

A Survey of Human-Object Interaction Detection With Deep Learning

Geng Han^{ID}, Jiachen Zhao^{ID}, Lele Zhang^{ID}, Member, IEEE, and Fang Deng^{ID}, Senior Member, IEEE

Abstract—Human-object interaction (HOI) detection has attracted significant attention due to its wide applications, including human-robot interactions, security monitoring, automatic sports commentary, etc. HOI detection aims to detect humans, objects, and their interactions in a given image or video, so it needs a higher-level semantic understanding of the image than regular object recognition or detection tasks. It is also more challenging technically because of some unique difficulties, such as multi-object interactions, long-tail distribution of interaction categories, etc. Currently, deep learning methods have achieved great performance in HOI detection, but there are few reviews describing the recent advance of deep learning-based HOI detection. Moreover, the current stage-based category of HOI detection methods is causing confusion in community discussion and beginner learning. To fill this gap, this paper summarizes, categorizes, and compares methods using deep learning for HOI detection over the last nine years. Firstly, we summarize the pipeline of HOI detection methods. Then, we divide existing methods into three categories (two-stage, one-stage, and transformer-based), distinguish them in formulas and schematics, and qualitatively compare their advantages and disadvantages. After that, we review each category of methods in detail, focusing on HOI detection methods for images. Moreover, we explore the development process of using foundation models for HOI detection. We also quantitatively compare the performance of existing methods on public HOI datasets. At last, we point out the future research direction of HOI detection.

Index Terms—Deep learning, visual relationship detection, human-object interaction, foundation models, attention mechanism, GNN, transformer.

I. INTRODUCTION

DEEP learning methods have achieved brilliant achievements in object recognition [1], [2] and detection [3], [4], which greatly reduce manual labor in processing mass

Received 2 July 2024; accepted 11 November 2024. This work was supported in part by the Key Program of National Natural Science Foundation of China under Grant 61933002, in part by the National Science Fund for Distinguished Young Scholars of China under Grant 62025301, in part by the National Natural Science Foundation of China Basic Science Center Program under Grant 62088101, and in part by China Postdoctoral Science Foundation under Grant BX20220186 and Grant 2023M741965. (Geng Han and Jiachen Zhao are co-first authors.) (Corresponding author: Fang Deng.)

Geng Han, Lele Zhang, and Fang Deng are with the Department of Automation, Beijing Institute of Technology, Beijing 100081, China, and also with the Chongqing Innovation Center, Beijing Institute of Technology, Chongqing 401120, China (e-mail: hangeng5446@163.com; zhanglele@bit.edu.cn; dengfang@bit.edu.cn).

Jiachen Zhao is with the National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: zhao_jiachen@163.com).

Recommended for acceptance by H. Cai.

Digital Object Identifier 10.1109/TETCI.2024.3518613

2471-285X © 2024 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See <https://www.ieee.org/publications/rights/index.html> for more information.

	Object recognition	Object detection	HOI detection
Examples			
Object recognition	Woman + Donut	Woman + Center	Woman + Center
Object detection	Donut + Lower Center	Donut + Lower Center	+ Eat + Donut + Lower Center
HOI detection		Boy + Center , Girl + Center Left ... Soccer + Lower Center	Boy + Center + Kick + Soccer + Lower Center

Fig. 1. Comparison between object recognition, object detection, and HOI detection.

visual information. Object recognition aims to answer “What is in the image?” while object detection aims to answer “What and where is in the image?” However, an expected intelligent machine should have a complete semantics understanding of a scene. Towards this goal, human-object interaction (HOI) detection is proposed to answer “What are the people doing with what objects?”. Fig. 1 gives two examples to show the different goals between object recognition, object detection, and HOI detection. From which we can see, HOI detection can provide more human-centered information at the semantics level. Therefore, HOI detection has plenty of application potential in human-robot interactions [5], [6], security monitoring [7], [8], automatic sports commentary [9], [10], action simulation [11], [12], recognition [13], [14], etc. At the same time, HOI detection plays a crucial role in the embodied artificial intelligence (AI) system [15], [16], which thinks that human intelligence needs to be formed through interaction and iteration with actual scenes.

However, HOI is a challenging visual task since it not only suffers from the common difficulties of machine vision. It also has to face some unique challenges:

- 1) *Multi-object interactions*: In a complex interaction scene, there may be multiple people performing interactions at the same time, one person interacting with multiple objects, and one object interacting with multiple people, such as a crowded party that involves various objects and interactions. Even if all objects and humans could be detected, assembling them into reasonable HOI triplets is still challenging. Some work [17], [18], [19], [20] focuses on solving this issue.
- 2) *Long-tail distribution of interaction categories*: Sample imbalance is particularly serious in HOI datasets. In some

datasets, the number of instance samples between different categories is unbalanced. Specifically, the number of common interaction samples can be tens or even hundreds of times that of uncommon interactions. This leads to the unreliable accuracy of the model after training due to overfitting, underfitting, and other problems. Some work [21], [22], [23], [24], [25] is trying to solve this problem by zero-shot and few-shot learning.

- 3) *Visual distraction under real-world settings*: Visual distraction is serious in HOI detection. In order to detect HOI, the model must identify objects. However, there are many interference factors in the natural environment, including occlusion, deformation, lighting changes, background clutter, shooting perspective, etc. This makes HOI detection challenging.

Before reviewing the literature, we first present the purpose of HOI detection. HOI detection aims to detect humans, objects, and interactions (verbs) between each human-object pair in a given image or video. Technically, an HOI instance is generally represented by a $\langle \text{human}, \text{verb}, \text{object} \rangle$ triplet and denoted by $\langle h, v, o \rangle$. Solving HOI detection consists of three sub-problems. Firstly, detecting an object instance o , which needs to indicate its location in the image (bounding box) b_o and its category c_o , i.e., $o = \langle b_o, c_o \rangle$. Similarly, the second is detecting a human instance $h = \langle b_h, c_h \rangle$. The third is predicting the action verb v , which is a classification problem. Considering the details of each sub-problem, some papers also present an HOI instance by a quintuple $\langle b_h, c_h, b_o, c_o, v \rangle$, but they are essentially the same.

Different work organizes and solves these three sub-problems in different frameworks. Some papers naturally regard HOI detection as three divided sub-problems and solve them in two subsequent stages [26], [27], [28], [29], [30], [31], [32]. In the first stage, they detect the object and human instances by an off-the-shelf detector. In the second stage, they generate the HOI hypothesis for all the human-object pairs. Some papers [33], [34], [35], [36] solve HOI detection in one stage by using point or union-box to represent humans, objects, and the interactions between them. Some papers [37], [38], [39], [40], [41], [42] treat HOI detection as a set prediction problem, then directly match the predicted and ground-truth HOI triples based on the transformer structure.

Based on these differences, most existing research papers classify their methods into one of the following three categories: two-stage methods, one-stage methods, and end-to-end methods. However, there is still no review literature to clearly define and distinguish between the three categories. Some papers discuss the categorization in their related work section, but they have disagreements with each other. For example, Zhong et al. [35] simply divide HOI detection methods into one-stage and two-stage. However, Antoun et al. [43] illustrate four categories: two-stage, one-stage, end-to-end transformer-based, and two-stage methods using transformers. Chen et al. [40] claim their transformer-based method as the one-stage network, while Zou et al. [37] call their transformer-based method as the end-to-end method. Such a confusing category is inconvenient for beginners to learn and for researchers to discuss. Therefore, one of the main contributions of this paper is to clearly define and

differentiate current methods that use deep learning for HOI detection. Instead of making a fresh start, we dig into the root reasons why methods contain a different number of stages, then divide existing methods into two-stage, one-stage, and transformer-based methods. Among them, we group all methods that use transformers into one category to avoid overlapping categories of some methods, such as methods that belong to both two-stage and end-to-end. The new category does not violate the category of most existing work but significantly makes the category clear.

Compared with the currently published HOI detection review papers, our contributions can be summarized as follows:

- 1) We review more than 200 references related to HOI detection and 13 datasets from 2015 to 2024, and compare the advantages and disadvantages of HOI detection methods and datasets. Then we summarize the pipeline of all three classes of HOI detection methods and clearly distinguish them in formulas and schematics.
- 2) We analyze the impact of foundation models on HOI detection methods, which is not covered in the previous HOI field review.
- 3) Based on the analyzed papers, we reasonably deduce and explore future research directions, analyze the current problems and limitations of each research direction, and propose our suggestions to solve these problems.

There have been five reviews on HOI detection. Bergstrom et al. claim their paper [44] to be the first general survey in this field, which mainly focuses on multi-stream CNN architectures but pays little attention to other architectures. Sunaina et al. [45] simply divide existing methods into two categories: handcrafted representation-based approaches and deep learning-based approaches, which do not reflect the specific differences in structure and algorithm of such methods. Li et al. [46] divide existing methods into sequential methods and parallel methods, which is similar to the definition of two-stage methods and one-stage methods. However, they do not consider the transformer-based methods. Antoun et al. [43] classify existing methods into two-stage methods, one-stage methods, end-to-end transformer-based methods, and two-stage methods using transformers. This classification method has partial overlap and does not consider the foundation models on HOI detection. Wang et al. [47] divided the existing HOI detection methods into one-stage and two-stage methods, and further divided the two-stage methods into HOI detection with multi-stream modeling, human parts and pose, compositional learning, graph-based modeling, and query-based modeling. This classification method also overlaps and causes confusion. For example, RPNN [28] is a method with graph-based modeling, and it also incorporates detailed feature information of human body parts and poses to refine the graph structure. This makes it difficult to clearly classify RPNN into HOI detection with graph-based modeling or HOI detection with human parts and pose. At the same time, Wang et al. did not pay attention to the impact of emerging foundation models on HOI detection. Differently, the classification method of our paper is more scientific. Moreover, we have focused on the influence of the foundation model, providing a detailed and comprehensive description of the foundation model methods

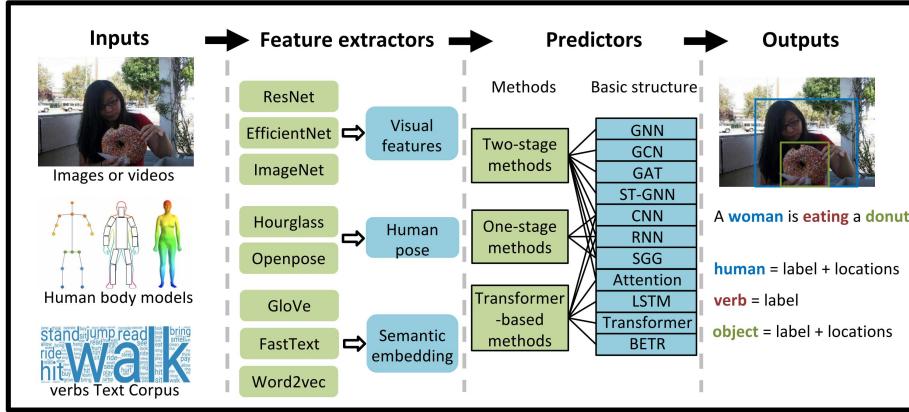


Fig. 2. Pipeline of HOI detection using deep learning methods.

and development directions. At the same time, the number of articles reviewed is far more than the papers mentioned above. We classify all transformer-based methods into one category instead of separating a two-stage method using transformers like Antoun et al. In conclusion, our review is more comprehensive and specific than existing reviews.

The rest of this paper is organized as follows. Section II introduces the HOI detection pipeline, classifies the existing HOI detection methods, defines them with formulas, and describes the advantages, disadvantages, and development process of each method. Section III defines the two-stage methods and further divides them into the attention-based methods and GNN-based methods; Section IV defines the one-stage methods and introduces the current advanced one-stage methods; Section V defines the transformer-based methods and introduces the current advanced transformer-based methods; Section VI introduces the definition of the foundation models and introduces methods using the foundation models to solve the long-tail distribution problem; Section VII introduces the datasets and evaluation metrics in the field of HOI in detail, and analyzes the performance of each method in the dataset; Sections VIII and IX present the future directions and conclusion, respectively.

II. OVERVIEW OF HOI DETECTION

In this section, we provide an overview of existing HOI detection, including their pipeline, category, and development time.

A. Pipeline of HOI Detection

Existing deep HOI detection methods follow a common four-step pipeline, as shown in Fig. 2. **Firstly**, the model takes an image as the main input. We aim to find out all the HOIs in the image. In addition to visual information, human body model has been used as prior knowledge to improve results (figures of human body models are cited from [48]). Text corpus has also been used as external clues to detect unseen objects or actions. **Secondly**, HOI detection methods utilize some off-the-shelf backbone networks to extract features from inputs. For example, ResNet [49], EfficientNet [50], ImageNet [51] are used to extract visual features; Hourglass [52], Openpose [53] are

used to estimate the human pose; GloVe [54], FastText [55], Word2Vec [56] are used to generate semantic embedding vectors of objects or verbs. Generally, these backbone networks have been pre-trained on large-scale datasets, and their weights are frozen during HOI detection training. An excellent pre-training method can affect the final detection accuracy [57]. **Thirdly**, the HOI predictor further learns HOI-specific features and then predicts the HOI triplets. The HOI predictor is the core of HOI algorithms, which could be based on various structures, such as CNN, LSTM, GCN, Transformer, etc. **Finally**, HOI detection model outputs the $\langle \text{human-verb-object} \rangle$ triplets existed in the image.

B. Category of HOI Detection

The current mainstream classification method for HOI detection is stage-based classification. However, we notice that after the emergence of transformer [58] in 2017, more and more researchers have begun to use transformer-based methods for HOI detection, and some of these papers are unsuitable for using stages for classification. Therefore, we separate the transformer-based methods into a separate category, which can intuitively reflect the progress and impact of transformers on traditional two-stage and one-stage methods.

Two-stage methods use the appearance of detected instances (either humans or objects) as cues to predict the interaction between them. Therefore, the two-stage methods generally consist of two sequential steps: instance detection and interaction classification. In the first stage, they use an object detector, such as a Faster RCNN in [59], to detect the human and object instances. The output of the first stage includes the labels, bounding box, and in-box features of the detected instances. In the second stage, they use features in the detected box to identify the interaction between each possible human-object pair. Note that the weights of the first-stage detector can be either fixed or updated during training.

Given an image \mathcal{I} , the computing process of a vanilla two-stage method can be formulated as (1)–(2):

$$\mathcal{H}, \mathcal{O} = f_{\theta_d}(\mathcal{I}), \quad (1)$$

$$\mathbf{v}_i = f_{\theta_v}(\mathbf{h}_i, \mathbf{o}_i, \mathbf{a}_i), \forall \mathbf{h}_i \in \mathcal{H}, \forall \mathbf{o}_i \in \mathcal{O}, \forall \mathbf{a}_i \in \mathcal{A}, \quad (2)$$

where f_{θ_d} is the object detector, f_{θ_v} is the interaction verb classifier. \mathcal{H} and \mathcal{O} are the human instance set and the object instance set, v_i is the predicted interaction verb between h_i and o_i . Note that auxiliary features have also been used in two-stage methods, therefore, we set auxiliary features a_i as the input for the interaction verb classifier and reflect it in (2). And \mathcal{A} is the set of auxiliary features, such as spatial configuration [26], [27], [60] of detected instances, human pose features [61], [62], and text [63], [64]. However, they all treat the detector output as an input of the interaction verb classifier.

One-stage methods aim to regress a region to represent the interaction. The interaction region could be a point [33], [34], dynamic points [35], a union box [36] or multi-scale boxes [65]. In other words, these methods simultaneously detect human instances, object instances, and some interaction areas or points, where the interaction areas are only used to predict interaction verbs. Therefore, they usually follow a parallel structure as (3),

$$\{\langle h, v, o \rangle_i\} = f_{\theta_d, \theta_v}(\mathcal{I}), \quad (3)$$

where $\{\langle h, v, o \rangle_i\}$ is the detected HOI triples set in the image \mathcal{I} . This parallel structure enables these methods to infer the HOI triples in one stage and be more efficient than most two-stage methods.

Transformer-based methods use trainable query vectors to represent HOI triplets [37], [38], [39]. Their basic architecture is a transformer encoder-decoder. The encoder uses an attention mechanism to extract features from the global image context. The decoder takes several learnable query vectors as input, and each query captures at most one interaction action of a human-object pair. Actually, these methods just extend the transformer-based detection model DETR [66] to capture HOI detection and treat HOI detection as a set prediction problem of matching the predicted and ground-truth HOI instances. Transformer-based methods can be formulated as (4).

$$\{\langle h, v, o \rangle_i\} = f_{\theta_t, Q}(\mathcal{I}), \quad (4)$$

where $\{\langle h, v, o \rangle_i\}$ is the HOI triples set, θ_t means the transformer weights, Q means the query vectors. Q is often treated as parameters of the transformer, but it represents the HOI triplet query vectors, so we list it separately in (4).

