Neural Networks 170 (2024) 242–253

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

# Human–Object Interaction detection via Global Context and Pairwise-level Fusion Features Integration

Haozhong Wang[1], Hua Yu[1], Qiang Zhang *

*Dalian University of Technology, Dalian, 116024, Liaoning, China*

ARTICLE INFO

*Keywords:*
Human–object interaction
Two-stage detector
Global context
Pairwise-level attention

ABSTRACT

Recent two-stage detector-based methods show superiority in Human-Object Interaction (HOI) detection along with the successful application of transformer. However, these methods are limited to extracting the global contextual features through instance-level attention without considering the perspective of human–object interaction pairs, and the fusion enhancement of interaction pair features lacks further exploration. The human–object interaction pairs guiding global context extraction relative to instance guiding global context extraction more fully utilize the semantics between human–object pairs, which helps HOI recognition. To this end, we propose a two-stage **G**lobal Context and Pairwise-level **F**usion Features **I**ntegration **N**etwork (GFIN) for HOI detection. Specifically, the first stage employs an object detector for instance feature extraction. The second stage aims to capture the semantic-rich visual information through the proposed three modules, Global Contextual Feature Extraction Encoder (GCE), Pairwise Interaction Query Decoder (PID), and Human-Object Pairwise-level Attention Fusion Module (HOF). The GCE module intends to extract the global context memory by the proposed crossover-residual mechanism and then integrate it with the local instance memory from the DETR object detector. HOF utilizes the proposed pairwise-level attention mechanism to fuse and enhance the first stage's multi-layer feature. PID outputs multi-label interaction recognition results with the input of the query sequence from HOF and the memory from GCE. Finally, comprehensive experiments conducted on HICO-DET and V-COCO datasets demonstrate that the proposed GFIN significantly outperforms the state-of-the-art methods. Code is available at https://github.com/ddwhzh/GFIN.

## 1. Introduction

Human-Object Interaction (HOI) detection is intended to detect interactive relations between humans and their surroundings by simplifying their visual relationships as a triple <human, interaction, object>. As a fundamental visual understanding task, HOI detection significantly advances research from perception to comprehension. It involves a wide range of research fields, such as action recognition (Sun et al., 2022; Xu, Ye, Zhong, & Xie, 2022) and visual question answering (Lee, Cheon, & Han, 2021; Zheng et al., 2021). Unfortunately, these methods do not provide a straightforward method for constructing a comprehensive framework to extract global semantic information from established interaction pairs. Additionally, there is a lack of guidance on creating a downstream network to merge the human–object features obtained in the first stage into interaction pairs. The human–object interaction pairs guiding global context extraction relative to instance guiding global context extraction more fully utilize the semantics between human–object pairs, which helps HOI recognition.

Existing HOI detection methods can be divided into one-stage and two-stage methods. One-stage methods (Chen et al., 2021; Liao et al., 2020) optimize both object detection and interaction recognition, depending on pre-specified matching algorithms for triplet relationship extraction, e.g., point matching (Zhong, Qu, Ding, & Tao, 2021), bipartite matching algorithm (Tamura, Ohashi, & Yoshinaga, 2021). Two-stage methods (Gao, Xu, Zou, & Huang, 2020; Zhang, Campbell & Gould, 2022) typically employ a pre-trained object detector to obtain a set of human–object instance results and instance features, which are paired up and processed by a downstream network for multi-label interaction recognition. Both methods have made potential progress with the application of transformer in vision (Kim, Lee, Kang, Kim, & Kim, 2021). However, there are still open challenges for HOI detection tasks. As shown in Fig. 1, one-stage HOI detection is a multi-objective optimization task that is required to optimize at least four sub-tasks, bounding boxes of people, bounding boxes of objects, classes of objects, and classes of action interactions (Tamura et al., 2021). The

* Corresponding author.
*E-mail addresses:* dlwhzh@mail.dlut.edn.cn (H. Wang), yhiccd@mail.dlut.edu.cn (H. Yu), zhangq@dlut.edu.cn (Q. Zhang).
[1] Co-author.

https://doi.org/10.1016/j.neunet.2023.11.002
Received 14 November 2022; Received in revised form 24 September 2023; Accepted 1 November 2023
Available online 13 November 2023

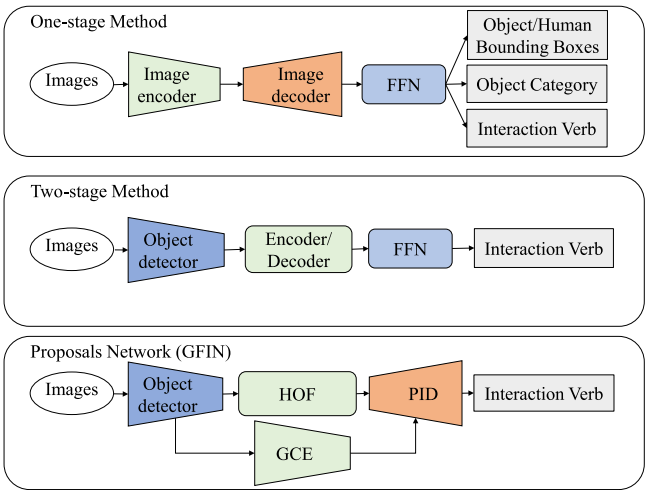


**Fig. 1.** Comparisons of conventional one-stage detectors, two-stage detectors, and our proposed GFIN. One-stage detectors adopt encoder–decoder framework to detect HOI triplets directly. Two-stage detectors utilize instance features to achieve interaction recognition by a downstream network, i.e., an encoder or a decoder. Our proposal introduces the encoder–decoder framework to two-stage methods, which aim to extract global contextual information.

obtained results tend to be converged to a sub-optimal solution due to inconsistent optimization directions for each sub-objective. In addition, most current one-stage detectors (Qu, Ding, Li, Zhong, & Tao, 2022) converge slowly and need to initialize the network with the parameters of End-to-End Object Detection with Transformers(DETR) (Carion et al., 2020) during the training process. To avoid issues introduced by multi-objective optimization, two-stage methods convert the detection problem into capturing the interaction of the human object. Nevertheless, many two-stage methods (Li et al., 2019; Zhang, Campbell & Gould, 2021) assume that the external environmental features have minimal influence on interaction recognition and ignore the contextual information.

