end-to-end) [5]–[8] methods, as shown in Figure 1. The majority of two-stage methods are based on serial models, dividing the HOI task into two steps: human-object detection and interaction classification. Specifically, in the human-object detection stage, a pre-trained object detection framework [9] is typically used to identify all humans and objects present in the image. Then, in the second stage, the detected humans and objects are paired, and their corresponding features are passed to the interaction classification network to determine the type of interaction. Although two-stage methods are relatively simple to understand, they require iterating over all the information extracted by the detection network for each human-object pair, leading to significant computational overhead. Moreover, while phased training and optimization can enhance the performance of the task at hand, it also results in the loss of contextual information.

Unlike two-stage methods, one-stage approaches can directly output the HOI triplets. Specifically, one-stage methods usually introduce a new HOI mediator, allowing the network to predict interaction relationships directly. Compared to two-stage methods, introducing an interaction mediator eliminates the need for separately matching humans and objects, significantly enhancing inference speed. However, one-stage algorithms typically adopt a multi-task collaborative learning architecture, which can lead to interference between tasks.

In real-world scenarios, the variability of human actions and the wide range of object types result in diverse HOI types. Additionally, the high cost of annotation further limits the performance of models. To effectively address these problems, new techniques, including zero-shot learning [10], [11], weakly supervised [8], [12], and large-scale language models [12], [13] are discussed.

The remainder of this review is organized as follows. In Section II, the mainstream datasets are introduced. Section III summarizes the milestone HOI detection technology. In Section IV, the strengths and weaknesses of these different methods are summarized, and the future development directions are implemented in the actual architecture. Section VII explores future development directions.

II. Datasets

In the past years, many excellent HOI detection datasets have emerged. Based on annotation granularity, the levels are instance-level, part-level, and pixel-level. The summary and analysis of datasets across these three annotation levels are listed in Table I. Below are their detailed descriptions:

V-COCO [14]: V-COCO is a dataset based on COCO [15], containing 10,346 images, with 5,400 used for training and 4,946 for testing. It includes 80 object categories and 29 action categories, 4 of which represent body actions that do not involve any interaction. These interactions cover various actions such as "eating", "reading", and "wearing" associated with objects like "bread", "book", and "clothes".

HICO-DET [2]: In 2018, the University of Michigan introduced HICO-DET, a HOI detection dataset with a larger number of images and more complex interaction relationships. The data is sourced from Flickr and comprises a total of 47,776 images, with 38,118 used for training and 9,658 for testing. The dataset includes 80 object categories and 117 action categories, resulting in 600 different interaction types, of which 462 are non-rare categories and 138 are rare categories. The HICO-DET and V-COCO datasets cover a wide range of object and action categories, which are the most commonly used benchmarks for HOI detection.

HCVRD [16]: The University of Adelaide in Australia constructed a large-scale human-centered visual relationship detection dataset called HCVRD. Compared to previously released datasets, this dataset contains a large number of relationship annotations, with nearly 10,000 categories. This extensive label space more accurately reflects real-world HOIs. Unlike HICO-DET and V-COCO, HCVRD not only focuses on interaction relationships but also includes the relative positional relationships between humans and objects.

PaStaNet-HOI [16]: Under coarse-grained instance-level annotation supervision, the model is prone to overfitting, which leads to poor generalization capabilities. Shanghai Jiao Tong University constructed the PaStaNet-HOI dataset. This dataset provides approximately 110,000 annotated images, with 77,260 images used as the training set, 11,298 images as the validation set, and 22,156 images as the test set. PaStaNet-HOI discards the "no interaction" category. It consists of 116 interaction relationships and 80 object categories, forming a total of 520 HOI relationship categories.