Fig. 3 compares the structural differences among the three categories of methods. Nearly all methods utilize a CNN backbone to extract high-level image features, so Fig. 3 does not show it. Two-stage methods generally follow a sequential structure, while one-stage methods follow a parallel structure. However, some two-stage methods [67], [68], [69] add a net stream to learn features from the global image, which leads to a hybrid structure. Therefore, we distinguish them by whether they predict an explicit region from the original image to represent the interaction verb or rely on the detected human-object pair to predict the interaction. The transformer-based methods use FFNs to predict HOI triplets based on the latent features extracted by a transformer.

The three types of methods have their own advantages and disadvantages:

- 1) Two-stage methods can effectively use the information in the picture, but they have the problem of low time

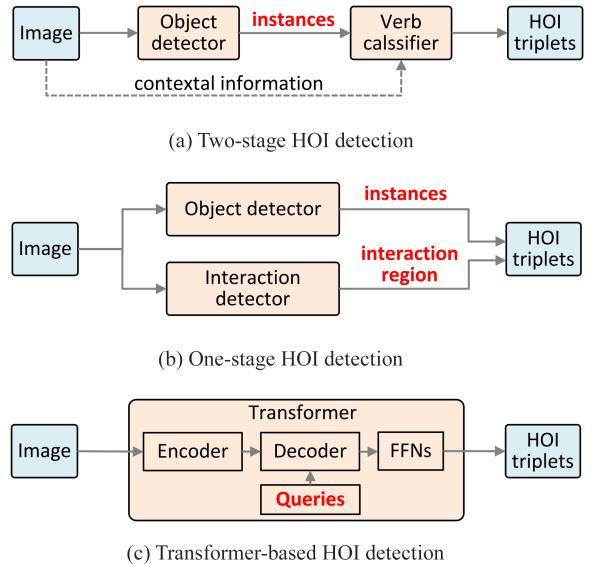


Fig. 3. Schematic comparison between different types of HOI detection methods.

efficiency and high computational complexity. They can effectively use the features of instance and additional information from context, body posture, and other aspects to carry out interaction classification. However, due to the separate architecture, detection algorithms require a lot of time to predict each possible human-object pair individually.

- 2) One-stage methods have better time efficiency, but they are difficult to quickly deploy to complex environments. They can detect HOI triples directly from the image due to the one-stage object detector [70], [71], [72] and parallel structure, so they have an obvious improvement in real-time performance. Although the efficiency has been greatly improved, the performance of the existing methods is limited by the complex handcrafted grouping strategies, which makes it difficult to quickly deploy it to complex or special real-world environments.
- 3) Transformer-based methods are currently the best method in terms of real-time performance and accuracy, but they lack interpretability. They can apply transformers with a good ability to capture remote dependencies. They can combine the advantages of one-stage methods and transformers to separately predict human-object proposals and interactions with multiple parallel decoders without requiring complex grouping strategies. This makes them the most studied method in the last two years. However, they lack interpretability and cannot easily configure parameters to fine-tune performance.

C. Development Timeline of HOI Detection

Fig. 4 is the development process of HOI detection methods. For clarity, Fig. 4 only parts but not all existing methods. The listed HOI methods meet the following two conditions: 1) solve a typical HOI detection problem; 2) can be clearly classified into

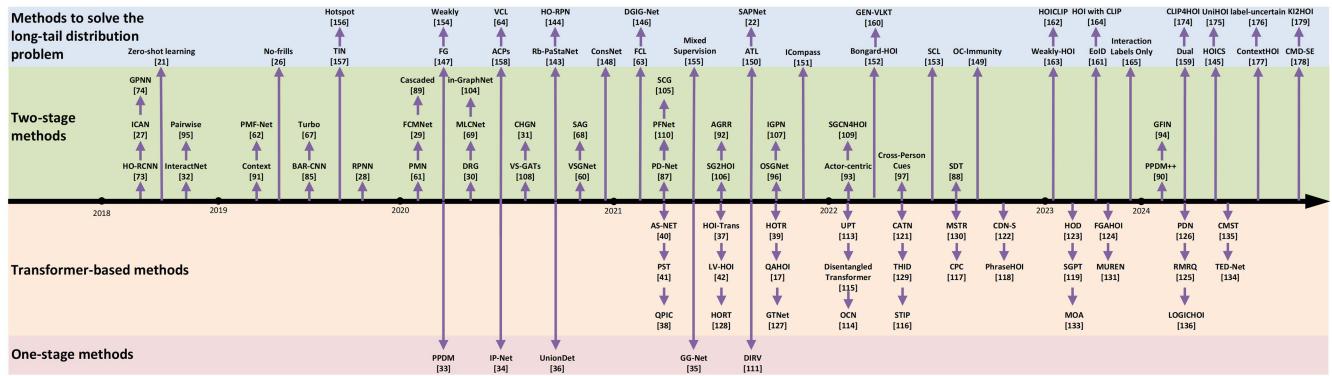


Fig. 4. Development timeline of HOI detection methods.

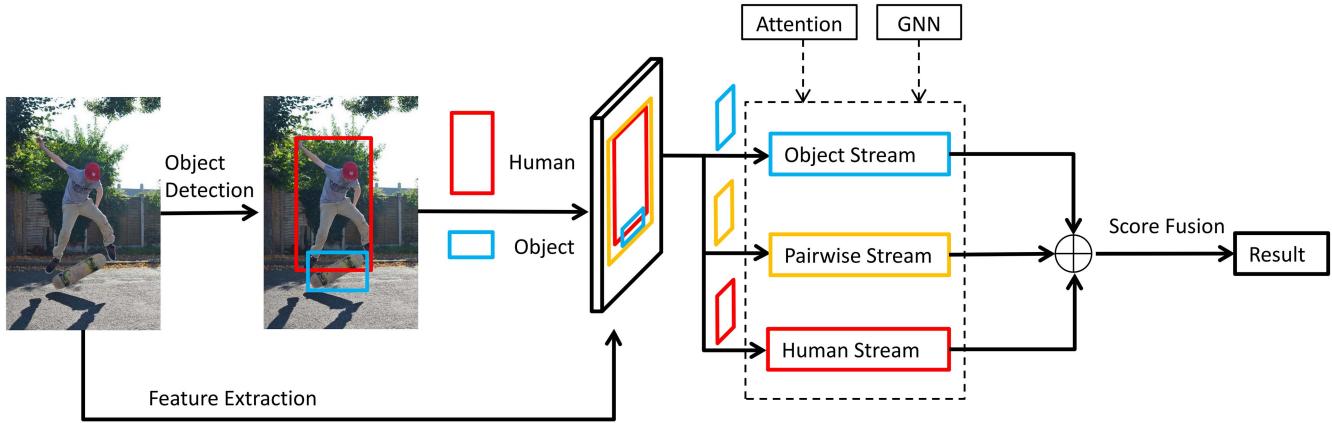


Fig. 5. Framework of two-stage methods.

one of the proposed three categories. From Fig. 4, we can find the following development trends:

- 1) The multi-stream architecture mainly used in the two-stage methods is proposed by HO-RCNN [73]. Later, attention mechanism [27] and graph neural network [74] are introduced into the field of HOI detection, further enriching the diversity of the two-stage methods and the detection accuracy.
- 2) The generation of the one-stage methods benefits from the development of the one-stage object detector. PPDM [33] is the first HOI detection method to achieve real-time performance, which has greatly improved the detection efficiency and accuracy compared with other methods in the same period.
- 3) For transformer-based methods, DETR [66] is one of the most advanced visual object detection methods. Most of the current transformer-based methods follow the structure of DETR and are further optimized based on DETR to improve the detection effect. At the same time, after the emergence of DETR, since it has greatly improved the real-time performance and accuracy of HOI detection, the two-stage methods and one-stage methods that do not use transformers have been greatly reduced, and many researchers have begun to pay attention to transformer-based methods.

- 4) After the large-scale visual language pre-training foundation model CLIP [75] is established, the mainstream research direction to solve the long-tail distribution problem began to shift to CLIP-based HOI detection.

III. TWO-STAGE METHODS

The timeline of two-stage methods is shown in Fig. 6. The human-object region-based convolutional neural networks (HO-RCNN) [73] proposed by Chao et al. in 2018 is known as the beginning of the two-stage methods. The multi-stream architecture proposed by them is of great significance for HOI detection research. Before the emergence of the transformer, almost all two-stage methods used multi-stream architecture as their structural framework. The multi-stream architecture includes human stream, object stream, and pairwise stream, as shown in Fig. 5. Among them, the human stream and object stream encode the appearance features of humans and objects, respectively, while the purpose of the pairwise stream is to encode the spatial relationship between humans and objects. The model first generates human-object region pair proposals using humans and object detectors. Then, it feeds each human-object pair proposal into a deep neural network (DNN) to generate HOI classification scores, and the scores in the three streams

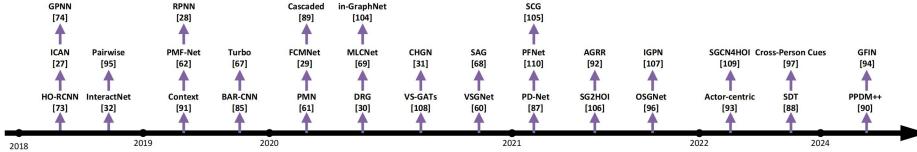


Fig. 6. Development timeline of two-stage methods.

are fused in a later fusion manner. Finally, the interaction classification is performed according to the fusion scores. Due to the outstanding performance of the attention mechanism and GNN in many aspects, such as feature extraction and object detection, researchers have begun to try to incorporate attention and GNN into the structural components of HOI detection, thereby improving detection accuracy.

The process of the two-stage methods divides the HOI detection task into two subtasks: instance detection and interaction classification. Generally, by sorting out relevant literature on HOI detection, we can divide the two-stage methods into two categories based on the model and overall framework of interaction classification: methods based on attention mechanism and methods based on graph neural networks (GNN).

403 A. Attention-Based Methods

404 The attention mechanism is a signal-processing mechanism
405 discovered by some scientists when they studied human vision
406 in the 1990s. At present, the attention mechanism has become
407 one of the most widely used “components” in the field of deep
408 learning, especially in the fields of natural language processing
409 (NLP) and computer vision (CV). Some prominent examples
410 include image classification [76], object detection [77], [78],
411 semantic segmentation [79], action recognition [80], [81], image
412 generation [82], pose estimation [83], and multi-modal task [84].
413 Since 2018, more and more researchers have integrated attention
414 mechanisms into the field of HOI detection.

415 Considering that some attention-based methods also incorporate
416 contextual information and body gestures, the attention-based
417 methods can be further subdivided into generalized
418 attention-based methods, attention-based methods incorporating
419 contextual information, and attention-based methods incorporating
420 body gestures. Among them, the generalized attention-based
421 methods are methods that do not deliberately focus on the
422 context information and body postures.

423 *1) The Generalized Attention-Based Methods:* For the generalized
424 attention-based methods, the most representative methods
425 include [27], [32], [85], [86], [87], [88], [89]. An instance-
426 centric attention module (ICAN) [27] proposed by Gao et al.
427 in 2018 highlights the important role of attention mechanism in
428 HOI detection for the first time. The module can dynamically
429 highlight important regions by learning the appearance of each
430 instance, which greatly improves the detection effect. Gkioxari
431 et al. [32] proposed a human-centric model named InteractNet
432 in 2018, which can classify actions at target object locations and
433 probability density estimates for specific actions. Box attention
434 R-CNN (BAR-CNN) [85] proposed by Kolesnikov et al. in 2019
435 is a joint probabilistic model for detecting visual relations. The

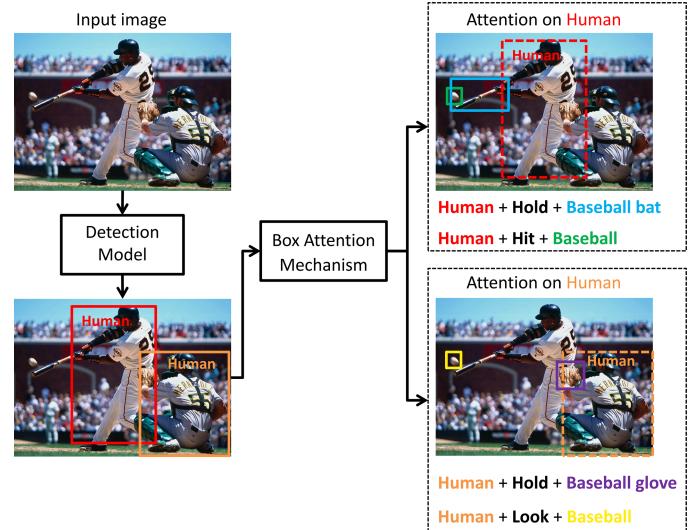


Fig. 7. Framework of BAR-CNN [85].

436 framework of BAR-CNN is shown in Fig. 7. Kolesnikov et al.
437 use the chain rule to decompose the probabilistic model into
438 two simpler models without using new hyperparameters. Zhong
439 et al. [86] proposed a novel Polysemy Deciphering Network
440 (PD-Net) in 2020. It proposes three methods for decoding HOI
441 to detect visual polysemy of related verbs. And they further
442 augmented PD-Net [87] in 2021. It is also the first attention
443 module that effectively uses language to address verb polysemy
444 in HOI detection. To optimize the performance on distant inter-
445 actions, SDT [88] proposes a spatially differentiated transformer
446 method and a Far-Near Distance Attention (FNDA) to effectively
447 model distant interactions. Considering the inherent complexity
448 of single-stage inference pipelines, Zhou et al. [89] proposed a
449 multi-stage cascaded architecture. At each stage, an instance lo-
450 calization network refines the HOI proposals and feeds them into
451 an interaction recognition network. The proposal of the PPDM
452 method represents the emergence of the one-stage methods, and
453 the specific content of PPDM will be elaborated in Section IV.
454 However, the author team believes that using only one interaction
455 point for its interaction feature representation is insufficient, so
456 they proposed a new two-stage method called PPDM++ [90].
457 PPDM++ breaks the limitation from redundant non-interactive
458 human-object proposals by directly locating interactive human-
459 object pairs in the first stage.

460 By focusing on areas closely related to HOI, the model can
461 capture key information more effectively and ignore irrelevant
462 background noise, allowing the model to better understand the
463 spatial relationship between humans and objects. Thus, the inter-
464 action between them can be judged more accurately. At the same

time, under limited computing resources, the model is allowed to allocate more computing resources to important information areas, which means that the model can use its processing power more efficiently.

In summary, the attention mechanism has significantly helped HOI detection by improving feature extraction capabilities, enhancing information processing capabilities, improving reasoning and judgment capabilities, optimizing model performance, and reducing background interference. These advantages make the attention mechanism an important research direction in HOI detection.

2) The Attention-Based Methods Incorporating Contextual Information: For the attention-based methods incorporating contextual information, the most representative methods include [29], [91], [92], [93]. Based on ICAN, Wang et al. [91] proposed a contextual attention framework for HOI detection in 2019. Unlike ICAN, which uses standard appearance features to build attention maps, their method builds attention maps by learning context-aware appearance features of instances. The context-aware appearance features are generated by the context-aware module and encoded by appearance and context information. The attention module can highlight regions that may contain HOI by selecting relevant contextual information. Liu et al. [29] proposed a Fine-grained layout-Context-Motion Network (FCMNet) for localizing and recognizing all HOI instances in an image in 2020. The main role of the model is to amplify and utilize key clues in the image. Considering the huge combination space and non-interactive pair domination problem of multi-stream information, Lin et al. [92] proposed an action-oriented attention mining and relational reasoning network (AGRR). They exploit the compatible consistency between the contexts of human-object pairs to filter non-interactive combinations for relational reasoning. In 2022, Xu et al. [93] proposed an actor-centric HOI detection framework to explore the relationship between one person and multiple objects to solve the ambiguity problem that there may be multiple interactions in images. The Global Context and Pairwise-level Fusion Features Integration Network (GFIN) [94] proposed by Wang et al. in 2024 utilizes an encoder to achieve global context memory extraction. Then, it uses a decoder to transform the query sequences representing human-object pairs into recognizable interactive features.