In addition, interaction tasks are not limited to the range of human–object locations and can be divided into direct and indirect contact actions. Some direct contact actions can be inferred from the surrounding environment. For example, the interaction between a human and his surfboard varies by context. Surfing can only be in the sea, so it is more likely to be holding a surfboard if the scene is the beach. The indirect contact actions are not within the bounding box of human–object, which requires additional inferential features. For example, matching a human and a kite should be more concerned with the kite string between them. Usually, the kite string is not wrapped within the human–object's bounding box. Some existing methods, such as iCAN (Gao, Zou, & Huang, 2018) TIN (Li et al., 2019), have considered introducing contextual information as a complement to human-stream or object-stream in multi-stream networks. However, it is usually used to complement a separate human stream and object stream feature rather than thinking about the whole in terms of human–object interaction. This leads to previous approaches that use contextual features to enhance the representation of a particular human or object pair for identifying interactions rather than extracting contextual features in the context of a particular human–object interaction pair. However, the mentioned methods lack semantic information between interaction pairs during the extraction of global context. In contrast, methods such as UPT (Zhang, Campbell et al., 2022) directly connect global features by pooling. However, the pooling operations cause a lack of spatial semantics in the contextual information, making the contextual semantic information too simple. Therefore, this paper discusses how to construct the overall structure for extracting global semantic information based on known interaction pairs and how to design downstream networks that fuse the human–object features provided in the first stage into interactive pairs. Taking these inspirations, we propose a novel two-stage Global Context and Pairwise-level Fusion Features Integration Network (GFIN) for Human-Object Interaction Detection for extracting context features through the mechanism of human–object interaction pairwise-level attention, and fusing human–object interaction pair features. In the first stage, a frozen pre-trained object detector is deployed to extract instance features, the bounding box, and the category results. The downstream network process instance features for interaction recognition. The second stage aims to achieve the fusion of instance feature pairs and extraction of global contextual features, which comprises three components. Specifically, Global Contextual Feature Extraction Encoder (GCE) is designed as an encoder to extract the global context memory. GCE enables the fusion of the local instance memory and the global context memory to mine features from both global and local perspectives through the crossover-residual mechanism, where the local instance memory is acquired by the encoder of the one-stage object detector. Moreover, Human-object Pairwise-level Attention Fusion (HOF) module with the pairwise-level attention mechanism generates pairwise-level HOI instance features as the query sequence of PID. Finally, with the input of query sequence and memory from the GCE, the Pairwise Interaction Query Decoder (PID) is designed as a decoder to output the interaction recognition of HOI instance pairs. Our contributions are summarized as follows:

- We propose a novel two-stage HOI detection architecture, which uses an encoder to implement global context memory extraction and a decoder to transform query sequences representing pairs of human–object into features that can recognize interactions.
- We exploit the prediction results of the DETR multi-layer decoder and propose the pairwise-level attention mechanism, which can adaptively learn the relationships between humans and objects.
- Comprehensive experiments are conducted on the HICO-DET and the V-COCO datasets. It can be demonstrated that our proposed GFIN achieves superior performance over the existing HOI detection methods.

The remainder of the paper is arranged as follows. Some preliminaries and the problem formulation are presented in Section 2. The proposed method and the process of training and inference are detailed in Section 3. The experimental results and analysis are presented in Section 4. The conclusion and the limitation of our work are described in Section 5.

## 2. Related work

We briefly cover the works most related to the proposed method in this section, including one-stage, two-stage methods and contextual methods.

### 2.1. One-stage methods

The one-stage approach takes full advantage of the capabilities of the neural network structure to accomplish multiple subtasks simultaneously. The pioneering work is UnionDet (Kim, Choi, Kang, & Kim, 2020) proposed by Kim et al. It achieves interaction detection by directly detecting external union regions containing people and objects. And it introduces the concept of matching confidence for the first time to realize how to determine the interactions between them when there are multiple objects in the external union region. IP-NET (Wang et al., 2020) and PPDM (Liao et al., 2020) were proposed almost simultaneously. They argue that the human–object interaction can be represented by the center of the connecting line between the center of the subject person and the center of the object. Meanwhile, the matching with subject person, object can be solved by using the offset of regression subject person and object center over the interaction

center. The difference mainly stems from the definition of the offset. The work of IP-NET considers that since the interaction representation point is the center of the line connecting the center of the subject and the center of the object, the offset can be represented by a vector of the same length and opposite direction, and only two values are regressed. While PPDM considers that the actual detected interaction center is not perfectly located at the center, it proposes that the offset of the subject and object should be represented by two vectors, and four values are regressed values. Experimental results show that PPDM gives superior results. Fang et al. propose the Dense Interaction Region Voting (DIRV) (Fang, Xie, Shao, & Lu, 2021) algorithm for end-to-end human–object interaction detection based on the EfficientDet (Tan, Pang, & Le, 2020) architecture and accepted the anchor design. Their approach focuses on densely sampled interaction regions at different scales for each human–object pair in order to capture the subtle visual features that are most important for the interaction. In addition, to compensate for the detection deficiency of a single interaction region, a new voting strategy is introduced that makes full use of those overlapping interaction regions instead of the traditional non-maximal suppression (NMS). Zhong et al. propose a novel one-stage approach, the Glance and Gaze Network (GGNet) (Zhong et al., 2021), which perceives a set of action points through glance and gaze steps for adaptive modeling. Meanwhile, the experimental results of GGNet show that the main reason why the one-stage human–object interaction detection method is weaker than the two-stage human–object interaction detection method comes from the low object detection ability.