PIC [17]: Liu et al. construct the pixel-level HOI database PIC (Person in Context) by performing pixel-level annotations on both the human body and objects. This enables more precise localization of humans and objects, even in cases of occlusion. PIC collects 17,122 human-centered images from the internet, which include a training set of 12,339 images, a validation set of 1,916 images, and a test set of 2,867 images. Additionally, as one of the most finely annotated databases for HOI detection. It includes a comprehensive range of annotation types, such as 141 object categories and 23 interaction relationships between humans and objects.

TABLE I
Dataset Statistics

Dataset	Total	Object	Relationship	Annotation
V-COCO	10346	80	29	instance-level
HICO-DET	47776	80	117	instance-level
HCVRD	52855	1824	927	instance-level
PaStaNet-HOI	110714	80	116	part-level
PIC	17122	141	23	pixel-level

III. The architectures of HOI Detection

Two-stage detection algorithms primarily focus on multi-stream models and graph models. One-stage detection algorithms can be divided into bounding box-based, relationship point-based, and query-based models. This section briefly reviews the above work and introduces new techniques such as zero-shot learning, weakly supervised, and large-scale language models.

A. Two-stage HOI detection architecture

Two-stage HOI detection is an instance-guided, bottom-up deep learning approach. Currently, two-stage methods are divided into multi-stream models and graph models.

Multi-stream models. The multi-stream model is an early attempt in the field of HOI detection. Firstly, the object detector generates region proposals about humans and objects. Secondly, an interaction relationship classification network extracts features for the targets and then fuses the classification results. In 2018, Chao et al. [2] construct a widely influential public dataset known as the HICO-A dataset. In addition, they propose a standard two-stage HOI network HO-RCNN for extracting features of spatial relationships between humans and objects. The HO-RCNN is composed of three key components: the human stream, the object stream, and the HOI stream. In work [1], Li et al utilize an interaction-aware network to learn general interaction knowledge from several HOI datasets. The non-interaction suppression strategy is employed before HOI classification inference, thereby improving network performance.

The subtle visual differences between various interaction relationships bring challenges to HOI detection. To address these challenges, Wan et al. [4] propose a multi-level relationship detection strategy that utilizes human pose cues to

capture interactions. Specifically, they employ a multi-branch network to learn pose-enhanced relationship representations across three semantic levels. By integrating these features, the model generates robust results. Zhong et al. [18] propose a new polysemy decoding network, PD-Net, which further mitigates the issue of verb polysemy through three strategies. In work [19], a novel network GTNet based on self-attention guidance, enhances detection performance by encoding spatial context information into the visual features of instances. To address the long-tail distribution problem in HOI, Hou et al. [20] design a novel Fabricated Compositional Learning framework.

Graph models. The aforementioned multi-stream models overlook the correlations between different human-object pairs. Additionally, processing each human-object pair individually increases time costs. To overcome these limitations, Qi et al. [3] first introduce the Graph Parsing Neural Network (GPNN). GPNN uses nodes and edges to identify instances and interaction relationships, respectively.

Previous graph-based algorithms treat humans and objects as the same type of node, failing to distinguish information exchanged between different entities. Therefore, Wang et al. [21] propose a heterogeneous graph network, CHGNet, which models humans and objects as distinct types of nodes. Gao et al. [22] utilize abstract spatial semantic representations to describe each human-object pair. They employ a dual relation graph to aggregate contextual information from the scene and capture discriminative cues, effectively addresses the prediction ambiguity issue. Most methods primarily focus on the visual and semantic features of instances. However, they do not leverage the high-level semantic relationships within the image. In order to solve this shortcoming, He et al. [23] embed scene graphs into global contextual cues. Additionally, a message passing module was developed to gather relational information from the neighbors of objects.

B. one-stage HOI detection architecture

One-stage algorithms have surpassed traditional two-stage models in both speed and accuracy. Unlike two-stage methods, one-stage approaches can directly output HOI triplets without an additional object detection process.