By incorporating contextual information, the model can better understand the scene, thereby optimizing the inference process. Contextual information also provides more clues about interaction relationships, helping the model identify interaction types more accurately. At the same time, contextual information can also enhance the generalization ability and robustness of the model, helping the model handle more diverse scenes and interaction types so that it can still accurately identify HOI even when the background noise is large or the target object is partially occluded. In summary, the attention-based methods incorporating contextual information have significantly helped HOI detection by improving detection accuracy, enhancing model generalization ability and robustness, and optimizing the reasoning process.

3) The Attention-Based Methods Incorporating Body Gestures: Existing research shows that only relying on the appearance features of humans and objects and the spatial relationship between them is far from achieving the needs of HOI detection. Therefore, some studies have begun introducing additional information to improve the accuracy of HOI detection, among which the body part and pose of a human are essential information. Traditional approaches treat the human body as a whole and pay equal attention to the entire body area. Still, they ignore that humans typically use only certain body parts to interact with objects.

For the attention-based methods incorporating body gestures, the most representative methods include [61], [62], [67], [95], [96], [97]. Fang et al. [95] believe that different body parts should be given different attention, and the correlation between different body parts should also be further considered, so they proposed a new pairwise body part attention model in 2018. This model is the first to apply an attention mechanism to body part correlation to detect HOI. Considering Fang et al. [95] only use the human body pose as the spatial constraint between the human body part and the object, Wan et al. [62] proposed a pose-aware multi-level feature network (PMFNet) in 2019, which uses the information of human posture to capture the global spatial configuration of relationships. Feng et al. [67] proposed a turbo-learning framework in 2019. They design an HOI-guided pose estimation module and a pose-aware HOI recognition module to enforce information transfer between tasks. Liang et al. [61] proposed a pose-based modular network (PMN) in 2020. It consists of a branch that independently processes the relative spatial pose features of each joint and another that uses graph convolution to update the absolute value of each joint. Lee et al. [96] proposed an on-the-fly stacked generalization deep neural network (OSGNet). To achieve dynamic stacking generalization, they train the submodel and meta-learner simultaneously. The former can provide complementary information, and the latter improves the generalization performance of unseen test data. Wu et al. [97] learn about body part interaction from a global perspective. It constructs a saliency map of body parts based on self-attention to mine cross-human information clues and learn the overall relationship between all body parts.

Body posture information can be combined with visual features to represent richer features. This combination can make up for the shortcomings of a single visual feature in describing interactive relationships and provide more comprehensive information. At the same time, body posture information can reflect the movements and postures of the human body, which is crucial for understanding HOI. By paying attention to body posture, the model can better understand the details and process of interaction and reduce misjudgments and missed detections. Body posture information is also stable and robust, providing valuable clues even when the target is occluded or the background is complex. In summary, the attention-based methods incorporating body gestures have significant advantages in feature representation, information richness, interactive recognition capabilities, and robustness.

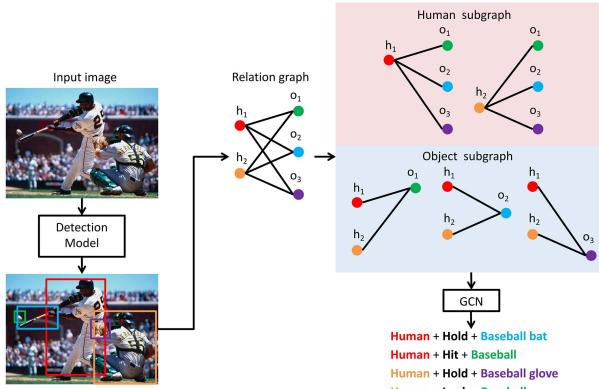


Fig. 8. Framework of DRG [30].

B. GNN-Based Methods

There has been growing interest in extending deep learning methods to graphs in recent years. Driven by the success of multiple factors, researchers have borrowed the ideas of convolutional networks, recurrent networks, and deep auto-encoders to define and design neural network structures for processing graph data. Thus, the graph neural network is born. The GNN mainly includes graph convolution networks (GCN), graph attention networks (GAT) [98], graph autoencoders, graph generative networks (GGN), and graph Spatio-temporal networks (ST-GNN) [99], [100]. One of the most extensive application areas of graph neural networks is computer vision, such as scene understanding [101], object detection and parsing [102], and visual question answering (VQA) [103].

Considering that some GNN-based methods also incorporate attention mechanisms and body gestures, the GNN-based methods can be further subdivided into generalized GNN-based methods, GAT-based methods, and GNN-based methods incorporating body gestures. Among them, the generalized GNN-based methods are methods that do not deliberately focus on attention mechanisms and body postures.

1) The Generalized GNN-Based Methods: For the generalized GNN-based methods, the most representative methods include [30], [74], [104], [105], [106], [107]. Qi et al. [74] first integrate graph models and neural networks to achieve HOI recognition, and they proposed a graph parsing neural network (GPNN), which is a generalization of message passing neural network (MPNN). The network inherits the learning ability of the neural network and the representation ability of the graph model. Unlike most previous graph or structured deep learning network models using pre-fixed graph structures, in order to seek better generalization ability, GPNN introduces an important connection function to solve the problem of graph structure learning. It learns to infer adjacency matrices in an end-to-end fashion and thus can infer analytic graphs that explicitly explain HOI relationships. In 2020, Gao et al. [30] first use abstract spatial semantic representation when describing each human-object pair and then utilize a dual relational graph (DRG) to aggregate contextual information of the scene. The framework of DRG is shown in Fig. 8. To exploit the implicit interaction semantics between visual objects, Yang et al. [104]

proposed a new graph-based interactive inference model (in-GraphNet) to infer HOIs. In 2021, to solve the problem existing in GNN methods, Zhang et al. [105] explored various methods of applying spatial conditioning under a multi-branch structure and demonstrated the advantages of spatial conditioning. Using scene graphs, He et al. [106] proposed HOI detection by exploiting the semantic relationships present in images (SG2HOI). Wang et al. [107] proposed a new model to learn interactive knowledge in Interaction Proposal Graph Network (IPGN).

GNN models HOI through a graph structure, which can not only effectively construct complex dependencies and capture high-order interaction information between nodes, but also naturally integrate visual features, semantic information, spatial relationships, and other information into the graph structure for unified modeling. GNN also makes the reasoning process of the model more intuitive and explainable. In HOI detection, this helps researchers better understand the decision-making process of the model, so as to make targeted improvements and optimizations.

In summary, GNN provides powerful help for HOI detection through its unique graph structure modeling capabilities, multi-stream information processing capabilities, high-order interaction information capture capabilities, and interpretability.

2) The GAT-Based Methods: GPNN's method of representing humans and objects with the same nodes is imperfect because humans and objects play different roles in HOI. There is an inter-class context between heterogeneous entities (human and object) in the activity scene. However, there is an intra-class context between homogeneous entities (human and human, object and object), meaning their relationship is not entirely the same.

For the GAT-based methods, the most representative methods include [31], [60], [68], [108], [109]. Given the above consideration, Wang et al. [31] proposed a context-heterogeneous graph network (CHGN) in 2020. Wang et al. combined context learning with graph attention methods to improve the effectiveness of nodes in gathering knowledge from their neighbors. By investigating HOI detection, Liang et al. [108] constructed a Visual-Semantic Graph Attention Network (VS-GATs), a dual-graph attention network aggregating visual-spatial and semantic information in parallel. The Visual-Spatial Graph Network (VS-GNet) [60] proposed by Ulutan et al. improves on the traditional three-branch network by exploiting the spatial configuration of human-object pairs to refine visual features and adding a graph convolution branch. However, Zhang et al. [68] test VSGNet and find that its bipartite graph performed worse when more than one message passing iteration is used. They then speculate that this is because adjacency values are not correctly normalized, resulting in node encoding being dominated by incoming messages. In their proposed spatiotemporal attention graph neural network (SAG) for detecting HOI, the message-passing algorithm is more stable and does not show this problem. SGCN4HOI [109] is a skeleton perceptual GCN for HOI detection. Based on VSGNet, it uses the spatial connection between human key points and object key points to capture their fine-grained structural interaction through graph convolution.

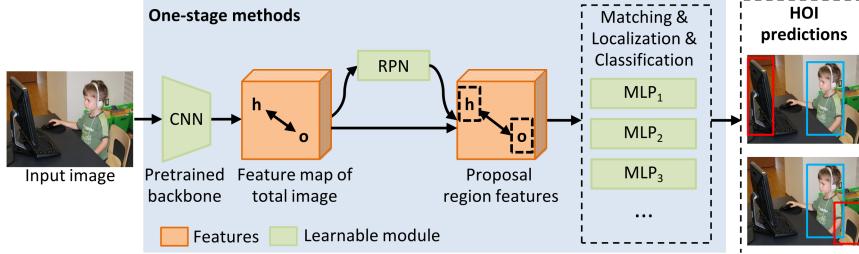


Fig. 9. Framework of one-stage methods.

The GAT enables the model to focus on the key information most relevant to HOI detection when processing graphic data and reduce the interference of irrelevant information, such as specific parts or interactions between the human body and objects. The GAT can also pay more attention to nodes and edges related to HOI during the inference process, thereby accelerating the inference process and reducing the use of computing resources. In summary, the GAT-based methods have significant advantages in focusing on key information, optimizing model reasoning, and reducing computing resource consumption.

3) The GNN-Based Methods Incorporating Body Gestures: For the GNN-based methods incorporating body gestures, the most representative methods include [28], [69], [110]. Zhou and Chi pioneered the combination of graph models with body parts and proposed relation parsing neural network (RPNN) [28] in 2019. RPNN is also the first study to focus on pairwise correlations between body parts and objects in HOI detection. RPNN introduces detailed body part features, and the model incorporates a graph structure for feature refinement instead of coarse appearance-based features in GPNN to extend learnable graph models for robust representations. Unlike ICAN, RPNN identifies objects and body parts as the most interesting areas to pay attention to. With further in-depth research, some problems with RPNN surface. Sun et al. [69] find that RPNN uses a backbone of convolutional neural networks (CNN) pre-trained on object detection datasets to extract visual features for HOI inference. This leads to significant deviations in the appearance distributions of interaction phrases and individual objects. Therefore, Sun et al. proposed a multi-level conditioned network (MLCNet) in 2020, which aims to fuse additional explicit knowledge with multi-level visual features. Liu et al. [110] also argue that the pairwise features of RPNN are not comprehensive enough, resulting in a better model of subtle interactions between body parts and objects. The multi-level pairwise feature network (PFNet) they proposed in 2021 contains more comprehensive pairwise features.

Body posture information provides a detailed description of human body movements and postures. Combined with GNN, it can provide a richer feature representation for HOI detection. This feature representation not only includes traditional visual features, but also includes relative position and angle information between various parts of the human body, which helps to understand HOI more accurately. At the same time, the GNN-based methods incorporating body gestures can more intuitively display the interactive relationship between various

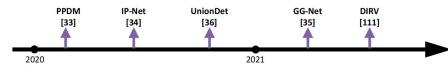


Fig. 10. Development timeline of one-stage methods.

parts of the human body and objects during the reasoning process, which makes the decision-making process of the model more transparent and explainable. In summary, the GNN-based methods incorporating body gestures have significant advantages in terms of richness of feature representation, optimization of model reasoning, and interpretability.

IV. ONE-STAGE METHODS

The common framework of the one-stage methods is shown in Fig. 9 and the timeline of one-stage methods is shown in Fig. 10. Due to the need to detect humans and objects first and then perform matching and interaction classification, the two-stage methods consume more computing resources and lack flexibility. At the same time, affected by its serial structure, the detection efficiency and real-time performance of the two-stage methods are also flawed. With the development of one-stage object detectors [70], [71], [72], scholars have studied more and more one-stage methods. As mentioned above, one-stage methods can reduce HOI detection as a parallel detection problem, detecting HOI triples directly from images. Therefore, compared with the two-stage methods, they have further improved detection efficiency, real-time performance, and accuracy.

In 2020, Liao et al. [33] proposed the first real-time one-stage HOI detection method, named parallel point detection and matching (PPDM). This method uses the center point of the detection frame to represent the human and object points and the midpoint between the human and object points to describe the interaction point. The model uses two parallel branches for point detection and matching, respectively. The point detection branch predicts humans, objects, and interaction points, and the point matching branch predicts two displacements from the human and object points to their corresponding interaction points. Human and object points originating from the same interaction point are treated as matched pairs. But Liao et al. also discovered the limitations of PPDM in interactive feature representation, so four years later, they proposed a two-stage method PPDM++ [90], which uses PPDM to detect HOI, and then extracts the regional features of each pair to predict

757 actions. More information has been elaborated on in the previous
 758 Section III.

759 Wang et al. [34] also use the idea of points to solve the
 760 problem of HOI detection in 2020, called IP-Net. By defining the
 761 interaction between humans and objects as interaction points,
 762 HOI detection is regarded as an interaction point estimation
 763 problem. IP-Net is the first method to treat HOI detection as
 764 a key point detection and grouping problem. The method learns
 765 to generate interaction vectors about the center points of humans
 766 and objects based on the interaction points.

767 Also, as a point-based method, Zhong et al. [35] proposed
 768 a novel one-stage method called Glance and Gaze Network
 769 (GGNet) in 2021. It models a set of action-aware points (Act-
 770 Points) through Glance and Gaze steps. The Glance step can
 771 quickly determine the distribution of interaction points in the
 772 feature map. The gaze step uses the feature map generated by
 773 the glance step to adaptively infer the ActPoint around each
 774 interaction point. Zhong et al. then devise an action-sensing
 775 approach to efficiently match each detected interaction with
 776 the relevant human-object pair and aggregate the features of
 777 ActPoints for interaction classification.

778 Different from the first three point-based methods, the union-
 779 level detector towards real-time HOI detection (UnionDet) [36]
 780 is proposed by Kim et al. in 2020. They regard HOI detection as
 781 a union box detection process. The obtained feature pyramid is
 782 sent to both the union branch and the instance branch. While the
 783 joint branch directly captures interaction regions, the instance
 784 branch performs traditional object detection and action classi-
 785 fication for more fine-grained HOI detection results. It directly
 786 detects interacting human-object pairs with the proposed joint
 787 detection framework instead of putting each object pair into a
 788 separate neural network. This eliminates the need for extensive
 789 reasoning after object detection and allows interactions to be
 790 detected with minimal extra time.

791 Fang et al. [111] proposed a region-based HOI detection
 792 method named DIRV in 2021. This method aims to solve the
 793 problem that the existing one-stage methods will introduce un-
 794 necessary visual disturbance information. Therefore, to capture
 795 the subtle visual features that are most important for interaction,
 796 it focuses more on densely sampled interaction regions at differ-
 797 ent scales for each human-object pair than previous one-stage
 798 methods. And it introduces a new voting strategy to replace the
 799 traditional Non-Maximum Suppression (NMS).

800 In summary, one-stage methods usually predict the input
 801 image directly and output all possible HOI triplets. Therefore,
 802 one-stage methods usually have higher detection speed and are
 803 more suitable for real-time applications. At the same time, the
 804 processing flow of one-stage methods is relatively simple, reduc-
 805 ing the need for intermediate steps and parameter adjustments.
 806 This makes one-stage methods easier to implement and deploy.

812 and position-wise feed-forward layers. Since its self-attention
 813 layer uses a multi-head attention mechanism to scan through all
 814 elements in the input sequence, the transformer can jointly attend
 815 to information from the whole sequence. The transformer's
 816 advantages of global computing and perfect memory make it
 817 popular in the machine-learning community. We recommend
 818 readers refer to paper [58], [66] for details about the self-
 819 attention mechanism and transformer model.

820 DETR [66] is one of the most advanced transformer-based
 821 visual object detection methods, which views object detection
 822 as a direct set prediction problem. It consists of a transformer
 823 to generate a set of object predictions and a set-based loss that
 824 forces correct matching between predictions and the ground-
 825 truth objects. The main advantage of DETR is eliminating the
 826 need for many hand-designed components. The tremendous
 827 success of DETR has led researchers to investigate its adap-
 828 tation to HOI detection. For example, HQM [112] can enhance
 829 DETR's robustness and improve detection accuracy by mining
 830 hard-positive queries.