The successful application of DETR in human–object interaction has inspired many researchers. For example, HOI-Trans (Zou et al., 2021) and QPIC (Tamura et al., 2021) added additional detection heads and relied on the bipartite matching algorithm to localize HOI instance pairs and recognize the interaction. However, sharing structures weaken the ability to learn from multiple tasks. AS-NET (Chen et al., 2021) proposed the instance-aware attention module to introduce the interaction-relevant instance features from the instance branch to the interaction branch. HOTR (Kim et al., 2021) proposed HO Pointers to associate outputs of two parallel decoders. Meanwhile, CDN (Zhang, Liao, Liu, Lu, Wang, Gao, et al., 2021) divided HOI detection into two cascade transformer-based decoders by initializing the Interaction Decoder's query sequence with the output of the Human-Object Pair Decoder (HO-PD). DT (Zhou, Liu, Wang, Wang, Hu, Ding, et al., 2022) went one step further by decoupling both the encoder and decoder, generating a unified representation for HOI triplets with a base decoder, and then utilizing it as input features of each disentangled decoder to predict results. HQM (Zhong, Ding, Li, & Huang, 2022) explicitly composes hard-positive queries according to the ground-truth position of labeled human–object pairs for each training image and proposes an alternate strategy that efficiently combines both types of hard queries and gets great results. The above methods note the multi-objective optimization problem of HOI detection and progressive decouple the encoders and the decoders to alleviate the problem. However, the training process is still multi-task in parallel, which relies heavily on pre-trained parameters of the object detector. This operation poses a significant challenge for HOI detection task.

### 2.2. Two-stage methods

Two-stage methods make the training process more stable by focusing on one optimization direction. The pioneer of HOI detection based on two-stage methods is the HORCNN (Chao, Liu, Liu, Zeng, & Deng, 2018). HORCNN is a two-stage three-stream network. The proposal of HORCNN is obtained from the existing object detector in one stage. The extracted features from three streams (i.e., the human stream, object stream, and pairwise stream) are summed up for HOI detection. InteractNet (Gkioxari, Girshick, Dollár, & He, 2018) aims to detect (human, interaction, object) triplets by fusing the predictions of an action-specific density over target object locations based on the appearance of a detected person. FCL (Hou, Yu, Qiao, Peng, & Tao, 2021) introduces an object fabricator to generate effective object representations, and then combines verbs and fabricated objects to compose new HOI samples. The two-stage method has achieved a new push based on the transformer method. STIP (Zhang, Pan et al., 2022) decomposed the process of HOI set prediction into two subsequent stages. Specifically, the first stage produces interaction proposals via the Interaction Proposal Network and the second stage maps interaction proposal queries into HOI predictions. IF (Liu et al., 2022) introduced Interactiveness Field to model the interactiveness distribution of an image and assigned a score for each pair. UPT (Zhang, Campbell et al., 2022) encoded the instance information into a unary and pairwise representation through the cooperative and competitive encoder to predict interaction. However, these methods ignore or trivialize global contextual features of the input, which is of significance to infer the human–object locations in two-stage detectors.

### 2.3. Contextual methods

There are many approaches to introduce contextual features in two-stage methods, but they focus on the instance-level feature stage and lack methods under matching pair constraints. iCAN (Gao et al., 2018) is the first approach to consider introducing attention mechanisms in HOI tasks. Its main idea is derived from a three-stream network and uses the proposed attention-based mechanism for prediction on a stream-by-stream basis. Its instance-based contextual attention mainly assists the person's features or object's features alone for inference. Also, TIN (Li et al., 2019) proposes a similar approach that uses the global features output from the backbone with pooled person and object features for attention operations. However, such methods lack the global attention mechanism under the matching pair constraint because there are many-to-many matches for people and objects, and the instance-level contextual attention mechanism treats the interacting objects/people as equal to the interaction context, which lacks the semantic information of the corresponding interaction instances and leads to confusion in contextual reasoning. Some other approaches use pooling to use scene features as additional features to assist inference. Both RPNN (Zhou & Chi, 2019) and UPT (Zhang, Campbell et al., 2022) use adaptive average pooling to obtain scene features. Still, the pooling operation destroys the features' spatial structure, leading to the need for more contextual semantics.

To tackle the limitation of the existing methods, we propose a transformer-based encoder architecture to extract global contextual memory for shallow features of images, which discards the traditional pooling operation (Zhang, Campbell et al., 2022) or the instance context methods (Gao et al., 2018). Meanwhile, the query sequence of the PID decoder uses instance-fused interaction pair features so that each interaction pair's attention feature map is entirely different. It can be demonstrated that the proposed method can promote the performance of the existing HOI detection models for a large margin.

## 3. Methodology

This section will briefly introduce the details of the proposed Global Context and Pairwise-level Fusion Features Integration Network (GFIN) for HOI detection. Sections 3.1–3.3 first explains how the encoder–decoder architecture inspired the multi-label interaction recognition for interaction pairs and outlines the pipeline of GFIN. Section 3.4 describes the proposed pairwise-level attention mechanism and the process of the designed HOF module. Section 3.5 explains the details of training and inference of the proposed method.

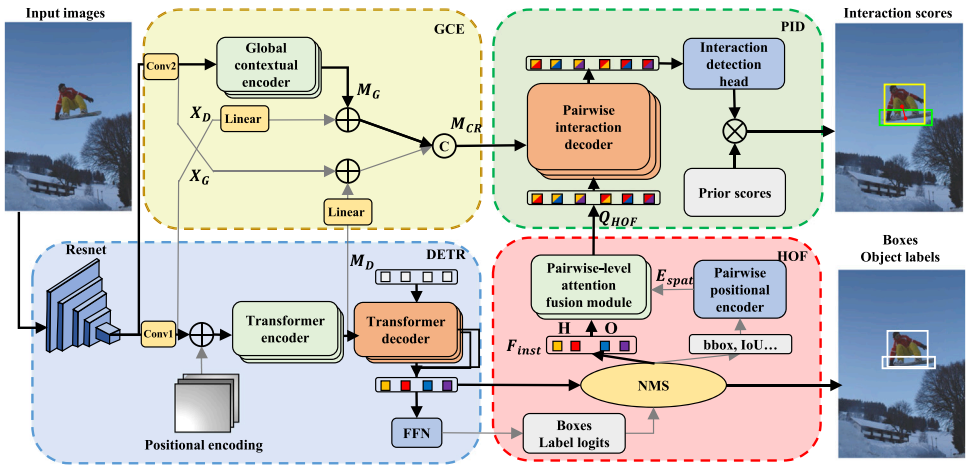


**Fig. 2.** Overview of GFIN pipeline. We first extract the instance multi-layer features by DETR, and then fuse them by the proposed pairwise-level attention mechanism of HOF. The GCE module intends to extract the global context memory by the proposed crossover-residual mechanism and then integrate it with the local instance memory from the DETR object detector. Finally, the HOI triplet is obtained by the proposed PID module.