Bounding box-based models. The bounding box-based algorithms directly detect the location and category of targets using a simple structure while simultaneously predicting potential interaction relationships. This simplified design significantly enhances the inference speed of the algorithm. Previous works first detect instances and then predict interaction actions, which results in longer inference times for HOI detection. To address this challenge, Kim et al. [24] propose UnionDet, a method that directly captures interaction regions, eliminating the need for additional inference stages. Compared to traditional two-stage algorithms, UnionDet significantly improves inference speed by 4 to 14 times. This innovation makes HOI detection more efficient and real-time.

Traditional one-stage methods typically focus on the joint region of interaction, which can introduce unnecessary visual noises. To tackle this issue, Fang et al. [25] propose DIRV, which focuses on the interaction regions of each human-object pair at different scales and extracts the most relevant subtle visual features. Additionally, DIRV develops a voting strategy that leverages the overlapping parts within the interaction region, replacing the traditional non-maximum suppression method.

Relationship point-based models. Inspired by anchor-free detectors, research based on relationship points opens a new era in one-stage methods. Wang et al. [26] argue that extracting appearance features only is insufficient to handle complex HOI sciences. Therefore, they propose IP-Net, the first algorithm to view HOI detection as a keypoint detection problem. PPDM [5] is the first real-time HOI detection method that redefines the HOI triplet as <human point, interaction point, object point>. As a novel parallel architecture, PPDM significantly reduces computational costs by filtering of interaction points.

Existing one-stage models lack a reasoning step for dynamic discriminative cues. Zhong et al. further improve PPDM by GGNet [27]. GGNet first determines whether a pixel is an interaction point, then infers action-aware points around the pixel to refine the point's position.

Query-based models. Tamura et al. [28] propose a transformer-based feature extractor, where the attention mechanism and query-based detection play key roles. One-stage HOI detection algorithms based on the transformer architecture are gradually emerging and developing. Unlike existing transformer-based models that query at a single level, Dong et al. [7] explicitly merge and sum queries to better model the relationships between parts and the whole, which are not directly captured within the transformer. Kim et al. [24] use HOTR to predict triplets from images directly. This method effectively leverages the inherent semantic relationships within the image, eliminating the post-processing strategies used in previous approaches. Moreover, Liao et al. [29] propose a dual-branch GEN-VLKT, which eliminates the need for post-matching. CLIP [30] is also embedded to initialize the classifier, effectively leveraging image and text information. Similarly, the category-aware transformer network (CATN) [31] enhances detector performance by initializing object queries with category-aware semantic information.

C. New techniques

This section introduces new techniques, including zero-shot learning, weakly supervised, and large-scale language models.

Zero-shot learning. Maraghi et al. [10] is the first to utilize zero-shot learning methods to address the long-tail problem in HOI detection. In this work, a decomposed model is used to separately infer verbs and objects, enabling the detection of new verb-object combinations during the testing phase. Eum et al. [11] propose an HOI detection method based on verb-object relationship reasoning, where semantic and spatial information is embedded into the visual stream. Due to the combinatorial nature of visual relationships, collecting a sufficiently large amount of trainable triplet data is hard. To address this, Peyre

TABLE II
PERFORMANCE COMPARISONS ON HICO-DET DATASET. * REPRESENTS THE RESULTS GIVEN IN [12], [13].