831 At present, most transformer-based HOI methods follow
 832 DETR's pipeline. Their common framework is shown in Fig. 11,
 833 which first uses a CNN backbone to extract preliminary visual
 834 features from the input image. Then, the 2D feature map is
 835 supplemented with a positional encoding and flattened into a 1D
 836 sequence. After that, a transformer encoder-decoder architecture
 837 further extracts HOI-specific features, where the decoder addi-
 838 tionally takes a fixed number of learnable HOI queries as input.
 839 Finally, the subsequent feed-forward networks (FFN) process
 840 the output of the decoder to generate n HOI prediction results.
 841 The timeline of transformer-based methods is shown in Fig. 12.

842 *Some papers adjust the output heads of DETR to predict
 843 HOI triplets:* HOI-Trans [37] is one of the most straightfor-
 844 ward methods based on the transformer for HOI detection. The
 845 network architecture of HOI-Trans contains three main parts:
 846 a ResNet backbone, a standard transformer encoder-decoder,
 847 and five FFNs to predict HOI instances. HOI-Trans defines each
 848 HOI instance with a quintuple of \langle human class, interaction class,
 849 object class, human box, object box \rangle and uses five parallel FFNs
 850 to predict them separately. QPIC [38] is another intuitive DETR-
 851 like method whose idea and proposed time are similar. The
 852 main difference between QPIC and HOI-Trans lies in that QPIC
 853 represents a HOI instance with a quadruple \langle human bounding
 854 box, object bounding box, object class, action class \rangle . Zhang
 855 et al. [113] find that their two-stage counterparts can improve
 856 performance and memory efficiency while requiring a fraction
 857 of the time to train when equipped with the same transformer.
 858 Therefore, they proposed a two-stage HOI detector consisting
 859 of unary encoding and pairwise encoding (UPT) and further
 860 demonstrated that the two encoder layers have complementary
 861 properties.

862 *Some papers improve the standard transformer encoder-
 863 decoder architecture to an HOI-dedicated architecture:*
 864 HOTR [39] realizes it by introducing a shared encoder and two
 865 parallel decoders (instance decoder and interaction decoder).
 866 Similarly, AS-Net [40] also builds an instance decoder branch
 867 and an interaction decoder branch in parallel following a shared
 868 transformer encoder. To overcome the problem of static semantic

807 V. TRANSFORMER-BASED METHODS

808 The transformer is first proposed in [58] as a novel attention
 809 model for NLP. It follows an encoder-decoder structure, which
 810 takes a sequence as input and generates an output sequence.
 811 Either encoder or decoder contains stacked self-attention layers

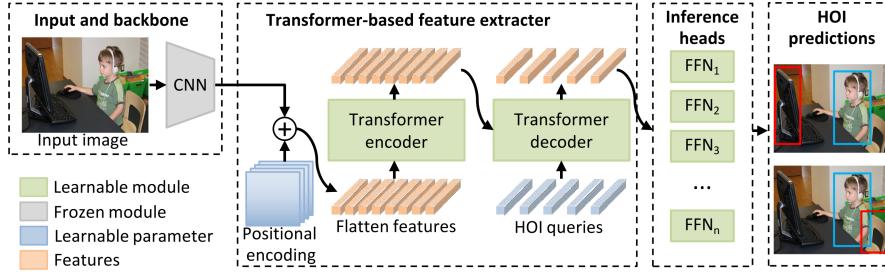


Fig. 11. Common framework of transformer-based HOI detection methods.

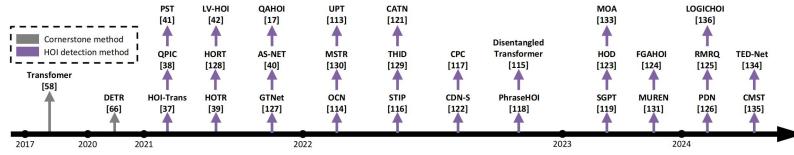


Fig. 12. Development timeline of transformer-based methods.

embeddings, Object-guided Cross-modal Calibration Network (OCN) [114] proposes to generate cross-modal perception visual and semantic features through cross-modal calibration (CMC). Zhang et al. [115] proposed a disentanglement strategy for HOI detection by decoupling triple prediction into human-object pair detection and interaction classification via instance flow and interaction flow. STIP [116] decomposes the process of HOI detection into two stages, i.e., interactive proposal generation is first performed, and then non-parametric interactive proposals are transformed into HOI predictions through a structure-aware transformer, which optimizes and strengthens the architecture. CPC [117] optimizes the interactive reasoning path. It improves the HOI detection of transformers by exploiting the enhanced decoding path. PhraseHOI [118] is the first method to creatively use HOI annotations as phrases. It proposes a novel relational phrase learning task based on HOI, which uses natural language phrases to describe interactions between humans and objects. Chen et al. [119] improved on the existing mainstream single-stream detection and proposed SGPT to use the secondary path to guide the primary path.

Some papers mainly focus on the query input of the transformer decoder to address HOI detection: PST [41] creatively introduces a composite transformer-based part-and-sum transformer decoder to replace the standard decoder. Based on this structure, PST passes composite queries into composite decoders to predict composite HOI sets. LV-HOI [42] employs a Fast-RCNN to detect person-bounding boxes and use them as queries for the subsequent transformer. The strategy chosen by Wang et al. [42] is to detect humans in images and use them as queries to search for corresponding interactions and target objects. In leveraging the model's object query capabilities, Category-Aware Transformer Network (CATN) [120] performs better by initializing object queries with category-aware semantic information. CDN-S [121] proposes an efficient knowledge distillation model, which shares parameters between teacher and student networks. QAHOI (Query-Based Anchors for Human-Object Interaction detection) [17] uses multi-scale architecture

to extract features from different spatial scales and uses query-based anchors to predict all elements of HOI instances. Zhang et al. [122] decoupled the human and object detection decoder and proposed a new human-object decoupling network (HOD). Zhang et al. [123] proposed a new end-to-end transformer-based framework (FGAHOI) to extract key features from images with complex background information and perform semantic alignment on the extracted features and query embeddings. In response to the potential overlap of multiple interaction actions and the lack of guidance information in existing query designs, Chan et al. [124] proposed Region Minning and Refined Query (RMRQ) in 2024. They utilized Ground Truth Mask Denoise (GTMD) to extract more features from different regions and proposed Dynamic Linguistic Query (DLQ) and Multi-label Focal Loss (MFL) to generate dynamic queries. The Parallel disentangling network (PDN) [125] proposed by Cheng et al. expands the naive query into triple explicit queries in parallel so that each query is focused on a specific task.

Some papers focus on the attention mechanism and the information obtained in the joint feature space to optimize performance: GTNet [126] encodes spatial context information in visual features through a self-attention mechanism and designs a guidance mechanism that combines object semantics and relative spatial configuration to guide the attention mechanism in the framework. The HORT [127] proposed by Ji et al. can greatly improve the understanding of human-object relationships by applying attention mechanisms among features distributed spatio-temporally. THID [128] proposes a new HOI visual encoder to detect interacting humans and objects and maps them to joint feature space for interaction recognition. MSTR [129] is a multi-scale transformer driven by two new deformable attention modules, entity-conditional contextual attention and dual-entity attention. Kim et al. [130] proposed a Multiple Relation Network (MUREN), which uses unary, pairwise, and ternary relationships of humans, objects, and interaction tokens to exchange context between three decoder branches. To solve the quantization problem that traditional feature extraction

943 methods (e.g., ROIAlign [131]) are difficult to directly apply
 944 when using the Visual Transformer backbone, Park et al.
 945 proposed a feature extraction method, named masking with
 946 overlapped area (MOA) module [132], to utilize the overlapped
 947 area between the given region and each patch in the attention
 948 function. Wang et al. proposed the Triple stream Enhanced
 949 encoder-decoder Dispersal Network (TED-Net) [133] in 2024.
 950 They designed a dispersal attention mechanism to capture indirect
 951 contact interaction information and an auxiliary discrimination
 952 mechanism to enhance the decoder's capability to recognize
 953 actions. Xia et al. proposed a cascade multi-scale transformer
 954 (CMST) [134] and used the multi-scale HOI attention mechanism,
 955 solving the problems of slow convergence and high
 956 computational complexity related to transformer architecture.
 957 Li et al. proposed LOGICHOI [135], which modified the self-
 958 attention mechanism in vanilla Transformer to enable it to infer
 959 $\langle \text{human}, \text{verb}, \text{object} \rangle$ triplet and constitute novel interactions.

960 The transformer-based method achieves end-to-end HOI de-
 961 tection. This means the model can predict HOI triplets directly
 962 from the input image without additional post-processing steps.
 963 This end-to-end solution simplifies the detection process and
 964 improves detection efficiency. Transformer has powerful mod-
 965 eling capabilities. Through its self-attention mechanism, it can
 966 not only capture long-distance dependencies in the image, but
 967 also pay attention to all positions in the input sequence, which
 968 enables the transformer to identify and understand HOI more
 969 accurately. At the same time, The transformer-based method is
 970 highly flexible and scalable. By adjusting the parameters and
 971 structure of the model, it can adapt to different HOI detection
 972 tasks and datasets.

973 In summary, the transformer-based method provides powerful
 974 assistance for HOI detection through its end-to-end solution,
 975 powerful modeling capabilities, ability to handle complex inter-
 976 actions, flexibility, and scalability. These advantages make the
 977 transformer one of the important technologies in HOI detection.

978 VI. METHODS RELATED TO FOUNDATION MODELS

979 The “Foundation Models” concept is first defined in the article
 980 “On the Opportunities and Risks of Foundation Models” [136]
 981 in 2021. Over 200 pages of text, Rishi Bommasani et al. provide
 982 a comprehensive introduction to the opportunities and risks of
 983 the foundation models, from their capabilities and technical
 984 principles to their applications and social impact. Foundation
 985 models are defined as an emerging paradigm for building AI
 986 systems based on a general class of models. A foundation
 987 model generally uses large-scale self-supervision so that it can
 988 be adapted to a wide range of downstream tasks. The current
 989 examples include BERT [137], GPT-3 [138] InstructGPT [139],
 990 GPT-4 [140], BLIP-2 [141] and CLIP [75]. The foundation
 991 models have multiple capabilities, such as language, vision,
 992 reasoning, interaction, and understanding, which shows that they
 993 have the potential to change the pattern of existing industries and
 994 expand the influence of AI in society.

995 We introduce the foundation models because, in the past two
 996 years, researchers have discovered that the foundation mod-
 997 els can be used to solve the long-tail distribution problem in

HOI detection. As we mentioned in Section I, the long-tail
 998 distribution problem refers to overfitting, underfitting, and other
 999 problems caused by an imbalance in the number of instance
 1000 samples between different categories in certain datasets. Since
 1001 2018, some scholars have begun to notice the seriousness of this
 1002 problem. In order to solve this problem, most current methods
 1003 can be divided into three categories: zero-shot/few-shot learning,
 1004 compositional learning, and weakly-supervised learning.

1005 *1) Zero-shot/Few-shot learning:* Shen et al. [21] introduced
 1006 a factorization model for HOI detection in 2018. Ji et al. [22]
 1007 formulated HOI as a few-shot task in a meta-learning framework.
 1008 Rb-PaStaNet [142] uses human prior knowledge to improve rare
 1009 HOI class detection; HO-RPN [143] designs a zero-shot classi-
 1010 fication module to identify new HOI class. Jiang et al. proposed
 1011 an innovative end-to-end parallel HOI detection framework
 1012 that incorporates compatibility self-learning (HOICS) [144]
 1013 for zero-shot HOI detection. Furthermore, DGIG-Net [145],
 1014 FG [146], ConsNet [147], and OC-Immunity [148] achieve
 1015 excellent results in zero-shot HOI detection experiments with
 1016 the new model.

1017 *2) Compositional learning:* Compositional learning facilitates
 1018 the affordance recognition of HOI models by enabling the learn-
 1019 ing of both known and unknown HOI concepts, which provides
 1020 new directions for solving the problem of long-tail HOI detec-
 1021 tion, such as FCL [63], VCL [64], ATL [149], ICompass [150],
 1022 Bongard-HOI [151] and SCL [152].

1023 *3) Weakly-supervised learning:* Baldassarre et al. [153]
 1024 proposed an explanation-based weakly-supervised method in
 1025 2020. In 2021, Kumaraswamy et al. [154] proposed a hybrid-
 1026 supervised HOI detection pipeline. The method proposed by Na-
 1027 garajan et al. [155] can infer weakly supervised spatial hotspot
 1028 maps from a given image.

1029 In addition, there are some methods for the long-tail distri-
 1030 bution problem that do not fall into the above three categories.
 1031 Gupta et al. [26] proposed a no-frills model for HOI detection
 1032 in 2019. Li et al. [156] proposed an interaction identification
 1033 method named Transferable Interactiveness Network (TIN) in
 1034 2019. Kim et al. [157] proposed a new concept of utilizing
 1035 co-occurring actions (ACPs) to train HOI detectors in 2020.
 1036 Gao et al. [158] proposed a dual-prior augmented decoding
 1037 network that utilizes external knowledge to enhance the model's
 1038 capabilities at a finer granularity.

1039 Since the large-scale visual language pre-training foundation
 1040 model CLIP is proposed, some studies have gradually focused on
 1041 CLIP-based HOI detection research. The timeline of methods to
 1042 solve the long-tail distribution problem is shown in Fig. 13. Liao
 1043 et al. designed the Guided-Embedding Network (GEN) [159]
 1044 in 2022. It is a dual-branch pipeline in which an instance de-
 1045 coder uses two independent query sets and a position-guided
 1046 embedding (p-GE) to detect humans and objects. Another in-
 1047 teractive decoder is used to classify interactions, wherein the
 1048 interactive query is generated by the instance-guided embedding
 1049 (i-GE) generated by the output of each instance decoder layer.
 1050 Then, they design a visual language knowledge transfer (VLKT)
 1051 training strategy to transfer knowledge from CLIP to HOI de-
 1052 tectors to enhance interactive understanding without additional
 1053 computational costs. The THID mentioned above also extracts

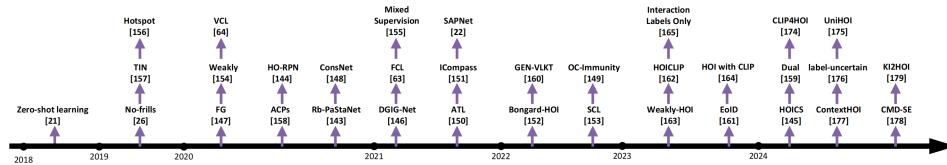


Fig. 13. Development timeline of methods to solve the long-tail distribution problem.

and leverages transferable knowledge from CLIP to perform zero-shot interaction detection. Wu et al. [160] proposed a new end-to-end zero-shot HOI detection (EoID) framework through visual language knowledge extraction in 2023. They extract interactive knowledge from CLIP and teach HOI models to recognize unseen actions. Ning et al. [161] proposed a new HOI detection framework (HOICLIP) that can effectively extract prior knowledge from CLIP and achieve better generalization. Wan et al. [162] developed a CLIP-guided HOI representation that can incorporate image-level and HOI instance-level prior knowledge and adopt a self-taught mechanism to prune incorrect human-object associations. At the same time, this is also a new weakly-supervised HOI detection strategy. It aims to integrate prior knowledge from pre-trained foundation models to facilitate HOI learning. Wan et al. [163] also design a multi-branch neural network and use CLIP to perform knowledge extraction at multiple levels. They do not use HOI annotations during training, which is a major leap forward in HOI detection. Unal et al. [164] tackle HOI detection in the weakest supervision setting in the system using only image-level interaction labels with the help of pre-trained large language models (LLM) and visual language models (VLM).

After entering 2024, research on HOI detection methods based on foundation models has further become the mainstream direction in the field of HOI. As of the writing of this paper, among the 15 papers related to HOI detection methods produced in 2024, 6 of them used foundation models, accounting for 40% of the total number of papers.

The CLIP4HOI [173] proposed by Mao et al. utilizes the generalizable knowledge of CLIP for unseen interaction recognition. By carefully designing the adaptation modules, CLIP is adjusted to a fine-grained HOI classifier, greatly improving the zero-shot transferability of CLIP4HOI and solving the problem of compromised generalization capability in joint human-object localization. Cao et al. proposed the first visual-textual HOI detection foundation model-based framework (UniHOI) [174], which utilizes LLM and Visual-Language (VL) foundation models to explore universal interaction recognition in open-world environments. Compared with HOI detectors that previously relied solely on specific datasets for training, the framework that combines foundation models significantly improves accuracy and versatility. Cao et al. also proposed HOI prompt-based learning, a high-level relation extraction method based on VL foundation models. The designed HOI Prompt-guided Decoder (HOPD) helps to associate high-level relation representations in the foundation model with various human-object pairs. In addition, Cao et al. used GPT for interaction interpretation to generate a richer linguistic understanding of complex HOIs. UniHOI effectively unleashes

the potential of LLMs and VL foundation models, Breaking through the limitations of traditional zero-shot/few-shot learning methods.