### 3.1. Overall pipeline

The overall architecture of GFIN is shown in Fig. 2. The GFIN consists of a DETR object detector and three proposed modules: GCE, HOF, and PID. The Global Contextual Feature Extraction Encoder (GCE) is designed to extract global context memory from images and fuse it with the local instance memory from the encoder of DETR. The Human-object Pairwise-level Attention Fusion (HOF) module encodes the location information obtained in the first stage into a query sequence and instance features into keys and values. The pairwise-level attention mechanism is proposed to generate the interaction pair's features. In addition, the Pairwise Interaction Query Decoder (PID) takes the memory output of GCE as memory input and the pairwise fusion features as query sequence input, and outputs multi-label interaction recognition results of each interaction pair. Memory is the output of the encoder, which can be considered as a recoding of the input visual features using the self-attention mechanism. And memory is provided to the decoder for cross-attention operations.

### 3.2. Global contextual feature extraction encoder

Conventional two-stage detectors are limited to processing the first stage's instance features and location information but ignore areas containing contextual features which are crucial for reasoning interactions. Therefore, to take full advantage of capturing global features, the proposed Global Contextual Feature Extraction Encoder (GCE) module is designed. Although specific supervised signals do not guide this process, we are confident that GCE can extract global contextual features. Because it receives constraints on matching pairs of query features, is guided by loss functions, and searches for features that can aid inference in the domain of visual features. The visual features are not constrained to a specific interaction region. Moreover, GCE employs the shared frozen backbone of DETR since freezing the pre-trained backbone is a necessary procedure to preserve feature discrimination in multi-task training (Dai, Cai, Lin, & Chen, 2021), which facilitates training efficiency and stabilizes the convergence direction. The ablation experiment in Section 4.4 proved its efficiency. Also, we use the same vanilla transformer encoder as GCE and DETR.

We also explores how to fuse global contextual features with local instance features. It is noted that the data distributions of the input and output are different for the shallow transformer structures (Zhou et al., 2021). (We also confirmed this through visualization in Section 4.5.) Therefore, we propose the crossover-residual mechanism to achieve cross-branch and cross-layer feature fusion. We fuse the inputs and outputs of different encoders to accommodate better the combination of global contextual and local instance information. Referring to the Eq. (1), we denote the input and output features of DETR and GCE as $\mathbf{X_D}$, $\mathbf{X_G}$, $\mathbf{M_D}$ and $\mathbf{M_G}$. Typically, the output feature of ResNet in DETR has 2048 dimensions and is fed into a convolution layer (Conv1 in Fig. 2) to downsample the feature from 2048 dimensions to 256 dimensions. In this paper, we introduce another convolution layer (Conv2 in Fig. 2) and get the output feature $\mathbf{X_G}$, which is a 256-dimensional feature as same as $\mathbf{X_D}$ (i.e., the result of Conv1 output). Our GCE module uses the vanilla transformer encoder, whose output is $\mathbf{M_G}$ in 256 dimensions, while DETR's transformer encoder output is $\mathbf{M_D}$ in 256 dimensions. The crossover-residual mechanism employs a linear mapping on $\mathbf{X_D}$ and adds the output embedding to $\mathbf{M_G}$. $\mathbf{M_D}$ is also mapped by a linear projection and added to $\mathbf{X_G}$. $\mathbf{M_D}$ and $\mathbf{X_D}$ are fixed during the training process. We train them through the linear layers that implement a 256-dimensional to 256-dimensional mapping to enhance the learning effect. By concatenating these output features, a fully fused feature containing neglected information is obtained, namely $\mathbf{M_{CR}}$. The neglected information is the global contextual information that can be used to identify specific feature pair interactions, which some two-stage methods ignore. In contrast, the instance context methods (Gao et al., 2018) ignore the contextual information between feature pairs.

$$\mathbf{M_{CR}} = \mathrm{Cat}\begin{pmatrix} \mathbf{M_G} + Linear\left(\mathbf{X_D}\right), \\ Linear\left(\mathbf{M_D}\right) + \mathbf{X_G} \end{pmatrix} \tag{1}$$

### 3.3. Pairwise interaction query decoder

The decoder of the vanilla transformer employs a fixed-length blank query sequence as input and integrates with memory from the cross-attention mechanism.

The decoder follows the vanilla architecture of the transformer (Carion et al., 2020), transforming $N$ embeddings of size $d$ using multi-headed self- and encoder–decoder attention mechanisms. The proposed decoder, whose query sequence is pair of fusion instance features, integrates with the memory output of GCE through the cross-attention mechanism and refines the interaction recognition results. Also, considering the significant difference in the number of interaction pairs per image, we discard the bipartite matching algorithm in favor of a query sequence of variable length.

Meanwhile, query embedding is generally considered to contain spatial semantics (Meng et al., 2021; Wang, Zhang, Yang, & Sun, 2022). However, our query sequence from GCE modules already contains positional encoding in the GCE and HOF modules, referring to Section 3.2, so the PID drops the query embedding. Finally, multiple layers are retained to calculate the auxiliary loss for refining, respectively.

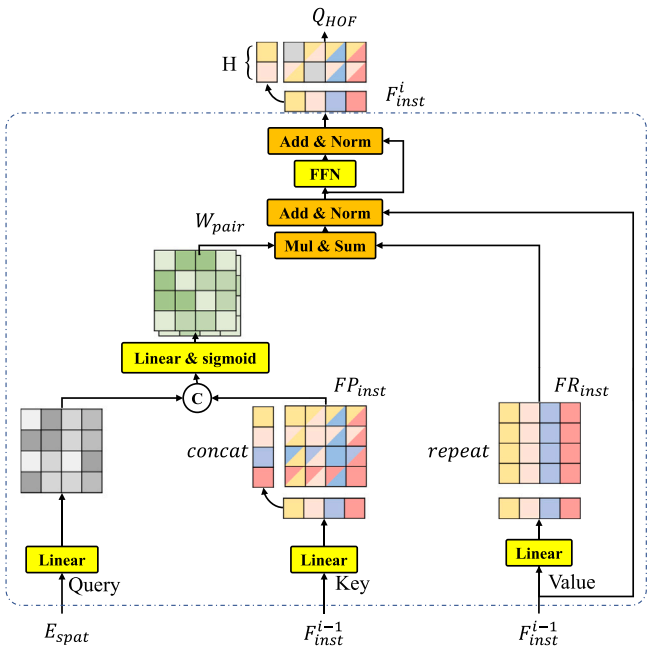


**Fig. 3.** Architecture of the HOF module and the pairwise-level attention mechanism.

Post-processing refers to the general practice in this field and obtains each HOI pair's interaction confidence by filtering through the existential matrix of object-action. The above process can be expressed by the following Eq. (2), where $\hat{\mathbf{S}}_{\mathbf{A}}$ represents the confidence of interaction actions(see Fig. 3).