Methods	Backbone	Source	Default (**mAP**↑)			Know Object (**mAP**↑)		
			Full	Rare	None-Rare	Full	Rare	None-Rare
Two-Stage Methods								
HO-RCNN	CaffeNet	WACV 2018	7.81	5.37	8.54	10.41	8.94	10.85
InteractNet	ResNet-50-FPN	CVPR 2018	9.94	7.16	10.77	-	-	-
GPNN	ResNet-101	ECCV 2018	13.11	9.34	14.23	-	-	-
iCAN	ResNet-50	BMCV 2018	14.84	10.45	16.15	16.26	11.33	17.73
No-Frills	ResNet-152	ICCV 2019	17.18	12.17	18.68	-	-	-
VSGNet	ResNet-152	CVPR2020	19.80	16.05	20.91	-	-	-
ACP	Res-DCN-152	ECCV 2020	20.59	15.92	21.98	-	-	-
PD-Net	ResNet-152-FPN	ECCV 2020	20.81	15.90	22.28	24.78	18.88	26.54
SG2HOI	ResNet-50	ICCV 2021	20.93	18.24	21.78	24.83	20.52	25.32
DJ-RN	ResNet-50	CVPR 2020	21.34	18.53	22.18	23.69	20.64	24.60
SCG	ResNet-50-FPN	ICCV 2021	21.85	18.11	22.97	-	-	-
IDN	ResNet-50	NIPS 2020	23.36	22.47	23.63	26.43	25.01	26.85
ATL	ResNet-50	CVPR 2021	23.81	17.43	25.72	27.38	22.09	28.96
UPT	ResNet-101-DC5	CVPR 2022	**32.62**	**28.62**	**33.81**	**36.08**	**31.41**	**37.47**
One-Stage Methods								
UnionDet	ResNet-50-FPN	ECCV 2020	17.58	11.72	19.33	19.76	14.68	21.27
IP-Net	Hourglass-104	CVPR 2020	19.56	12.79	21.58	22.05	15.77	23.92
HRNet	ResNet-152	TIP 2021	21.93	16.30	23.62	25.22	18.75	27.15
PPDM-Hourglass	Hourglass-104	CVPR 2020	21.94	13.97	24.32	24.81	17.09	27.12
HOI-Trans	ResNet-50	CVPR 2021	23.46	16.91	25.41	26.15	19.24	28.22
GG-Net	Hourglass-104	CVPR 2021	23.47	16.48	25.60	27.36	20.23	29.48
HOTR	ResNet-50	CVPR 2021	25.10	17.34	27.42	-	-	-
AS-Net	ResNet-50	CVPR 2021	28.87	24.25	30.25	31.74	27.07	33.14
QPIC	ResNet-50	CVPR 2021	29.07	21.85	31.23	31.68	24.14	33.93
FGAHOI	Swin-Tiny	TPAMI 2023	29.94	22.24	32.24	32.48	24.16	34.97
MSTR	ResNet-50	CVPR 2022	31.17	25.31	33.92	34.02	28.83	35.57
PR-Net	ResNet-50	arXiv 2023	31.17	25.66	32.82	-	-	-
CDN-S	ResNet-50	NIPS 2021	31.44	27.39	32.64	34.09	29.63	35.42
DisTr	ResNet-50	CVPR 2022	31.75	27.45	33.03	34.50	30.13	35.81
RCL	ResNet-50	CVPR 2023	32.87	28.67	34.12	35.52	30.88	36.45
GEN-VLKT	ResNet-50	CVPR 2022	33.75	29.25	35.10	36.78	32.75	37.99
HOICLIP	ResNet-50	CVPR 2023	34.59	31.12	35.74	37.61	34.47	38.54
PBLQG	ResNet-50	TIP 2023	31.64	26.23	33.25	34.61	30.16	35.93
SG2HOI	ResNet-50	TIP 2023	33.14	29.27	35.72	35.73	32.01	36.43
TED-Net	ResNet-50	TCSVT 2024	34.00	29.88	35.24	37.13	33.63	38.18
PViC	Swin-L	ICCV 2023	**44.32**	44.61	44.24	**47.81**	**48.38**	**47.64**
RLIPv2	Swin-L	ICCV 2023	43.23	**45.64**	**45.09**	-	-	-
Weakly+ Supervised Methods								
Explanation-HOI*	ResNeXt101	ECCV 2020	10.63	8.71	11.20	-	-	-
MAX-HOI	ResNet101	WACV 2021	16.14	12.06	17.50	-	-	-
Align-Former	ResNet-101	arXiv 2021	20.85	18.23	21.64	-	-	-
PPR-FCN*	ResNet-50	ICCV 2017	17.55	15.69	18.41	-	-	-
PGBL	ResNet-50	ICLR 2023	22.89	**22.41**	23.03	-	-	-
FreeA	ResNet-50	arXiv 2024	**24.57**	21.45	**25.51**	**26.52**	**23.64**	**27.38**
Weakly Supervised Methods								
SCG*	ResNet-50	ICCV 2021	7.05	-	-	-	-	-
VLHOI	ResNet-50	CVPR 2023	8.38	-	-	-	-	-
FreeA	ResNet-50	arXiv 2024	**16.96**	**16.26**	**17.17**	**18.89**	**18.11**	**19.12**