The Bongard-HOI [151] mentioned earlier reflects the problem of few-shot in HOI, and novel label-uncertain query augmentation techniques are introduced in the method proposed by Lei et al. [175] to enhance the diversity of query inputs. By distinguishing positive and negative HOI classes through the pre-trained CLIP image encoder and text encoder, effective learning of Bongard-HOI from a limited number of samples is promoted. To overcome the limitations of previous zero-shot learning methods in understanding context information and comprehensive reasoning, Gao et al. proposed a new HOI learning framework named ContextHOI [176], which serves as an effective contextual HOI detector, including a novel context-mining decoder and a powerful interaction reasoning LLM. The former aims to extract linguistic contextual information from a pre-trained VLM. Based on the extracted context information, the proposed interaction reasoning LLM utilizes rich linguistic knowledge to further improve the zero-shot reasoning ability. Lei et al. proposed an end-to-end open-vocabulary HOI detection framework, which features conditional multi-level decoding and fine-grained semantic enhancement (CMD-SE) [177], to address the issue of suboptimal performance of previous zero-shot methods in scenes containing human-object pairs with extensive distances. The framework employs various levels of feature maps from VLM to model HOIs at varying distances in the open-vocabulary scenario. At the same time, by utilizing the extensive world knowledge advantages of LLMs such as GPT, descriptions of human body part states are generated, enhancing the understanding of a large vocabulary of interactions. Xue et al. proposed a novel framework called Knowledge Integration to HOI (KI2HOI) [178], which effectively integrates the knowledge of VLM for zero-shot HOI detection. KI2HOI effectively utilizes prior knowledge and achieves excellent zero-shot transferability. Meanwhile, to deal with zero-shot learning in low-data, Xue et al. utilized prior knowledge from the CLIP text encoder to initialize the linear classifier, enhancing interaction understanding.

LLM has rich semantic knowledge. It acquires a lot of language knowledge through pre-training and provides text descriptions for HOI detection, which helps to understand the semantics of the interaction. GPT has strong zero-shot and few-shot learning capabilities. It can learn and reason without or with only a small amount of specific data, which is particularly important for solving the long-tail distribution problems. At the same time, GPT also has generative capabilities and can generate natural language descriptions to enhance the understanding of interactions.

TABLE I
HOI DATASETS

Reference	Year	Author	Name	Images	Action categories	Object categories	HOI categories	Parent datasets	Characteristics
[165]	2015	Gupta et al.	V-COCO	10346	26	80	-	MS-COCO	Early, Basic, Widely used
[166]	2015	Chao et al.	HICO	47774	116	80	600	-	New benchmark, Image-level annotations
[73]	2018	Chao et al.	HICO-DET	47776	117	80	600	HICO	New large-scale benchmark, Widely used
[167]	2018	Zhuang et al.	HCVRD	788160	927	1824	28323	Visual Genome	More types, Solve long-tail distribution problem
[33]	2020	Liao et al.	HOI-A	38668	10	11	17	-	Practical application
[168]	2020	Li et al.	HAKE	118000	156	76	-	HICO, HICO-DET, et al.	Human body part-level
[169]	2020	Li et al.	Ambiguous-HOI	8996	48	40	87	V-COCO, HCVRD, et al.	2D pose, Appearance ambiguities
[87]	2021	Zhong et al.	HOI-VP	29190	15	517	5825	HCVRD	Verb polysemy problem
[42]	2021	Wang et al.	SWiG-HOI	59000	406	1000	-	SWiG and DOH	Large vocabulary objects
[170]	2022	Bhatnagar et al.	BEHAVE	15200	-	20	-	-	3D annotations, RGBD sequences
[171]	2022	Li et al.	HAKE-HOI	110714	117	80	520	HICO-DET, HAKE	Larger than HICO-DET
[123]	2023	Ma et al.	HOI-SDC	46821	93	74	321	HAKE-HOI	Two new metrics
[172]	2023	Chen et al.	HOT	35750	-	-	-	V-COCO, HAKE, et al.	2D contact area heatmaps, Associated human part labels

VLM, such as BLIP-2 and CLIP, has strong cross-modal learning capabilities and can process visual and language information at the same time. This helps to combine visual features and natural language descriptions in HOI detection, which improves detection accuracy. At the same time, for open vocabulary HOI detection, the VLM model can use natural language descriptions to identify new and unseen interaction relationships, thereby expanding the detection capabilities.

The combination of LLM and VLM can provide richer interpretability and higher accuracy for HOI detection. It can not only intuitively display the interaction relationship, but also perform well in dealing with complex scenes and rare interaction relationships, and is better than the first three methods in dealing with the long-tail distribution problems.

In summary, the foundation model has rich semantic knowledge, strong zero-shot and few-shot learning capabilities, generative capabilities, cross-modal learning capabilities, richer interpretability, and higher accuracy. Observing Fig. 13, we can find that after 2022, there are more and more methods using the foundation model to solve long-tail distribution problems, which reflects the impact and potential of the foundation model on HOI detection.

VII. DATASET, METRICS AND PERFORMANCE

A. Dataset

This section summarizes the information of popular and widely used HOI detection datasets, as shown in Table I. From 2015 to 2024, a total of 13 datasets for HOI detection have emerged, including HICO, V-COCO, HICO-DET, HCVRD, HOI-A, HAKE, Ambiguous-HOI, HOI-VP, SWiG-HOI, BEHAVE, HAKE-HOI, HOI-SDC, and HOT.

V-COCO (Verbs in Common Objects in Context) dataset [165] is constructed by Saurabh Gupta et al. in 2015. It is a subset of COCO, with annotations of 160 K images with 80 different object classes. Its authors add additional annotations to COCO by using Amazon Mechanical Turk (AMT) to connect interacting people and objects and label their semantic roles. The V-COCO dataset contains 10,346 images with 16,199 people

instances, of which 2,533 images are used for training, 2,857 images are used for validating, and 4,946 images are used for testing. Each annotated person has 26 different binary action labels, and each image is annotated with 80 object categories. The richness of the data and the fact that all images inherit the annotations from COCO make V-COCO more suitable for HOI detection. This dataset is the earliest dataset used for HOI detection, so it is also the most basic and commonly used dataset.

HICO (Humans Interacting with Common Objects) [166] is a benchmark dataset proposed by Chao et al. at ICCV (IEEE International Conference on Computer Vision) in 2015, which includes 117 common behaviors from 80 objects. In natural scenes, people may interact with multiple objects simultaneously, so this dataset is object-centric with a large extension of label annotations for interaction categories. Although HICO proposes a new benchmark, it only contains image-level annotations, which makes it inconvenient to use.

HICO-DET [73] is a large-scale benchmark dataset for HOI detection, proposed by the same author as HICO in 2018. They augment HICO with instance annotations by setting up annotation tasks on AMT. HICO-DET has a total of 47,776 images, of which 38,118 are used for training and 9,658 are used for testing. It has 151,276 human instances and 600 HOI categories, including the same 80 object categories as COCO. Those HOI categories with fewer than ten training samples are called “rare” classes, and the rest are called “non-rare” categories; specifically, there are 138 rare categories and 462 non-rare categories. There are two modes of mean average precision on HICO-DET: the Default (DT) mode and the Known-Object (KO) mode. In DT mode, each HOI category is evaluated on all test images, while in KO mode, one HOI is evaluated only on images containing its associated object category. Compared with the V-COCO dataset, the HICO-DET dataset is larger in scale and has more diverse HOI categories. Therefore, it is one of the most commonly used datasets together with the V-COCO dataset.

HCVRD (Human-Centered Visual Relationship Detection) dataset [167] is proposed by Zhuang et al. in 2018, which is constructed based on the Visual Genome dataset [179]. It is currently the largest human-centric HOI dataset, with 788,160

1229 images and 927 action categories. At the same time, HCVRD has
 1230 more types, and it is also the first to use the zero-shot method
 1231 to monitor the labels appearing in the test set, which makes
 1232 it more suitable and easier to solve the long-tail distribution
 1233 problem.

1234 *HOI-A* (Human-Object Interaction for Application) dataset
 1235 [33] is proposed by Liao et al. in 2020. It consists of 38,668
 1236 annotated images with 11 interactive objects and 10 interactive
 1237 actions. To solve the problem that the V-COCO dataset and
 1238 the HICO-DET dataset have limited HOI categories that need
 1239 special attention in practical applications, HOI-A covers dif-
 1240 ferent appearance types, low resolution, and images with severe
 1241 occlusions that are difficult to identify. In addition, to expand the
 1242 intra-class variation of the data, each type of interaction in the
 1243 HOI-A dataset is divided into three scenarios: indoor, outdoor,
 1244 and in-vehicle, including dark, natural, and intense lighting
 1245 conditions and a variety of different angles. Therefore, this
 1246 dataset is suitable for testing in the above practical application
 1247 scenarios.

1248 *HAKE* [168] is the latest dataset proposed by Li et al. in the
 1249 field of HOI detection in 2020. It uses a large number of human
 1250 part states (PaSta) [180] labels to infer human body part states,
 1251 becoming the first large-scale instance dataset with fine-grained
 1252 annotations. It covers 247 K human instances, 220 K object
 1253 instances, and 7000 K local action labels. This dataset is suitable
 1254 for testing at the human body part level.

1255 *Ambiguous-HOI* [169] is generated to examine the ability
 1256 to process 2D pose and appearance ambiguities. It can better
 1257 evaluate the 2D ambiguity processing capacity of models. It
 1258 mainly obtains data from V-COCO, HCVRD, and other datasets,
 1259 forming a dataset with 8996 images, 87 HOI categories, 48
 1260 action categories, and 40 object categories.

1261 *HOI-VP* [87] proposed by Zhong et al. in 2021 is the first
 1262 database designed explicitly for the verb polysemy problem in
 1263 HOI detection. Specifically, it consists of 15 common verbs and
 1264 contains 517 common objects in natural scenes, thus ensuring the
 1265 richness and variety of semantics. Among them, the images and
 1266 labels of HOI-VP are collected based on the HCVRD dataset.

1267 *SWiG-HOI* [42] is a dataset for studying human interactions
 1268 with large vocabulary objects, consisting mainly of data from the
 1269 SWiG [181] and DOH datasets [182]. SWiG is initially collected
 1270 for grounded situational recognition (GSR) missions. SWiG-
 1271 HOI extracts the top 1000 common object categories and obtains
 1272 406 human actions, and there are 45 k training images and 14 k
 1273 test images.

1274 *BEHAVE* [170] is a HOI detection dataset with 3D annotations
 1275 of humans, objects, and interactions and is the largest dataset of
 1276 RGB-D sequences. The dataset contains 20 3D objects, eight
 1277 subjects (five males, three females), and five different locations,
 1278 totaling about 15.2 k records frames. They also provide ground
 1279 truth SMPL [183] and 3D object meshes. This dataset will be
 1280 helpful when studying HOI detection related to 3D annotations
 1281 and RGB-D sequences.

1282 *HAKE-HOI* [171] is split from the HAKE dataset. To al-
 1283 levi ate the impact of missing annotations in HICO-DET, Li
 1284 et al. exclude 80 non-interacting categories. This dataset has

1285 larger training and test sets than HICO-DET, providing 110,714
 1286 images.

1287 *HOI-SDC* [123] is established to solve the two difficulties of
 1288 the uneven size distribution of humans and objects and excessive
 1289 distance between humans and objects. Ma et al. proposed two
 1290 metrics to quantify these two difficulties in this dataset. At the
 1291 same time, the data of this dataset is selected from the HAKE-
 1292 HOI. This dataset is suitable for situations where the above two
 1293 difficulties also need to be solved.

1294 *HOT* [172] is a new dataset of HOI formed by collecting
 1295 data from four datasets: PROX [184], V-COCO, HAKE, and
 1296 Watch-n-Patch [185]. Among them, PROX is a 3D human
 1297 meshes dataset moving in 3D scenes. Chen et al. used this
 1298 dataset to automatically annotate 2D image areas for contact and
 1299 annotate relevant body parts of the human body. Watch-n-Patch
 1300 is an indoor action recognition dataset, which Chen et al. used
 1301 along with V-COCO and HAKE to draw polygons around the 2D
 1302 image areas where contact occurred. HOT has 2D contact area
 1303 heatmaps and related human body part labels as annotations,
 1304 using automatic generation and manual annotation.

B. Metrics

1305 In this section, we introduce the accuracy evaluation metrics
 1306 and calculation process for HOI detection in detail.

1307 Early HOI detection mainly uses the accuracy rate (A_{cc}) to
 1308 evaluate, which is defined as the proportion of correct samples
 1309 detected in all samples. However, when the sample data distri-
 1310 bution is uneven, the use of accuracy evaluation is prone to bias,
 1311 and the type of error cannot be objectively described.

1312 Currently, HOI detection uses mean average precision (mAP)
 1313 and average precision (AP) as the standard evaluation criteria.
 1314 The AP used in the HOI detection task is the AP of the triplet
 1315 $\langle \text{human}, \text{verb}, \text{object} \rangle$, called “role AP”(AP_{role}). In the object
 1316 detection task, if the Intersection over Union (IoU) of the object
 1317 bounding box and the ground-truth bounding box is 0.5 or higher,
 1318 it is considered a true positive (TP). The HOI detection task is
 1319 modified based on this discriminant. A triplet is considered as a
 1320 true positive if: i) the predicted human box b_h has IoU of 0.5 or
 1321 higher with the ground-truth human box, ii) the predicted object
 1322 box b_o has IoU of 0.5 or higher with the ground-truth target
 1323 object, and iii) the predicted and ground-truth actions match.

1324 To calculate AP, a confusion matrix is needed. The confu-
 1325 sion matrix includes four calculation variables: TP, FN (false
 1326 negative), FP (false positive), and TN (true negative).

1327 Precision refers to the proportion of true positive samples to
 1328 all positive samples predicted by the HOI detection model. It
 1329 is defined as (5), in which DE represents the number of all
 1330 prediction boxes.

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{DE} \quad (5)$$

1332 Recall refers to the proportion of positive samples predicted
 1333 by the human-object interaction detection model to be correct
 1334 among all real positive samples. It is defined as (6), in which GT

TABLE II
TEST RESULTS OF TWO-STAGE METHODS ON THE V-COCO DATASET AND THE HICO-DET DATASET

Reference	Year	Name	Backbone	Structure	AProle	Default			Known Object		
						full	rare	non-rare	full	rare	non-rare
[73]	2018	HO-RCNN	CaffeNet	Other	-	7.81	5.37	8.54	10.41	8.94	10.85
[27]	2018	ICAN	ResNet-50	Attention	45.3	14.84	10.45	16.15	16.26	11.33	17.73
[156]	2019	TIN	ResNet-50	Other	47.8	17.03	13.42	18.11	19.17	15.51	20.26
[28]	2019	RPNN	ResNet-50	GNN	47.5	17.35	12.78	18.71	-	-	-
[29]	2020	FCMNet	ResNet-50	Attention	53.1	20.41	17.34	21.56	22.04	18.97	23.12
[104]	2020	in-GraphNet	ResNet-50	GNN	48.9	17.72	12.93	19.31	-	-	-
[110]	2021	PFNet	ResNet-50	GNN	52.8	20.05	16.66	21.07	24.01	21.09	24.89
[92]	2021	AGRR	ResNet-50	Attention	48.1	16.63	11.30	18.22	19.22	14.56	20.61
[106]	2021	SG2HOI	ResNet-50	GNN	53.3	20.93	18.24	21.78	24.83	20.52	25.32
[97]	2022	Cross-Person Cues	ResNet-50	Attention	63.0	35.15	33.71	35.58	37.56	35.87	38.06
[32]	2018	InteractNet	ResNet-50-FPN	Attention	40.0	9.94	7.16	10.77	-	-	-
[62]	2019	PMFNet	ResNet-50-FPN	Attention	52.0	17.46	15.65	18.00	20.34	17.47	21.20
[30]	2020	DRG	ResNet-50-FPN	GNN	51.0	19.26	17.74	19.71	23.40	21.75	23.89
[31]	2020	CHGN	ResNet-50-FPN	GNN	52.7	17.57	16.85	17.78	21.00	20.74	21.08
[61]	2020	PMN	ResNet-50-FPN	Attention	51.8	21.21	17.60	22.29	-	-	-
[108]	2020	VS-GATs	ResNet-50-FPN	GNN	50.6	20.27	16.03	21.54	-	-	-
[69]	2020	MLCNet	ResNet-50-FPN	GNN	55.2	17.95	16.62	18.35	22.28	20.73	22.74
[147]	2020	ConsNet	ResNet-50-FPN	GNN	53.2	22.15	17.55	23.52	-	-	-
[105]	2021	SCG	ResNet-50-FPN	GNN	54.2	31.33	24.72	33.31	-	-	-
[107]	2021	IGPN	ResNet-50-FPN	GNN	53.8	21.26	18.47	22.07	-	-	-
[88]	2022	SDT	ResNet-101	Attention	61.8	32.97	28.49	34.31	36.32	31.90	31.64
[94]	2024	GFN	ResNet-101	Attention	70.1	35.28	31.91	36.29	38.80	35.48	39.79
[74]	2018	GPNN	ResNet-152	GNN	44.0	13.11	9.34	14.23	-	-	-
[26]	2019	No-frills	ResNet-152	Other	-	17.18	12.17	18.68	-	-	-
[60]	2020	VSGNet	ResNet-152	GNN	51.8	19.80	16.05	20.91	-	-	-
[87]	2021	PD-Net	ResNet-152	Attention	52.2	22.37	17.61	23.79	26.86	21.70	28.44
[96]	2021	OSGNet	ResNet-152	Attention	53.4	21.40	18.12	22.38	-	-	-
[93]	2022	Actor-centric	HRNet-W32	Attention	51.7	27.39	21.34	29.20	30.87	24.20	32.87
[90]	2024	PPDM++	Hourglass-104	Attention	54.3	30.10	23.73	32.00	31.80	24.93	33.85

represents the number of true bounding boxes for all objects.