$$\hat{\mathbf{S}}_{\mathbf{A}} = \text{Sigmoid}(\text{FFN}\left(\text{PID}\left(\mathbf{Q}_{\mathbf{HOF}}, \mathbf{M}_{\mathbf{CR}}\right)\right)) \tag{2}$$

### 3.4. Pairwise-level attention fusion module

The major challenge of a two-stage detector is how to correctly process the instance features provided by the first stage to fulfill combination and enhancement, which is crucial for improving recognition accuracy. The DETR detector is selected as the object detector, which is different from previous Faster-RCNN methods (Ren, He, Girshick, & Sun, 2015). Since the DETR decoder is trained with multiple layers and utilizes auxiliary loss, each layer of the DETR decoder can make predict results independently. While early layer prediction is full of low-quality predictions, providing richer prediction information can help improve recall and increase the generalization of the network. Nevertheless, concatenating the results of predicting through multiple layers directly will lead to the repeated prediction of the same instance, so we first apply NMS post-processing and threshold, leaving a set that consists of the box coordinates, the instance class, and features.

We believe that implicit knowledge can be learned among and between different matching pair instances, which can guide us in predicting human–object interaction detection. In order to enable the pairwise positional encoding with semantics to guide the fusion between instance features, the pairwise-level attention mechanism and the Human-object Pairwise-level Attention Fusion (HOF) module are proposed (in Fig. 3). The specific process refers to the Eq. (3)–(5), which is inspired by SENet (Hu, Shen, & Sun, 2018) and the cross-attention mechanism (Devlin, Chang, Lee, & Toutanova, 2018). In detail, refer to UPT, the pairwise positional encoding $\mathbf{E}_{\mathbf{spat}}$ consists of the center coordinates of the bounding box, the height and the width, the ratio of box areas, IoU and the distance between box centers. The pairwise positional encoding $\mathbf{E}_{\mathbf{spat}}$ with size $N \times N \times C$ represented as the query sequence of the HOF module, where $N$ represents the total number of instances, and $C$ represents the number of channels. Next, instance features called $\mathbf{F}_{\mathbf{inst}}$ whose size is $N \times C$ are represented as the key and the value of the HOF module. Unlike the cross-attention mechanism, the pairwise-level attention mechanism does not apply the traditional self-attention method. The HOF is designed to enhance the fusion of single instance features $\mathbf{F}_{\mathbf{inst}}$ into pairwise-level query features, so we introduce HOF to achieve early fusion between instance features. Hence, the operation of HOF is more similar to the fusion of features by attaching the inter-instance features and positional encoding to the original features through linear mapping as pairwise-level attention. Usually, the query input of the crossed self-attention mechanism is a blank or a query sequence containing semantics. At the same time, memory can be chosen from features of the same modality and different modalities. Assuming that we use the traditional cross-attention mechanism, we input the positional encoding $\mathbf{E}_{\mathbf{spat}}$ as the memory. The positional encoding $\mathbf{E}_{\mathbf{spat}}$ as an embedding of location information between each query lacks corresponding meaning as a memory input. Each instance is concatenated with all detected instances. The query sequence is also concatenated to introduce the information of position, and the output is a feature matrix, which is defined as $\mathbf{FP}_{\mathbf{inst}}$. It is altered into the matrix of association of $N \times N \times 1$ by MLP, which we call $\mathbf{W}_{\mathbf{pair}}$. Softmax is discarded, and sigmoid is utilized instead because softmax will cause individual results to stand out, which is not what we want. Finally, the repeated value $\mathbf{FR}_{\mathbf{inst}}$ is multiplied, summed, and enhanced by FFN so that each instance feature includes the feature information of the corresponding instance. The multi-layer stacking and the multi-head attention mechanism are also used to enhance the performance by refining features of each layer $\mathbf{F}^{\mathbf{i}}_{\mathbf{inst}}$.

The preceding can be described as the early fusion stage. In the later fusion stage, different human and object features are concatenated to form deeply integrated interaction pair features $\mathbf{Q}_{\mathbf{HOF}}$ for subsequent input as PID. With the combined effect of the early and later fusion, pairs of the interaction acquire features of associated instances as well as features with similar semantics. It will help the interaction recognition of the interaction pairs.

There are several differences between the HOF module and the attention module. Firstly, we use different linear layers to learn various features. For the modified transformer encoder layer in UPT (Zhang, Campbell et al., 2022), it has no obvious $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ distinction, which leads to learning the attention between $\mathbf{FR}_{\mathbf{inst}}$ and $\mathbf{E}_{\mathbf{spat}}$ using only one layer of linear and learning the mapping of features, affecting the final result. Secondly, we do not multiply $\mathbf{FR}_{\mathbf{inst}}$ with $\mathbf{E}_{\mathbf{spat}}$; the idea of UPT is to assign positional encoding $\mathbf{E}_{\mathbf{spat}}$ to each feature to break the symmetry. Instead, we use positional encoding $\mathbf{E}_{\mathbf{spat}}$ as a query to compute attention by pairwise-level attention. Finally, instead of using softmax as the normalization method, we use sigmoid as the normalization method because softmax causes individual attention results to stand out, and we want to accept more information. We also found that using multiple attentions to take the mean can improve the experimental effect. Considering all the above differences, our attentional graph results are also different from UPT (in Section 4.5).

$$\mathbf{W}_{\mathbf{pair}} = \text{Sigmoid}\left(\text{Cat}\left(\mathbf{FP}_{\mathbf{inst}}, \mathbf{E}_{\mathbf{spat}}\right)\right) \tag{3}$$

$$\mathbf{F}^{\mathbf{i}}_{\mathbf{inst}} = \text{FFN}\left(\text{Sum}\left(\mathbf{W}_{\mathbf{pair}} \times \mathbf{FR}_{\mathbf{inst}}\right) + \mathbf{F}^{\mathbf{i-1}}_{\mathbf{inst}}\right) \tag{4}$$

$$\mathbf{Q}_{\mathbf{HOF}} = \text{Cat}\left(\mathbf{F}_{\mathbf{inst}}{}^{i}[\mathbf{X}! = \mathbf{Y}\ \textit{and}\ \mathbf{X} = \mathbf{Human}]\right) \tag{5}$$