et al. [32] develop a model that successfully merges semantic and visual spaces.

Weakly supervised models. Most existing HOI detection models require fully annotated data for supervised training. However, obtaining complete data labels remains a challenging task. To effectively address this issue, weakly supervised HOI detection typically uses image-level interaction labels for training. For example, Peyre et al. [33] introduce a weakly supervised discriminative clustering model that learns relationships solely from image-level labels. They also propose a new and challenging dataset, UnRel, to accurately evaluate visual relationships. In work [34], Sarullo et al. model the relationships between actions and objects in the form of a graph. Align-Former [35] is equipped with an HOI alignment layer that generates pseudo aligned human-object pairs based on weakly supervised. Furthermore, Baldassarre et al. [36] attempt to leverage a multi instance learning framework to detect instances and then use image-level labels to supervise the interaction classifier.

Large-scale language models. With the help of large language models, Unal et al. [13] use only image-level labels to query possible interactions between human and object categories. The graph-based ProposalCLIP [37] addresses the limitations of CLIP cues by predicting the categories of objects without annotations, effectively enhancing performance. Recently, Wan et al. [12] develop a CLIP-guided

HOI representation that integrates prior knowledge at both the image level and the human-object pair level. This dual-layer framework is designed to more effectively utilize image-level information through a shared HOI knowledge base, thereby further enhancing the learning of interaction features.

IV. Complex problem of HOI detection

Table II and Table III present the results of different methods on the HICO-DET and V-COCO datasets, respectively. The advantage of two-stage methods is they can decouple object detection and interaction classification, allowing each stage to focus on optimizing its specific task. However, this approach encounters many obstacles. There are imbalanced positive and negative sample distributions, additional computational, and insufficient information exchange.

One-stage HOI detection algorithms significantly improve efficiency and accuracy by directly predicting interactions. However, these models typically use a multi-task learning approach to share features, which may lead to interference between tasks, preventing the model from achieving optimal performance.

Significant progress has been made in the field of HOI detection, especially with the emergence of new technologies. These methods can directly predict interactions, avoiding the complexity of the two-stage process. Nevertheless, despite these advancements, several issues remain in the field of HOI detection. Current models may perform poorly when handling multiple interaction types. Additionally, environmental conditions and lighting issues can further affect the detection of small objects.

Peering into the future, solutions to these issues could begin with improving the diversity and quality of datasets. Techniques such as transfer learning can also be utilized to enhance the generalizability of models. More effective multi-task learning strategies can be explored to better balance the relationships between different tasks and improve overall performance. Moreover, incorporating more contextual information, such as semantic or temporal data, could further enhance the accuracy of interaction detection.

TABLE III
Performance comparisons on V-COCO dataset.