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{GT} \quad (6)$$

AP refers to the proportion of all the accuracy rates and the number of images in the category, and it measures the quality of the model's judgment results in a single category. It is defined as (7):

$$AP = \int_0^1 p(r)dr \quad (7)$$

In the above formula, p represents Precision, r represents Recall, and $p(r)$ is a function with r as a parameter, which is called Precision-Recall curve (PR curve). The integral of the function represents the average precision rate. The AP value represented by this formula can also be regarded as the area under the PR curve. PR curve is widely used in object detection, saliency detection, and other fields. AP is under a single category, and mAP is the mean of AP values under all categories, which measures how well the model judges the results across all categories.

C. Performance

Considering that the V-COCO dataset and the HICO-DET dataset are the two most commonly used datasets, and basically all HOI detection methods have been tested on them, we extract

the test results on these two datasets from the articles mentioned above. We rank the test results of each method according to backbones as the main influencing factor and paper publication time as the secondary influencing factor. The test results of the two-stage, one-stage, transformer-based, and foundation model methods are presented in Tables II, III, IV, and V.

By analyzing the accuracy of each method in the above tables, we can draw the following conclusions.

- With the advancement of technology, the accuracy of HOI detection methods has been continuously improved, showing an overall upward trend. It can be inferred that with the further development of technology in the future, the accuracy of HOI detection will be further improved.
- From Table II, we can see that for the attention-based and GNN-based methods, the methods of incorporating body gestures, the methods of incorporating contextual information, and the GAT-based methods introduce additional information and features, making the accuracy significantly improved compared to the previous generalized methods.
- Compared with the two-stage and one-stage methods, we can find that the overall accuracy of the transformer-based methods and the foundation model methods are the best. The transformer-based methods and the foundation model methods have the highest accuracy in both the V-COCO

TABLE III
TEST RESULTS OF ONE-STAGE METHODS ON THE V-COCO DATASET AND THE HICO-DET DATASET

Reference	Year	Name	Backbone	AP _{Role}	Default			Known Object		
					full	rare	non-rare	full	rare	non-rare
[34]	2020	IP-Net	Hourglass-104	51.0	19.56	12.79	21.58	22.05	15.77	23.92
[33]	2020	PPDM	Hourglass-104	-	21.73	13.78	24.10	24.58	16.65	26.84
[35]	2021	GGNet	Hourglass-104	54.7	29.17	22.13	30.84	33.50	26.67	34.89
[36]	2020	UnionDet	ResNet-50-FPN	47.5	17.58	11.72	19.33	19.76	14.68	21.27
[111]	2021	DIRV	EfficientDet-d3	56.1	21.81	16.35	23.44	25.84	21.02	27.28

TABLE IV
TEST RESULTS OF TRANSFORMER-BASED METHODS ON THE V-COCO DATASET AND THE HICO-DET DATASET

Reference	Year	Name	Backbone	AP _{Role}	Default			Known Object		
					full	rare	non-rare	full	rare	non-rare
[39]	2021	HOTR	ResNet-50	55.2	25.10	17.34	27.42	-	-	-
[41]	2021	PST	ResNet-50	-	23.93	14.98	26.60	26.42	17.61	29.05
[40]	2021	AS-Net	ResNet-50	53.9	28.87	24.25	30.25	31.74	27.07	33.14
[116]	2022	STIP	ResNet-50	66.0	32.22	28.15	33.43	35.29	31.43	36.45
[117]	2022	CPC	ResNet-50	63.1	29.63	23.14	31.57	-	-	-
[120]	2022	CATN	ResNet-50	60.1	31.86	25.15	33.84	34.44	27.69	36.45
[121]	2022	CDN-S	ResNet-50	63.5	33.28	29.19	34.50	-	-	-
[129]	2022	MSTR	ResNet-50	62.0	31.17	25.31	32.92	34.02	28.83	35.57
[130]	2023	MUREN	ResNet-50	68.8	32.87	28.67	34.12	35.52	30.88	36.91
[124]	2024	RMRQ	ResNet-50	61.1	31.11	25.16	32.88	33.89	27.78	35.72
[134]	2024	CMST	ResNet-50	63.9	32.20	27.95	33.56	34.90	30.04	36.33
[133]	2024	TED-Net	ResNet-50	63.4	34.00	29.88	35.24	37.13	33.63	38.18
[135]	2024	LOGICHOI	ResNet-50	64.4	35.47	32.03	36.22	38.21	35.29	39.03
[37]	2021	HOI-Trans	ResNet-101	52.9	26.61	19.15	28.84	29.13	20.98	31.57
[38]	2021	QPIC	ResNet-101	58.8	29.90	23.92	31.69	32.38	26.06	34.27
[114]	2022	OCN	ResNet-101	65.3	31.43	25.80	33.11	-	-	-
[118]	2022	PhraseHOI	ResNet-101	-	30.03	23.48	31.99	33.74	27.35	35.64
[119]	2023	SGPT	ResNet-101	60.3	30.08	24.00	31.89	32.40	26.49	34.16
[122]	2023	HOD	ResNet-101	64.0	34.48	30.38	35.71	-	-	-
[125]	2024	PDN	ResNet-101	64.7	33.18	27.95	34.75	35.86	30.57	37.43
[113]	2022	UPT	ResNet-101-DC5	61.3	32.62	28.62	33.81	36.08	31.41	37.47
[126]	2023	GTNet	ResNet-152	58.3	29.71	23.23	31.64	31.64	24.42	33.81
[17]	2023	QAHOI	Swin-Large [*] ₊	-	35.78	29.80	37.56	37.59	31.66	39.36
[123]	2024	ViPLO	ViT-B/16	62.2	37.22	35.45	37.75	40.61	38.82	41.15

dataset and the HICO-DET dataset. Moreover, the mAP average value of the listed foundation model methods is 61.68, slightly higher than the mAP average value of 61.07 for the transformer-based methods. And both of these methods are significantly higher than the mAP average value of 52.21 for the two-stage methods and the mAP average value of 52.33 for the one-stage methods. In summarizing the research time of each method, it can be found that in the past two years, there have been more and more papers on transformer-based methods and foundation model methods, indicating that these two types of methods have become the main research directions for HOI detection.

4) Since the backbone network will directly affect the feature extraction process in the HOI detection process, the selection of the backbone network will also affect the accuracy.

At present, for the two-stage methods, transformer-based methods, and foundation model methods, the mainstream backbone networks are ResNet [49] and the feature pyramid networks (FPN) [186] based on ResNet. For the one-stage methods, Hourglass-104 [52] is mainly used to extract features.

Regarding the detection speed, not all methods currently provide their own real-time test results. Moreover, the GPU model, parameters, test datasets, etc., used in the papers that provide detection speed are inconsistent. PPDM [33] tested its efficiency on the HOI-A dataset using Titan Xp GPU and CUDA 9.0, outperforming ICAN and TIN. UnionDet [36] used NVIDIA GTX1080Ti GPU to test the additional inference time on V-COCO and HICO-DET datasets, and the results show that it outperformed ICAN, TIN, GPNN, and InteractNet. DIRV [111] outperformed PMFNet, InteractNet, VSGNet,

TABLE V
TEST RESULTS OF FOUNDATION MODEL METHODS ON THE V-COCO DATASET AND THE HICO-DET DATASET

Reference	Year	Name	Backbone	AP _{role}	Default			Known Object		
					full	rare	non-rare	full	rare	non-rare
[161]	2023	HOICLIP	ResNet-50	63.5	34.69	31.12	35.74	37.61	34.47	38.54
[173]	2024	CLIP4HOI	ResNet-50	66.3	35.33	33.95	35.74	37.19	35.27	37.77
[178]	2024	KI2HOI	ResNet-50	63.9	34.20	32.26	36.10	37.85	35.89	38.78
[159]	2022	GEN-VLKT	ResNet-101	63.6	34.95	31.18	36.08	38.22	34.36	39.37
[162]	2023	Weakly-HOI	ResNet-101	44.7	25.70	24.52	26.05	-	-	-
[174]	2024	UniHOI	ResNet-101	68.1	40.95	40.27	41.32	43.26	43.12	43.25

PPDM, and UnionDet by testing inference time on V-COCO and HICO-DET datasets using NVIDIA RTX2080Ti GPU. AS-Net [40] tested the inference speed on the HICO-DET dataset using GeForce GTX 1080Ti GPU and CUDA 9.0, outperforming InteractNet, ICAN, no-frills, PMFNet, DRG, VSGNet, PPDM, and UnionDet. OCN [114] tested the detection time on the HICO-DET dataset using Tesla V100, and its performance was better than PPDM, HOTR, and ASNet, while QPIC performed slightly better than OCN. HOICLIP [161] tested the inference time using NVIDIA A100 GPUs, and its performance (55.52 ms/img) was almost the same as GEN-VLKT (52.80 ms/img). ViPLO [132] tested the inference speed on the HICO-DET dataset using a Geforce RTX 3090 GPU. It used 3119 MiB of GPU memory and had a speed of 131 ms, compared to SCG which used 3423 MiB of GPU memory and had a speed of 106 ms.

Although the GPU model, parameters, and test datasets are inconsistent, we can still find from the results of PPDM, Union-Det, and DIRV that the one-stage method has better detection speed than the two-stage method due to the influence of the one-stage detector and parallel detection. The results of AS-Net and OCN also further show that the one-stage method has no advantage in detection speed compared with the transformer-based method. This is also the main reason why the progress of the one-stage methods has stagnated after the emergence of DETR.

In summary, the overall results of the two datasets and different papers on detection time show that HOI detection is developing towards improving detection accuracy and speed. Accuracy and real-time performance indicate whether the HOI detection method is excellent.

VIII. FUTURE DIRECTIONS

Deep learning-based HOI detection methods are still in development, and there are many difficulties that need to be explored and solved. Future research can be carried out from the following aspects.

1) *Domain adaptations for various scenes*: The future HOI detection algorithm should be both general and professional. Generality refers to the ability to apply models trained on existing data to unseen scenarios. Professional refers to the ability to train models for specific usage scenarios based on existing data. Domain adaptations can effectively reduce the dependence on labeled training data [187], [188]. At present, the most advanced

research directions in this area are few-shot and zero-shot learning, both of which are the application of Meta-Learning in the field of supervised learning. The two hope that the machine learning model can quickly learn new categories with only a small number of samples or zero samples. At present, some early work has begun to explore their application value in the field of HOI detection, such as [142], [143], [160].

2) *Transformer models are getting increasing attention*: Transformer's advantages of global computing and perfect memory make it popular in the machine-learning community. As we mentioned above, the transformer-based methods are slightly more efficient than the one-stage methods in grouping results and interaction classification, and they are also better than the two-stage methods in terms of real-time and accuracy performance. We can find that transformer-based HOI detection methods have become the main research direction in 2024 [124], [125], [133], [134], [135] and they are likely to get more attention in the future. At present, transformer-based methods still face difficulties such as long convergence time, poor interpretability, high computational complexity, weak deployment ability in natural environments, and substantial memory needed for model computation. In the future, further research can be conducted on optimizing decoder computing power, mining more interaction information between human pose and contact environment, extracting features from objects that are too small in images, simplifying configuration parameters and fine-tuning processes, and enhancing real-time application performance and generalization performance.

3) *Foundation models to solve the long-tail distribution problem*: Nowadays, the popularity of ChatGPT has caused more and more researchers to focus on the foundation models. In the field of HOI detection, there have been some articles using the foundation models, such as CLIP, GPT, and BLIP-2, to solve the long-tail distribution problem, such as [173], [174], [175], [176], [177], [178]. The foundation models also perform excellently in solving problems such as excessive reliance on a large number of HOI text labels for classification, poor transferability, complex and diverse HOIs in the real world, poor ability to understand contextual information, and weak comprehensive reasoning ability in previous HOI detection methods. It is conceivable that with the further development of foundation models, more researchers will try to use foundation models to optimize existing HOI detection methods. In the future, further research can be conducted through foundation models to generate and capture uncommon interactions in the long-tail distribution problem, enhance the interaction interpretation and language

1499 understanding of complex HOIs using prior knowledge, enhance
 1500 contextual context and zero-shot reasoning ability, and generate
 1501 human body part state descriptions using the extensive world
 1502 knowledge.

1503 *4) More comprehensive and professional datasets:* The existing
 1504 mainstream datasets all have room for further optimization.
 1505 For example, in [73], [165], [166], the number of instance
 1506 samples between different categories is unbalanced, which will
 1507 cause long-tail distribution. At the same time, there is also the
 1508 problem that the interaction categories are not comprehensive.
 1509 Although the current largest dataset HCVRD [167] has 9852
 1510 behavior categories, it still cannot cover the relationship category
 1511 between all humans and objects. In addition, datasets with more
 1512 professional characteristics also have research value.
 1513 For example, to solve the common occlusion problem in visual
 1514 perception, Hu et al. [189] proposed a hidden human object
 1515 detection dataset inspired by physiological features; In the field
 1516 of human-robot handover, Wiederhold et al. [190] proposed the
 1517 HOH (Human-Object-Human) Handover dataset to overcome
 1518 the problem of previous datasets requiring specific body tracking;
 1519 MECCANO [191] and EgoISM-HOI [192] are proposed to study
 1520 HOIs in the industrial environment; Li et al. developed V-HICO
 1521 [193] to solve the problem of traditional video datasets Charades
 1522 [102], EpicKitchens [194], VidVRD [195],
 1523 VidOR [196], YouCook [197], [198], being unsuitable for HOI
 1524 detection, while VidHOI [199], D3D-HOI [200], HOI4D [201],
 1525 etc. were also proposed to study the video field. Extracting
 1526 synthetic data from games is also an excellent way to augment
 1527 the dataset.

1528 *5) Multi-modal optimizes HOI detection performance:* In
 1529 recent years, multi-modal fusion target detection has become
 1530 a new direction to solve problems such as the poor performance
 1531 of traditional target detection in complex environments.
 1532 Similarly, multi-modal fusion can also use information from
 1533 different modalities to more comprehensively capture interactive
 1534 information, further improving the robustness, accuracy and
 1535 versatility of the HOI detection model. In addition to traditional
 1536 visual information, the available modal information can also be
 1537 obtained from sensor [202], radar [203], LiDAR [204],
 1538 sound [205], 3D pose and mesh reconstruction [206], etc.

1539 *6) Beyond HOI detection: HOI generation:* In the HOI motion
 1540 generation field, various innovative methods have emerged in
 1541 recent years. These methods aim to simulate and generate HOI
 1542 motions that conform to real physical rules and semantic logic
 1543 based on the characteristics of scenes [207], text [208], [209]
 1544 or objects [210]. Combining physical simulation, reinforcement
 1545 learning, modular design, and deep learning technologies makes
 1546 the HOI motion generation more natural, accurate, and widely
 1547 applicable. These research results provide strong support for virtual
 1548 reality (VR), augmented reality (AR), and human-machine
 1549 interaction (HMI).