### 3.5. Training and inference

We assume that the presence of a human and an object are mutually independent. $P(HO) = P(H) \times P(O)$, $P(AHO) = P(A|HO) \times P(HO)$, where $HO$ indicates that both a human and an object exist in the image, and $AHO$ indicates that a human and an object have a specific interaction action. The score function of two-stage detectors can be defined as $\hat{\mathbf{S}}_{\mathbf{HOI}} = \hat{\mathbf{S}}_{\mathbf{H}} \times \hat{\mathbf{S}}_{\mathbf{O}} \times \hat{\mathbf{S}}_{\mathbf{A}}$, where $\hat{S}_H$ represents the confidence of human, $\hat{S}_O$ represents the confidence of object, and $\hat{S}_A$ represents the

confidence of interaction action. We optimize the $\hat{\mathbf{S}}_{\mathbf{HOI}}$ with focal loss, refers to Eq. (6), where $\alpha$ and $\gamma$ are hyper-parameters from focal loss, $N_{pos}$ and $N_{neg}$ represent the number of positive samples and negative samples respectively, $\mathbf{Y_i}$ indicates if this is the label of this action.

$$\mathbf{L} = \frac{1}{N_{pos}}\Big(\sum_{\mathbf{Y_i}=1}^{N_{pos}} -\alpha \times (1-\hat{\mathbf{S}}_{\mathbf{HOI}})^{\gamma}\log(\hat{\mathbf{S}}_{\mathbf{HOI}}) + \sum_{\mathbf{y_i}=0}^{N_{neg}} -(1-\alpha)\times \hat{\mathbf{S}}_{\mathbf{HOI}}^{\gamma}\log(1-\hat{\mathbf{S}}_{\mathbf{HOI}})\Big) \tag{6}$$

The inference follows UPT,

$$\hat{\mathbf{S}}_{\mathbf{HOI}} = \left(\hat{\mathbf{S}}_{\mathbf{H}}\right)^{\lambda} \times \left(\hat{\mathbf{S}}_{\mathbf{O}}\right)^{\lambda} \times \hat{\mathbf{S}}_{\mathbf{A}} \tag{7}$$

where the constant $\lambda = 2.8$ is used to suppress overconfident objects during inference.

## 4. Experiments

We first introduce the adopted datasets, metrics, and implementation details to help reproduce experimental results exactly. Next, we compare our GFIN method with previous state-of-the-art methods on V-COCO and HICO-DET datasets. Finally, through qualitative and quantitative analyses, our results provide experimental justification for development.

### 4.1. Datasets and metric

We adopt two HOI detection benchmarks to evaluate the proposed method: V-COCO (Gupta & Malik, 2015), and HICO-DET (Chao et al., 2018).

V-COCO is an evaluation dataset commonly used in HOI detection. It contains 10,346 images, including 16,199 people instances. Meanwhile, some of the images are unlabeled. Each annotated person has binary labels for 26 available actions, and three actions distinguish different roles, like obj or instr. Moreover, there are no roles for four actions. The presence of the role in images is not necessary for each action. We use the mean average precision (mAP) for the evaluation metrics. Scenario 1 is fit for missing roles due to the occlusion. An role prediction is correct if the action is accurate, the IoU of the person is $> 0.5$, and the corresponding role is empty. Scenario 2 fits the cases with roles outside the COCO categories (Lin et al., 2014). An role prediction is accurate if the action is correct and the IoU of the person is $> 0.5$.

HICO-DET is another evaluation dataset for HOI detection. It consists of a total of 37,633 training images and more than $150K$ annotated instances of human–object pairs, spanning the 600 HOI categories. It declares a true positive if the minimum of human overlap $IoU_h$ and object overlap $IoU_o$ exceeds 0.5 and considers two different evaluation settings. Known object setting evaluates the detection only containing the object category on images. Default settings evaluate the detection of images containing or not containing the object category.

### 4.2. Implementation details

The experiments are implemented on two versions of GFIN: GFIN-R50 with ResNet-50 (He, Zhang, Ren, & Sun, 2016) and GFIN-R101 with ResNet-101 as the backbone of DETR. To compare with the baseline method UPT, the parameters of DETR in the first stage are initialized as same as UPT but frozen.

For the interaction head, we filter out detections from the first stage with scores lower than 0.2 and sample each human and object for at least 4 and up to 15, prioritizing high-scoring ones. There are four layers for HOF to enhance and combine features and four encoder–decoder layers for GCE and PID with the vanilla transformer in the proposed GFIN. For the hidden dimension of HOF and GCE, $m = 256$ is used, and $m = 512$ is set for PIQ to fit the output of HOF and GCE.

In addition, for hyper-parameters used in the focal loss (Lin, Goyal, Girshick, He, & Dollár, 2017), $\alpha$ and $\gamma$ are set to 0.5 and 0.2, which is consistent with UPT. We optimize our networks with AdamW (Loshchilov & Hutter, 2017) with an initial learning rate of $2 \times 10^{-4}$. All models are trained for 20 epochs, and the learning rate is decreased by a factor of 10 at 10 epochs. The training is conducted on 4 GeForce GTX V100 devices, and the batch is set to 32, which averages a batch size of 8 per GPU.

### 4.3. Comparisons with state-of-the-art methods

Table 1 shows the comparison results of GFIN with other state-of-the-art methods on V-COCO and HICO-DET. For the HICO-DET dataset, GFIN outperforms state-of-the-art two-stage and one-stage methods. Specifically, GFIN-R101 yields a significant gain of 2.98 mAP (relatively 9.2%) compared with baseline UPT in default settings. Compared to one-stage detectors QPIC and CDN, commonly recognized as a comparison with transformer-based detectors, relative gains of 18.0% and 10.0% are obtained, respectively. The SOTA method GEN-VLKT (Liao et al., 2022) enhances interaction understanding by transferring knowledge from a visual-linguistic pre-trained model CLIP (Radford et al., 2021). The proposed method outperforms it without any extra features.