Method	Source	AP1 (**mAP↑**)	AP2 (**mAP↑**)
Two-stage Methods			
InteractNet	CVPR 2018	40.0	-
GPNN	ECCV 2018	44.0	-
iCAN	BMCV 2018	45.3	52.4
TIN	CVPR 2019	47.8	54.2
VCL	ECCV 2020	48.3	-
DRG	ECCV 2020	51.0	-
IP-Net	CVPR 2020	51.0	-
VSGNet	CVPR 2020	51.8	57.0
PMFNet	ICCV 2019	52.0	-
PD-Net	ECCV 2020	52.6	-
FCMNet	ECCV 2020	53.1	-
ACP	ECCV 2020	53.2	-
IDN	NIPS 2020	53.3	60.3
UPT	CVPR 2022	**61.3**	**67.1**
One-stage Methods			
UnionDet	ECCV 2020	47.5	56.2
HOI-Trans	CVPR 2021	52.9	-
AS-Net	CVPR 2021	53.9	-
GG-Net	CVPR 2021	54.7	-
HOTR	CVPR 2021	55.2	64.4
QPIC	CVPR 2021	58.8	61.0
FGAHOI	TPAMI 2023	60.5	61.2
Iwin-L	ECCV 2022	60.9	-
PR-Net	arXiv 2023	61.4	-
CDN-S	NIPS 2021	61.7	63.8
GEN-VLKT	CVPR 2022	62.4	64.5
TED-Net	TCSVT 2024	63.4	65.0
PViC	ICCV 2023	64.1	70.2
RLIPv2	ICCV 2023	**72.1**	**74.1**
Weakly+ Supervised Methods			
Align-Former	arXiv 2021	15.8	16.3
PGBL	ICLR 2023	43.0	48.1
FreeA	arXiv 2024	**50.2**	**52.1**
Weakly Supervised Methods			
SCG*	ICCV 2021	20.1	-
VLHOI†	CVPR 2023	17.7	-
VLHOI	CVPR 2023	29.6	-
FreeA	arXiv 2024	**30.8**	**32.6**

V. Analysis of Current Datasets

While significant progress has been made in HOI detection, there remain substantial limitations in the current datasets and methods. One major challenge is the lack of diversity in existing datasets. Most datasets primarily consist of interactions between humans and a limited set of object categories, which may not be representative of real-world scenarios. Additionally, many datasets suffer from annotation inconsistencies and limited coverage of occlusions or complex interactions, which can hinder model performance.

VI. Future Directions in HOI

Looking forward, there are several promising directions for further research in HOI detection. One key area is the integration of multi-modal data, such as depth maps, thermal images, or even audio, which can provide complementary information that is valuable for detecting interactions in more complex environments. Another exciting direction is the development of more advanced weakly-supervised or semi-supervised learning techniques that reduce the reliance on large annotated datasets.

VII. Conclusion

HOI detection is a research hotspot in computer vision, with widespread applications in action recognition, pose estimation, autonomous driving, and other scenarios. This paper discusses the latest advancements in two-stage and one-stage HOI detection tasks, providing an overview of the cutting-edge developments. At the same time, new techniques including zero-shot learning, weakly supervised learning, and large-scale language methods are introduced. Finally, the current challenges faced by HOI detection, from the perspective of technical difficulties, are summarized, and its future development trends are predicted.

References

[1] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, and C. Lu, "Transferable interactiveness knowledge for human-object

interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3585–3594.

[2] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *2018 ieee winter conference on applications of computer vision (wacv)*. IEEE, 2018, pp. 381–389.

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

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

[5] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. 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*, 2020, pp. 482–490.

[6] D. Zhou, Z. Liu, J. Wang, L. Wang, T. Hu, E. Ding, and J. Wang, "Human-object interaction detection via disentangled transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 19 568–19 577.

[7] Q. Dong, Z. Tu, H. Liao, Y. Zhang, V. Mahadevan, and S. Soatto, "Visual relationship detection using part-and-sum transformers with composite queries," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3550–3559.

[8] Y. Wang, Z. Wei, X. Jiang, Y. Lei, W. Xue, J. Liu, and Q. Liu, "FreeA: Human-object interaction detection using free annotation labels," *arXiv preprint arXiv:2403.01840*, 2024.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[10] V. O. Maraghi and K. Faez, "Scaling human-object interaction recognition in the video through zero-shot learning," *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, p. 9922697, 2021.