IX. CONCLUSION

1551 HOI detection plays an essential role in many research fields,
 1552 but there is no comprehensive and precise method to define and
 1553 classify all HOI detection methods. This paper fills this gap
 1554 with the following content. Firstly, this paper introduces the

1555 definition, application, and main problems of HOI detection.
 1556 At the same time, we describe the innovations of this paper
 1557 compared with previous HOI detection reviews. Then, this paper
 1558 gives our classification criteria and formula definitions for HOI
 1559 detection methods and expounds on each method's advantages,
 1560 disadvantages, and development process. After that, starting
 1561 from the overall structure of the model, this paper divides the
 1562 HOI detection methods into the two-stage methods, one-stage
 1563 methods, and transformer-based methods and expounds them,
 1564 respectively. At the same time, we introduce the definition of the
 1565 popular foundation models and a series of methods successfully
 1566 using the foundation model to solve the long-tail distribution
 1567 problem faced by HOI detection. Then, this paper introduces the
 1568 datasets and metrics commonly used in the HOI field specifically
 1569 and analyzes the performance of each method on the V-COCO
 1570 dataset and the HICO-DET dataset. Finally, this paper points out
 1571 future development directions.

REFERENCES

- [1] N. Wang, Y. Wang, and M. J. Er, "Review on deep learning techniques for marine object recognition: Architectures and algorithms," *Control Eng. Pract.*, vol. 118, 2022, Art. no. 104458.
- [2] S. Rani, K. Lakhwani, and S. Kumar, "Three dimensional objects recognition & pattern recognition technique; related challenges: A review," *Multimedia Tools Appl.*, vol. 81, pp. 17303–17346, 2022.
- [3] L. Jiao et al., "New generation deep learning for video object detection: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3195–3215, Aug. 2022.
- [4] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.
- [5] W. Hui, H. Li, M. Chen, and A. Song, "Robotic tactile recognition and adaptive grasping control based on CNN-LSTM," *Chin. J. Sci. Instrum.*, vol. 40, pp. 211–218, 2019.
- [6] S. Haresh, X. Sun, H. Jiang, A. X. Chang, and M. Savva, "Articulated 3D human-object interactions from RGB videos: An empirical analysis of approaches and challenges," in *Proc. Int. Conf. 3D Vis.*, 2022, pp. 312–321.
- [7] S. Tang, D. Roberts, and M. Golparvar-Fard, "Human-object interaction recognition for automatic construction site safety inspection," *Automat. Construction*, vol. 120, 2020, Art. no. 103356.
- [8] C. Jiang, Z. Ni, Y. Guo, and H. He, "Pedestrian flow optimization to reduce the risk of crowd disasters through human–robot interaction," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 4, no. 3, pp. 298–311, Jun. 2020.
- [9] A. Erkul, "Automated analysis of crossing actions in football commentary using large language models," M.S. thesis, Inf. Syst. Dept., Middle East Tech. Univ., Çankaya/Ankara, Türkiye, 2024.
- [10] M. H. Sarkhoosh, S. Gautam, C. Midoglu, S. S. Sabet, and P. Halvorsen, "Multimodal AI-based summarization and storytelling for soccer on social media," in *Proc. 15th ACM Multimedia Syst. Conf.*, 2024, pp. 485–491.
- [11] H. Hu, X. Yi, Z. Cao, J.-H. Yong, and F. Xu, "Hand-object interaction controller (Hoic): Deep reinforcement learning for reconstructing interactions with physics," in *Proc. ACM SIGGRAPH Conf. Papers*, 2024, pp. 1–10.
- [12] W. Du, "The computer vision simulation of Athlete's wrong actions recognition model based on artificial intelligence," *IEEE Access*, vol. 12, pp. 6560–6568, 2024.
- [13] K. Ehsani, S. Tulsiani, S. Gupta, A. Farhadi, and A. Gupta, "Use the force, luke! learning to predict physical forces by simulating effects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 224–233.
- [14] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with R* CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1080–1088.
- [15] M. Savva et al., "Habitat: A platform for embodied ai research," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9339–9347.

- [16] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, "A survey of embodied AI: From simulators to research tasks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 2, pp. 230–244, 2022.
- [17] J. Chen and K. Yanai, "QAHOI: Query-based anchors for human-object interaction detection," in *Proc. 18th Int. Conf. Mach. Vis. Appl.*, 2023, pp. 1–5.
- [18] N. Bodla, G. Shrivastava, R. Chellappa, and A. Shrivastava, "Hierarchical video prediction using relational layouts for human-object interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12146–12155.
- [19] T. Qiao, Q. Men, F. W. Li, Y. Kubotani, S. Morishima, and H. P. Shum, "Geometric features informed multi-person human-object interaction recognition in videos," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 474–491.
- [20] R. Moraes, V. Le, S. Venkatesh, and T. Tran, "Learning asynchronous and sparse human-object interaction in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16041–16050.
- [21] L. Shen, S. Yeung, J. Hoffman, G. Mori, and L. Fei-Fei, "Scaling human-object interaction recognition through zero-shot learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1568–1576.
- [22] Z. Ji, X. Liu, Y. Pang, W. Ouyang, and X. Li, "Few-shot human-object interaction recognition with semantic-guided attentive prototypes network," *IEEE Trans. Image Process.*, vol. 30, pp. 1648–1661, 2021.
- [23] Y.-L. Li, X. Liu, X. Wu, Y. Li, and C. Lu, "HOI analysis: Integrating and decomposing human-object interaction," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 5011–5022.
- [24] W. Yang, Y. Song, Z. Zhao, and F. Su, "Instance search via fusing hierarchical multi-level retrieval and human-object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2323–2327.
- [25] X. Liu, Y.-L. Li, X. Wu, Y.-W. Tai, C. Lu, and C.-K. Tang, "Interactiveness field in human-object interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20113–20122.
- [26] T. Gupta, A. Schwing, and D. Hoiem, "No-frills human-object interaction detection: Factorization, layout encodings, and training techniques," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9677–9685.
- [27] C. Gao, Y. Zou, and J.-B. Huang, "iCAN: Instance-centric attention network for human-object interaction detection," 2018, *arXiv:1808.10437*.
- [28] P. Zhou and M. Chi, "Relation parsing neural network for human-object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 843–851.
- [29] Y. Liu, Q. Chen, and A. Zisserman, "Amplifying key cues for human-object-interaction detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 248–265.
- [30] C. Gao, J. Xu, Y. Zou, and J.-B. B. Huang, "DRG: Dual relation graph for human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, vol. 12357, pp. 696–712.
- [31] H. Wang, W.-s. Zheng, and L. Yingbiao, "Contextual heterogeneous graph network for human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 248–264.
- [32] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8359–8367.
- [33] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 482–490.
- [34] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4116–4125.
- [35] X. Zhong, X. Qu, C. Ding, and D. Tao, "Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13234–13243.
- [36] B. Kim, T. Choi, J. Kang, and H. J. Kim, "Uniondet: Union-level detector towards real-time human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 498–514.
- [37] C. Zou et al., "End-to-end human object interaction detection with hoi transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11825–11834.
- [38] M. Tamura, H. Ohashi, and T. Yoshinaga, "QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10410–10419.
- [39] B. Kim, J. Lee, J. Kang, E.-S. Kim, and H. J. Kim, "HOTR: End-to-end human-object interaction detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 74–83.
- [40] M. Chen, Y. Liao, S. Liu, Z. Chen, F. Wang, and C. Qian, "Reformulating HOI detection as adaptive set prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9004–9013.
- [41] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3550–3559.
- [42] S. Wang, K.-H. Yap, H. Ding, J. Wu, J. Yuan, and Y.-P. Tan, "Discovering human interactions with large-vocabulary objects via query and multi-scale detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13475–13484.
- [43] M. Antoun and D. Asmar, "Human object interaction detection: Design and survey," *Image Vis. Comput.*, vol. 130, 2023, Art. no. 104617.
- [44] T. Bergstrom and H. Shi, "Human-object interaction detection: A quick survey and examination of methods," in *Proc. 1st Int. Workshop Hum.-Centric Multimedia Anal.*, 2020, pp. 63–71.
- [45] S. Sunaina, R. Kaur, and D. V. Sharma, "A review of vision-based techniques applied to detecting human-object interactions in still images," *J. Comput. Sci. Eng.*, vol. 15, no. 1, pp. 18–33, 2021.
- [46] F. Li, S. Wang, S. Wang, and L. Zhang, "Human-object interaction detection: A survey of deep learning-based methods," in *Proc. CAAI Int. Conf. Artif. Intell.*, 2022, pp. 441–452.
- [47] J. Wang, H.-H. Shuai, Y.-H. Li, and W.-H. Cheng, "Human-object interaction detection: An overview," *IEEE Consum. Electron. Mag.*, vol. 13, no. 6, pp. 56–72, Nov. 2024.
- [48] Q. Dang, J. Yin, B. Wang, and W. Zheng, "Deep learning based 2D human pose estimation: A survey," *Tsinghua Sci. Technol.*, vol. 24, no. 6, pp. 663–676, 2019.
- [49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [50] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, Jun. 2019, pp. 10691–10700.
- [51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2012, vol. 25, pp. 1097–1105.
- [52] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 483–499.
- [53] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2019.
- [54] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.
- [55] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "FastText.zip: Compressing text classification models," 2016, *arXiv:1612.03651*.
- [56] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
- [57] H. Yuan et al., "RLIP: Relational language-image pre-training for human-object interaction detection," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 37416–37431.
- [58] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [59] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
- [60] O. Ulutan, A. Iftekhar, and B. S. Manjunath, "VSGNet: Spatial attention network for detecting human object interactions using graph convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13617–13626.
- [61] Z. Liang, J. Liu, Y. Guan, and J. Rojas, "Pose-based modular network for human-object interaction detection," 2020, *arXiv:2008.02042*.
- [62] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9469–9478.
- [63] Z. Hou, B. Yu, Y. Qiao, X. Peng, and D. Tao, "Detecting human-object interaction via fabricated compositional learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14646–14655.

- [64] Z. Hou, X. Peng, Y. Qiao, and D. Tao, "Visual compositional learning for human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 584–600.
- [65] H.-S. Fang, Y. Xie, D. Shao, and C. Lu, "DIRV: Dense interaction region voting for end-to-end human-object interaction detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1291–1299.
- [66] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, vol. 12346, pp. 213–229.
- [67] W. Feng, W. Liu, T. Li, J. Peng, C. Qian, and X. Hu, "Turbo learning framework for human-object interactions recognition and human pose estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 01, pp. 898–905.
- [68] F. Z. Zhang, D. Campbell, and S. Gould, "Spatio-attentive graphs for human-object interaction detection," 2020, *arXiv:2012.06060*.
- [69] X. Sun, X. Hu, T. Ren, and G. Wu, "Human object interaction detection via multi-level conditioned network," in *Proc. 2020 Int. Conf. Multimedia Retrieval*, 2020, pp. 26–34.
- [70] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
- [71] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
- [72] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Le-Cun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," 2013, *arXiv:1312.6229*.
- [73] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 381–389.
- [74] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 401–417.
- [75] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
- [76] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
- [77] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
- [78] R. Cong, W. Song, J. Lei, G. Yue, Y. Zhao, and S. Kwong, "PSNet: Parallel symmetric network for video salient object detection," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 2, pp. 402–414, Apr. 2023.
- [79] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.
- [80] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1347–1360, Mar. 2018.
- [81] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 33–44.
- [82] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.
- [83] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1831–1840.
- [84] W. Su et al., "VL-BERT: Pre-training of generic visual-linguistic representations," 2019, *arXiv:1908.08530*.
- [85] A. Kolesnikov, A. Kuznetsov, C. Lampert, and V. Ferrari, "Detecting visual relationships using box attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 1749–1753.
- [86] X. Zhong, C. Ding, X. Qu, and D. Tao, "Polysemy deciphering network for human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 69–85.
- [87] X. Zhong, C. Ding, X. Qu, and D. Tao, "Polysemy deciphering network for robust human-object interaction detection," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1910–1929, 2021.
- [88] G. Wang, Y. Guo, Y. Wong, and M. Kankanhalli, "Distance matters in human-object interaction detection," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 4546–4554.
- [89] T. Zhou, W. Wang, S. Qi, H. Ling, and J. Shen, "Cascaded human-object interaction recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4263–4272.
- [90] Y. Liao et al., "PPDM++: Parallel point detection and matching for fast and accurate HOI detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 10, pp. 6826–6841, Oct. 2024.
- [91] T. Wang et al., "Deep contextual attention for human-object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5694–5702.
- [92] X. Lin, Q. Zou, and X. Xu, "Action-guided attention mining and relation reasoning network for human-object interaction detection," in *Proc. 29th Int. Conf. Int. Joint Conf. Artif. Intell.*, 2021, pp. 1104–1110.
- [93] K. Xu et al., "Effective actor-centric human-object interaction detection," *Image Vis. Comput.*, vol. 121, 2022, Art. no. 104422.
- [94] H. Wang, H. Yu, and Q. Zhang, "Human-object interaction detection via global context and pairwise-level fusion features integration," *Neural Netw.*, vol. 170, pp. 242–253, 2024.
- [95] H.-S. Fang, J. Cao, Y.-W. Tai, and C. Lu, "Pairwise body-part attention for recognizing human-object interactions," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 51–67.
- [96] G. Lee, K. Yun, and J. Cho, "Improved human-object interaction detection through on-the-fly stacked generalization," *IEEE Access*, vol. 9, pp. 34251–34263, 2021.
- [97] X. Wu, Y.-L. Li, X. Liu, J. Zhang, Y. Wu, and C. Lu, "Mining cross-person cues for body-part interactiveness learning in HOI detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 121–136.
- [98] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Represent.*, 2018. [Online]. Available: <https://openreview.net/forum?id=rJXMpkCZ>
- [99] A. Nicolicchio, I. Duta, and M. Leordeanu, "Recurrent space-time graph neural networks," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 12838–12850.
- [100] O. Köpüklü, F. Herzog, and G. Rigoll, "Comparative analysis of CNN-based spatiotemporal reasoning in videos," in *Proc. Int. Conf. Pattern Recognit.*, 2021, pp. 186–202.
- [101] R. Li, M. Tapaswi, R. Liao, J. Jia, R. Urtasun, and S. Fidler, "Situation recognition with graph neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4173–4182.
- [102] Y. Yuan, X. Liang, X. Wang, D.-Y. Yeung, and A. Gupta, "Temporal dynamic graph LSTM for action-driven video object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1801–1810.
- [103] D. Teney, L. Liu, and A. Van Den Hengel, "Graph-structured representations for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1–9.
- [104] D. Yang and Y. Zou, "A graph-based interactive reasoning for human-object interaction detection," 2020, *arXiv:2007.06925*.
- [105] F. Z. Zhang, D. Campbell, and S. Gould, "Spatially conditioned graphs for detecting human-object interactions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13319–13327.
- [106] T. He, L. Gao, J. Song, and Y.-F. Li, "Exploiting scene graphs for human-object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15984–15993.
- [107] H. Wang et al., "IPGN: Interactiveness proposal graph network for human-object interaction detection," *IEEE Trans. Image Process.*, vol. 30, pp. 6583–6593, 2021.
- [108] Z. Liang, J. Liu, Y. Guan, and J. Rojas, "Visual-semantic graph attention networks for human-object interaction detection," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, 2021, pp. 1441–1447.
- [109] M. Zhu, E. S. Ho, and H. P. Shum, "A skeleton-aware graph convolutional network for human-object interaction detection," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2022, pp. 275–281.
- [110] H. Liu, T.-J. Mu, and X. Huang, "Detecting human—object interaction with multi-level pairwise feature network," *Comput. Vis. Media*, vol. 7, no. 2, pp. 229–239, 2021.
- [111] H.-S. Fang, Y. Xie, D. Shao, and C. Lu, "DIRV: Dense interaction region voting for end-to-end human-object interaction detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 2, pp. 1291–1299.
- [112] X. Zhong, C. Ding, Z. Li, and S. Huang, "Towards hard-positive query mining for DETR-based human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 444–460.
- [113] F. Z. Zhang, D. Campbell, and S. Gould, "Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20104–20112.
- [114] H. Yuan, M. Wang, D. Ni, and L. Xu, "Detecting human-object interactions with object-guided cross-modal calibrated semantics," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 3, pp. 3206–3214.