For the V-COCO dataset, refering to Table 2, UPT ignores the actions that roles do not present on the images, which causes low detection metrics. To solve this problem, we set the bounding boxes of such actions of the roles to the person itself and ignore the prediction of objects in the inference. The quantitative results demonstrate the effectiveness of our proposed strategy. GFIN with ResNet-101 outperforms UPT with a large margin of mAP as 9.4 (15.5%) in scenario 1. With the above strategy, UPT-R50 has 66.5 mAP in scenario 1, and 68.6 mAP in scenario 2. And UPT-R101 has about 68.5 mAP in scenario 1, and 70.5 mAP in scenario 2. In this case, GFIN gains 1.8 (2.7%) of mAP compared with UPT in scenario 1. The result also outperforms the SOTA method, DT (Zhou et al., 2022). Furthermore, in scenario 2, GFIN with ResNet-50 is slightly lower than STIP (Zhang, Pan et al., 2022). Considering the significant advantages we have in scenario 1, STIP's abundant definition of kinds of inter-interaction semantic dependencies help it capture features in the case of wrong or unseen roles.

### 4.4. Ablation study

To understand the critical ingredients of GFIN's superiority, we conduct a series of ablation experiments with GFIN-101 in HICO-DET datasets.

Table 3 shows the influence of the three aspects on the detection performance. Specifically, mAP is only 22.41, which is lower than the base model without GCE, PID, and HOF simultaneously. The UPT method is a two-stage encoder-encoder method. Differently, our proposed GFIN method is a two-stage encoder–decoder method. This difference leads us not to enhance the UPT method but to retain its one-stage skeleton and loss function, etc. Therefore, in the first line of our Table 3, we show the results of direct inference prediction of the features of the one-stage output. Therefore, relying only on the instance features of the first stage is not enough for HOI detection. Introducing HOF to our architecture promotes mAP to 28.65, with a gain of 6.24, which lies in both rare and non-rare classes. It confirms that the combination of features can also significantly improve the performance of interaction recognition. Experiments also show that adding GCE-PID improves mAP by around 11.33. We infer that a reasonable encoder–decoder architecture can regain semantic features from an image contributing to the interaction rather than lots of additional parameters and computations because the experimental results prove that hyper-parameter like the number of layers of GCE, HOF, and PID affect little. "1x, 2x" is the number of stacking layers in Table 3.

**Table 1**
Comparison results of GFIN with other state-of-the-art methods on HICO-DET. The best and the second best results are **highlighted** and underlined. The letters in the Extra column indicate the extra input features: **T** (Linguistic features of label semantic embeddings), **P** (Human pose features), **X** (No extra features).

| Method | Backbone | Epochs | Extra | HICO-DET | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Default | | | Known objects | | |
| | | | | Full | Rare | Non-Rare | Full | Rare | Non-Rare |
| One-stage detectors: | | | | | | | | | |
| PPDM (Liao et al., 2020) | HG104 | 120 | X | 21.94 | 13.97 | 24.32 | 24.81 | 17.09 | 27.12 |
| GGNet (Zhong et al., 2021) | HG104 | 120 | X | 23.47 | 16.48 | 25.60 | 27.36 | 20.23 | 29.48 |
| AS-Net (Chen et al., 2021) | R50 | 90 | X | 28.87 | 24.25 | 30.25 | 31.74 | 27.07 | 33.14 |
| QPIC (Tamura et al., 2021) | R101 | 150 | X | 29.90 | 23.92 | 31.69 | 32.38 | 26.06 | 34.27 |
| MSTR (Kim et al., 2022) | R50 | 50 | X | 31.17 | 25.31 | 32.92 | 34.02 | 28.83 | 35.57 |
| SSRT (Iftekhar et al., 2022) | R101 | 150 | T | 31.34 | 24.31 | 33.32 | – | – | – |
| DT (Zhou et al., 2022) | R50 | 80 | X | 31.75 | 27.45 | 33.03 | 34.50 | 30.13 | 35.81 |
| CDN-L (Zhang et al., 2021) | R101 | 90 | X | 32.07 | 27.19 | 33.53 | 34.79 | 29.48 | 36.38 |
| CDN-S+HQM (Zhong et al., 2022) | R50 | 80 | X | 32.47 | 28.15 | 33.76 | 35.17 | 30.73 | 36.50 |
| DOQ (Qu et al., 2022) | R50 | 80 | T | 33.28 | 29.19 | 34.5 | – | – | – |
| GEN-VLKT-L (Liao et al., 2022) | R101 | 90 | T | 34.95 | 31.18 | 36.08 | 38.22 | 34.36 | 39.37 |
| Two-stage detectors: | | | | | | | | | |
| InteractNet (Gkioxari et al., 2018) | R50 | – | X | 9.94 | 7.16 | 10.77 | – | – | – |
| iCAN (Gao et al., 2018) | R50 | – | X | 14.84 | 10.45 | 16.15 | 16.26 | 11.33 | 17.73 |
| TIN (Li et al., 2019) | R50 | 25 | P | 17.03 | 13.42 | 18.11 | 19.17 | 15.51 | 20.26 |
| DRG (Gao et al., 2020) | R50 | – | T | 24.53 | 19.47 | 26.04 | 27.98 | 23.11 | 29.43 |
| SCG (Zhang, Campbell et al., 2021) | R101 | 10 | X | 29.26 | 24.61 | 30.65 | 32.87 | 27.89 | 34.35 |
| UPT-R50(Zhang, Campbell et al., 2022) | R50 | 20 | X | 31.66 | 25.94 | 33.36 | 35.05 | 29.27 | 36.77 |
| STIP (Zhang, Pan et al., 2022) | R50 | 30 | T | 32.22 | 28.15 | 33.43 | 35.29 | 31.43 | 36.45 |
| UPT-R101(Zhang, Campbell et al., 2022) | R101 | 20 | X | 32.30 | 28.55 | 33.44 | 35.65 | 31.60 | 36.86 |
| IF (Liu et al., 2022) | R50 | 30 | X | 33.51 | 30.30 | 34.46 | 36.28 | 33.16 | 37.21 |
| GFIN-R50 | R50 | 20 | X | 34.03 | 31.44 | 34.80 | 37.70 | 34.53 | 38.65 |
| GFIN-R101 | R101 | 20 | X | **35.28** | **31.91** | **36.29** | **38.80** | **35.48** | **39.79** |