[11] S. Eum and H. Kwon, "Semantics to space (s2s): Embedding semantics into spatial space for zero-shot verb-object query inferencing," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 1384–1391.

[12] B. Wan, Y. Liu, D. Zhou, T. Tuytelaars, and X. He, "Weakly-supervised hoi detection via prior-guided bi-level representation learning," *International Conference on Learning Representations*, 2023.

[13] M. E. Unal and A. Kovashka, "Vlms and llms can help detect human-object interactions with weak supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, url: https://asu-apg.github.io/odrum/posters_2023/poster_6.pdf*, 2023.

[14] S. Gupta and J. Malik, "Visual semantic role labeling," *arXiv preprint arXiv:1505.04474*, 2015.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[16] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. van den Hengel, "Hcvrd: A benchmark for large-scale human-centered visual relationship detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[17] S. Liu, Z. Wang, Y. Gao, L. Ren, Y. Liao, G. Ren, B. Li, and S. Yan, "Human-centric relation segmentation: Dataset and solution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4987–5001, 2021.

[18] X. Zhong, C. Ding, X. Qu, and D. Tao, "Polysemy deciphering network for human-object interaction detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 69–85.

[19] A. Iftekhar, S. Kumar, R. A. McEver, S. You, and B. Manjunath, "Gtnet: Guided transformer network for detecting human-object interactions," in *Pattern Recognition and Tracking XXXIV*, vol. 12527. SPIE, 2023, pp. 192–205.

[20] Z. Hou, B. Yu, Y. Qiao, X. Peng, and D. Tao, "Detecting human-object interaction via fabricated compositional learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 646–14 655.

[21] H. Wang, W.-s. Zheng, and L. Yingbiao, "Contextual heterogeneous graph network for human-object interaction detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 2020, pp. 248–264.

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

[23] T. He, L. Gao, J. Song, and Y.-F. Li, "Exploiting scene graphs for human-object interaction detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 984–15 993.

[24] B. Kim, T. Choi, J. Kang, and H. J. Kim, "Uniondet: Union-level detector towards real-time human-object interaction detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 498–514.

[25] H.-S. Fang, Y. Xie, D. Shao, and C. Lu, "Dirv: Dense interaction region voting for end-to-end human-object interaction detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1291–1299.

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

[27] X. Zhong, X. Qu, C. Ding, and D. Tao, "Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 234–13 243.

[28] M. Tamura, H. Ohashi, and T. Yoshinaga, "Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 410–10 419.

[29] Y. Liao, A. Zhang, M. Lu, Y. Wang, X. Li, and S. Liu, "Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 123–20 132.

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[31] L. Dong, Z. Li, K. Xu, Z. Zhang, L. Yan, S. Zhong, and X. Zou, "Category-aware transformer network for better human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 538–19 547.

[32] J. Peyre, I. Laptev, C. Schmid, and J. Sivic, "Detecting unseen visual relations using analogies," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1981–1990.

[33] J. Peyre, J. Sivic, I. Laptev, and C. Schmid, "Weakly-supervised learning of visual relations," in *Proceedings of the ieee international conference on computer vision*, 2017, pp. 5179–5188.

[34] A. Sarullo and T. Mu, "Zero-shot human-object interaction recognition via affordance graphs," *arXiv preprint arXiv:2009.01039*, 2020.

[35] M. Kilickaya and A. Smeulders, "Human-object interaction detection via weak supervision," *arXiv preprint arXiv:2112.00492*, 2021.

[36] F. Baldassarre, K. Smith, J. Sullivan, and H. Azizpour, "Explanation-based weakly-supervised learning of visual relations with graph networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*. Springer, 2020, pp. 612–630.

[37] H. Shi, M. Hayat, Y. Wu, and J. Cai, "Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9611–9620.