- [115] D. Zhou et al., "Human-object interaction detection via disentangled transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19568–19577.
- [116] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, and C.-W. Chen, "Exploring structure-aware transformer over interaction proposals for human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19548–19557.
- [117] J. Park, S. Lee, H. Heo, H. K. Choi, and H. J. Kim, "Consistency learning via decoding path augmentation for transformers in human object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1019–1028.
- [118] Z. Li, C. Zou, Y. Zhao, B. Li, and S. Zhong, "Improving human-object interaction detection via phrase learning and label composition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 1509–1517.
- [119] S. Chan, W. Wang, Z. Shao, and C. Bai, "SGPT: The secondary path guides the primary path in transformers for HOI detection," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 7583–7590.
- [120] L. Dong et al., "Category-aware transformer network for better human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19538–19547.
- [121] X. Qu, C. Ding, X. Li, X. Zhong, and D. Tao, "Distillation using oracle queries for transformer-based human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19558–19567.
- [122] H. Zhang, S. Wan, W. Guo, P. Jin, and M. Zheng, "HOD: Human-object decoupling network for HOI detection," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2023, pp. 2219–2224.
- [123] S. Ma, Y. Wang, S. Wang, and Y. Wei, "FGAHOI: Fine-grained anchors for human-object interaction detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2415–2429, Apr. 2024.
- [124] S. Chan, W. Wang, Z. Shao, Z. Wang, and C. Bai, "Region mining and refined query improved HOI detection in transformer," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 6, pp. 4237–4249, Dec. 2024.
- [125] Y. Cheng, H. Duan, C. Wang, and Z. Chen, "Parallel disentangling network for human-object interaction detection," *Pattern Recognit.*, vol. 146, 2024, Art. no. 110021.
- [126] A. Iftekhar, S. Kumar, R. A. McEver, S. You, and B. Manjunath, "GTNet: Guided transformer network for detecting human-object interactions," *Proc. SPIE*, vol. 12527, pp. 192–205, 2023.
- [127] J. Ji, R. Desai, and J. C. Niebles, "Detecting human-object relationships in videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8106–8116.
- [128] S. Wang, Y. Duan, H. Ding, Y.-P. Tan, K.-H. Yap, and J. Yuan, "Learning transferable human-object interaction detector with natural language supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 939–948.
- [129] B. Kim, J. Mun, K.-W. On, M. Shin, J. Lee, and E.-S. Kim, "MSTR: Multi-scale transformer for end-to-end human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19578–19587.
- [130] S. Kim, D. Jung, and M. Cho, "Relational context learning for human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2925–2934.
- [131] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
- [132] J. Park, J.-W. Park, and J.-S. Lee, "ViPLO: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17152–17162.
- [133] Y. Wang, Q. Liu, and Y. Lei, "TED-Net: Dispersal attention for perceiving interaction region in indirectly-contact HOI detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 5603–5615, Jul. 2024.
- [134] L. Xia and X. Ding, "Human-object interaction detection based on cascade multi-scale transformer," *Appl. Intell.*, vol. 54, pp. 2831–2850, 2024.
- [135] L. Li, J. Wei, W. Wang, and Y. Yang, "Neural-logic human-object interaction detection," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2024, vol. 36, pp. 21158–21171.
- [136] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.
- [137] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
- [138] T. Brown et al., "Language models are few-shot learners," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 1877–1901.
- [139] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 27730–27744.
- [140] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.
- [141] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19730–19742.
- [142] S. Zhang, Z. Zhu, and Q. Bao, "Rb-PaStaNet: A few-shot human-object interaction detection based on rules and part states," 2020, *arXiv:2008.06285*.
- [143] S. Wang, K.-H. Yap, J. Yuan, and Y.-P. Tan, "Discovering human interactions with novel objects via zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11652–11661.
- [144] M. Jiang, M. Li, J. Ren, and W. Huang, "HOICS: Zero-shot HOI detection via compatibility self-learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 1–5.
- [145] X. Liu, Z. Ji, Y. Pang, J. Han, and X. Li, "DGIG-Net: Dynamic graph-in-graph networks for few-shot human-object interaction," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 7852–7864, Aug. 2022.
- [146] A. Bansal, S. S. Rambhatla, A. Shrivastava, and R. Chellappa, "Detecting human-object interactions via functional generalization," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 07, pp. 10460–10469.
- [147] Y. Liu, J. Yuan, and C. W. Chen, "ConsNet: Learning consistency graph for zero-shot human-object interaction detection," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 4235–4243.
- [148] X. Liu, Y.-L. Li, and C. Lu, "Highlighting object category immunity for the generalization of human-object interaction detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 1819–1827.
- [149] Z. Hou, B. Yu, Y. Qiao, X. Peng, and D. Tao, "Affordance transfer learning for human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 495–504.
- [150] D. Huynh and E. Elhamifar, "Interaction compass: Multi-label zero-shot learning of human-object interactions via spatial relations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8472–8483.
- [151] H. Jiang, X. Ma, W. Nie, Z. Yu, Y. Zhu, and A. Anandkumar, "Bongard-HOI: Benchmarking few-shot visual reasoning for human-object interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19056–19065.
- [152] Z. Hou, B. Yu, and D. Tao, "Discovering human-object interaction concepts via self-compositional learning," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 461–478.
- [153] F. Baldassarre, K. Smith, J. Sullivan, and H. Azizpour, "Explanation-based weakly-supervised learning of visual relations with graph networks," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 612–630.
- [154] S. K. Kumaraswamy, M. Shi, and E. Kijak, "Detecting human-object interaction with mixed supervision," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1228–1237.
- [155] T. Nagarajan, C. Feichtenhofer, and K. Grauman, "Grounded human-object interaction hotspots from video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8688–8697.
- [156] Y.-L. Li et al., "Transferable interactiveness knowledge for human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3585–3594.
- [157] D.-J. Kim, X. Sun, J. Choi, S. Lin, and I. S. Kweon, "Detecting human-object interactions with action co-occurrence priors," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 718–736.
- [158] J. Gao, K. Liang, T. Wei, W. Chen, Z. Ma, and J. Guo, "Dual-prior augmented decoding network for long tail distribution in HOI detection," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 3, pp. 1806–1814.
- [159] Y. Liao, A. Zhang, M. Lu, Y. Wang, X. Li, and S. Liu, "GEN-VLK: Simplify association and enhance interaction understanding for HOI detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20123–20132.
- [160] M. Wu, J. Gu, Y. Shen, M. Lin, C. Chen, and X. Sun, "End-to-end zero-shot HOI detection via vision and language knowledge distillation," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 3, pp. 2839–2846.
- [161] S. Ning, L. Qiu, Y. Liu, and X. He, "HOICLIP: Efficient knowledge transfer for HOI detection with vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23507–23517.
- [162] B. Wan, Y. Liu, D. Zhou, T. Tuytelaars, and X. He, "Weakly-supervised HOI detection via prior-guided bi-level representation learning," 2023, *arXiv:2303.01313*.
- [163] B. Wan and T. Tuytelaars, "Exploiting clip for zero-shot HOI detection requires knowledge distillation at multiple levels," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 1805–1815.

- [164] M. E. Unal and A. Kovashka, "Weakly-supervised HOI detection from interaction labels only and language/vision-language priors," 2023, *arXiv:2303.05546*.
- [165] S. Gupta and J. Malik, "Visual semantic role labeling," 2015, *arXiv:1505.04474*.
- [166] Y.-W. W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "HICO: A benchmark for recognizing human-object interactions in images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1017–1025.
- [167] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. Van den Hengel, "HCVRD: A benchmark for large-scale human-centered visual relationship detection," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, pp. 7631–7638, 2018.
- [168] Y.-L. Li et al., "Hake: Human activity knowledge engine," 2019, *arXiv:1904.06539*.
- [169] Y.-L. Li et al., "Detailed 2D-3D joint representation for human-object interaction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10166–10175.
- [170] B. L. Bhatnagar, X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "Behave: Dataset and method for tracking human object interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15935–15946.
- [171] Y.-L. Li, X. Liu, X. Wu, X. Huang, L. Xu, and C. Lu, "Transferable interactiveness knowledge for human-object interaction detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 07, pp. 3870–3882, Jul. 2022.
- [172] Y. Chen, S. K. Dwivedi, M. J. Black, and D. Tzionas, "Detecting human-object contact in images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17100–17110.
- [173] Y. Mao, J. Deng, W. Zhou, L. Li, Y. Fang, and H. Li, "CLIP4HOI: Towards adapting clip for practical zero-shot HOI detection," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2024, vol. 36, pp. 45895–45906.
- [174] Y. Cao et al., "Detecting any human-object interaction relationship: Universal HOI detector with spatial prompt learning on foundation models," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2024, vol. 36, pp. 739–751.
- [175] Q. Lei, B. Wang, and R. T. Tan, "Few-shot learning from augmented label-uncertain queries in bongard-HOI," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 4, pp. 2974–2982.
- [176] J. Gao, K.-H. Yap, K. Wu, D. T. Phan, K. Garg, and B. S. Han, "Contextual human object interaction understanding from pre-trained large language model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 13436–13440.
- [177] T. Lei, S. Yin, and Y. Liu, "Exploring the potential of large foundation models for open-vocabulary HOI detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16657–16667.
- [178] W. Xue et al., "Towards zero-shot human-object interaction detection via vision-language integration," 2024, *arXiv:2403.07246*.
- [179] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.
- [180] Y.-L. Li et al., "PaStaNet: Toward human activity knowledge engine," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 382–391.
- [181] S. Pratt, M. Yatskar, L. Weihs, A. Farhadi, and A. Kembhavi, "Grounded situation recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 314–332.
- [182] D. Shan, J. Geng, M. Shu, and D. F. Fouhey, "Understanding human hands in contact at internet scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9869–9878.
- [183] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–16, 2015.
- [184] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black, "Resolving 3D human pose ambiguities with 3D scene constraints," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2282–2292.
- [185] C. Wu, J. Zhang, S. Savarese, and A. Saxena, "Watch-N-patch: Unsupervised understanding of actions and relations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4362–4370.
- [186] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.
- [187] J. Zhao, F. Deng, H. He, and J. Chen, "Local domain adaptation for cross-domain activity recognition," *IEEE Trans. Hum.-Mach. Syst.*, vol. 51, no. 1, pp. 12–21, Feb. 2021.
- [188] J. Zhao, L. Li, F. Deng, H. He, and J. Chen, "Discriminant geometrical and statistical alignment with density peaks for domain adaptation," *IEEE Trans. Cybern.*, vol. 52, no. 2, pp. 1193–1206, Feb. 2022.
- [189] M. Hu et al., "Physiological characteristics inspired hidden human object detection model," *Displays*, vol. 81, 2024, Art. no. 102613.
- [190] N. Wiederhold, A. Megyeri, D. Paris, S. Banerjee, and N. Banerjee, "HOH: Markerless multimodal human-object-human handover dataset with large object count," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2024, vol. 36, pp. 68736–68748.
- [191] F. Ragusa, A. Furnari, S. Livatino, and G. M. Farinella, "The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1569–1578.
- [192] R. Leonardi, F. Ragusa, A. Furnari, and G. M. Farinella, "Exploiting multimodal synthetic data for egocentric human-object interaction detection in an industrial scenario," *Comput. Vis. Image Understanding*, vol. 242, 2024, Art. no. 103984.
- [193] S. Li, Y. Du, A. Torralba, J. Sivic, and B. Russell, "Weakly supervised human-object interaction detection in video via contrastive spatiotemporal regions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1845–1855.
- [194] D. Damen et al., "Scaling egocentric vision: The epic-kitchens dataset," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 720–736.
- [195] X. Shang, T. Ren, J. Guo, H. Zhang, and T.-S. Chua, "Video visual relation detection," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 1300–1308.
- [196] X. Shang, D. Di, J. Xiao, Y. Cao, X. Yang, and T.-S. Chua, "Annotating objects and relations in user-generated videos," in *Proc. Int. Conf. Multimedia Retrieval*, 2019, pp. 279–287.
- [197] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 7590–7598.
- [198] L. Zhou, N. Louis, and J. J. Corso, "Weakly-supervised video object grounding from text by loss weighting and object interaction," 2018, *arXiv:1805.02834*.
- [199] M.-J. Chiou, C.-Y. Liao, L.-W. Wang, R. Zimmermann, and J. Feng, "ST-HOI: A spatial-temporal baseline for human-object interaction detection in videos," in *Proc. Workshop Intell. Cross-Data Anal. Retrieval*, 2021, pp. 9–17.
- [200] X. Xu, H. Joo, G. Mori, and M. Savva, "D3D-HOI: Dynamic 3D human-object interactions from videos," 2021, *arXiv:2108.08420*.
- [201] Y. Liu et al., "HOI4D: A 4D egocentric dataset for category-level human-object interaction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21013–21022.
- [202] L. Liu, J. Tian, Z. Shi, J. Fan, and Y. Rui, "Multi-sensor fusion for multi-target detection and tracking," in *Autonomous Vehicles and Systems*. Aalborg, Denmark: River Publishers, 2024, pp. 175–217.
- [203] S. Wang et al., "Multi-modal fusion sensing: A comprehensive review of millimeter-wave radar and its integration with other modalities," *IEEE Commun. Surveys Tuts.*, early access, May 8, 2024, doi: 10.1109/COMST.2024.3398004.
- [204] Y. Li et al., "Fully sparse fusion for 3D object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 11, pp. 7217–7231, Nov. 2024.
- [205] Y. Jin et al., "Audio matters too! Enhancing markerless motion capture with audio signals for string performance capture," *ACM Trans. Graph.*, vol. 43, no. 4, pp. 1–10, 2024.
- [206] J. Zhao, T. Yu, L. An, Y. Huang, F. Deng, and Q. Dai, "Triangulation residual loss for data-efficient 3D pose estimation," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2024, vol. 36, pp. 12721–12732.
- [207] J. Li, A. Clegg, R. Mottaghi, J. Wu, X. Puig, and C. K. Liu, "Controllable human-object interaction synthesis," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2025, pp. 54–72.
- [208] X. Peng, Y. Xie, Z. Wu, V. Jampani, D. Sun, and H. Jiang, "HOI-Diff: Text-driven synthesis of 3D human-object interactions using diffusion models," 2023, *arXiv:2312.06553*.
- [209] J. Cha, J. Kim, J. S. Yoon, and S. Baek, "Text2HOI: Text-guided 3D motion generation for hand-object interaction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 1577–1585.
- [210] J. Braun, S. Christen, M. Kocabas, E. Aksan, and O. Hilliges, "Physically plausible full-body hand-object interaction synthesis," in *Proc. Int. Conf. 3D Vis.*, 2024, pp. 464–473.

2219
2220
2221
2222
2223
2224
2225
2226

Geng Han received the B.E. degree in automation in 2021 from the Beijing Institute of Technology, Beijing, China, where he is currently working toward the Ph.D. degree in control science and engineering. His researches include HOI detection, foundation models, visual language navigation, and human-robot interactions.

2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240

Jiachen Zhao received the B.Eng. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the Beijing Institute of Technology, Beijing, China, in 2016 and 2022, respectively. He is currently as a Postdoctoral Researcher with the Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China. From 2018 to 2020, he was a Visiting Researcher with the Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island, Kingston, RI, USA. His research interests include spatiotemporal data mining, intelligent machine vision, and their practical applications.



Lele Zhang (Member, IEEE) received the B.E. degree in automation from Inner Mongolia University, Inner Mongolia Autonomous Region, China, in 2013, and the Ph.D. degree in control science and engineering from the Beijing Institute of Technology, Beijing, China, in 2019. From November 2016 to November 2017, he engaged in research work as a Visiting Ph.D. degree Student with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, and as a Postdoctoral Researcher in control science and engineering from the Beijing Institute of Technology, from January 2021 to January 2023. He is currently an Assistant Professor with the School of Automation, Beijing Institute of Technology. His research interests include camera calibration, 3D reconstruction and deep learning, and specifically in the area of aerial and ground robotics.



Fang Deng (Senior Member, IEEE) received the B.E. and Ph.D. degrees in control science and engineering from the Beijing Institute of Technology, Beijing, China, in 2004 and 2009, respectively. He is currently a Professor with the School of Automation, Beijing Institute of Technology. His research interests include intelligent fire control, intelligent information processing, and smart wearable devices.

2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
22562257
2258
2259
2260
2261
2262
2263
2264
2265