**Table 2**
Experimental results of V-COCO dataset. The best and the second best results are **highlighted** and underlined.

| Method | Backbone | Params(M) | Frames | V-COCO | |
|---|---|---|---|---|---|
| | | | | Scenario 1 | Scenario 2 |
| One stage methods | | | | | |
| UnionDet (Kim et al., 2020) | R50-FPN | – | 15.7 | 47.5 | – |
| HOI-Trans (Zou et al., 2021) | R50 | 60.4 | 14.3 | 52.9 | – |
| DIRV (Fang et al., 2021) | ED-d3 | – | – | 56.1 | – |
| QPIC-R101(Tamura et al., 2021) | R101 | 60.2 | 14.5 | 58.3 | 60.7 |
| GEN-VLKT-S (Liao et al., 2022) | R50 | 42 | – | 62.4 | 64.5 |
| GEN-VLKT-L (Liao et al., 2022) | R101 | 70.5 | – | 63.6 | 65.9 |
| CDN-S+HQM (Liao et al., 2022; Zhong et al., 2022) | R50 | – | – | 63.6 | – |
| CDN-L (Zhang et al., 2021) | R101 | 60.4 | 14.1 | 63.9 | 65.9 |
| OCN (Yuan, Wang, Ni, & Xu, 2022) | R101 | – | – | 65.3 | 67.1 |
| DT (Zhou et al., 2022) | R50 | – | – | 66.2 | 68.5 |
| Two stage methods: | | | | | |
| InteractNet (Gkioxari et al., 2018) | R50-FPN | 40.4 | – | 40.0 | 48.0 |
| GPNN (Qi, Wang, Jia, Shen, & Zhu, 2018) | R101 | – | – | 44.0 | – |
| iCAN (Gao et al., 2018) | R50 | 39.8 | 6.0 | 45.3 | 52.4 |
| TIN (Li et al., 2019) | R50 | – | – | 47.8 | – |
| DRG (Gao et al., 2020) | R50-FPN | 46.1 | 6.1 | 51.0 | – |
| SCG (Zhang, Campbell et al., 2021) | R101 | 53.9 | 5.5 | 54.2 | 60.9 |
| UPT-R50(Zhang, Campbell et al., 2022) | R50 | 50.0 | 14.4 | 59.0 | 64.5 |
| UPT-R101(Zhang, Campbell et al., 2022) | R101 | 68.9 | 9.6 | 60.7 | 66.2 |
| GFIN-R50(ours) | R50 | 55.2 | 11.7 | 68.3 | 70.5 |
| GFIN-R101(ours) | R101 | 74.1 | 9.0 | **70.1** | **72.3** |

HOF can stack multiple layers to boost the results, GCE and PID are the stacking of encoder and decoder, where the stacking of PID also increases the auxiliary loss per layer. We also measured the effect of increasing the number of parameters per layer, where GCE-PID adds 3.1M parameters per additional layer, while HOF adds 1.2M parameters per layer. It can be seen that, from efficiency considerations, more layers of HOF would be better for the results. Finally, combining these three parts increases mAP to 35.28.

We also study the impact of the sharing ResNet backbone, and experimental results show that sharing or not does not affect the performance too much but increases the training and inference time significantly without a shared backbone. In the SBR (GCE utilizes the shared backbone ResNet) case, $\mathbf{X_D}$ and $\mathbf{X_G}$ are the outputs of different convolution layers (Conv1, Conv2). And the non-SBR case, $\mathbf{X_D}$ and $\mathbf{X_G}$ are the outputs of different ResNet networks. As shown in Table 3, not using SBR has 42.4M more parameters, which is the ResNet101 parameter, while not using GCE-PID has 13.3M fewer parameters, which is the actual size of the 4-layers GCE-PID network. The reason for this is speculated that the CNN backbone serves as an initial feature extractor that has extracted enough features for downstream network learning. Also, Its stability helps the transformer structure to be learned.

As mentioned above, the GCE module combines the global context memory and the local instance memory with the crossover-residual mechanism to bring performance improvements. In this part, we mainly verify the necessity of the crossover-residual mechanism and the impact of global and local memory combinations. As Table 4 shows, without global context and local instance memory, mAP decreased by 2.04 and 0.75 separately, whereas MD is the memory provided by DETR. While

**Table 3**
Ablation studies of our proposed method and the effect of layer hyper-parameters on the HICO-Det. SBR represents that GCE utilizes the shared backbone ResNet. "1x, 2x" is the number of stacking layers.

| SBR | GCE-PID | HOF | Params(M) | Full | Rare | Non-Rare |
|---|---|---|---|---|---|---|
| – | – | – | 55.8 | 22.41 | 17.73 | 23.81 |
| ✓ | – | ✓ | 60.8 | 28.65 | 24.29 | 29.95 |
| ✓ | ✓ | – | 69.0 | 33.74 | 29.39 | 35.04 |
| – | ✓ | ✓ | 116.5 | 35.02 | **32.07** | 35.9 |
| ✓ | ✓ | ✓ | 74.1 | **35.28** | 31.91 | **36.29** |
| – | 1× | 4× | 64.6 | 34.77 | 31.11 | 35.86 |
| – | 2× | 4× | 67.8 | 35.14 | **32.28** | 35.96 |
| – | 3× | 4× | 70.9 | 34.69 | 31.50 | 35.64 |
| – | 4× | 2× | 71.7 | 34.82 | 30.82 | 36.01 |
| – | 4× | 3× | 72.9 | 35.08 | 31.75 | 36.07 |
| – | 4× | 4× | 74.1 | **35.28** | 31.91 | **36.29** |

**Table 4**
Evaluation of the influence of three modules with different settings. The CR represents the crossover-residual mechanism.

| Module | Strategy | Params(M) | Full | Rare | Non-Rare |
|---|---|---|---|---|---|
| GCE | w/o $\mathbf{M_G}$ | 69.4 | 33.24 | 30.65 | 34.02 |
| | w/o $\mathbf{M_D}$ | 74.0 | 34.53 | 31.19 | 35.52 |
| | $g(.)$ w/o CR | 74.0 | 34.6 | 31.39 | 35.56 |
| | $g(.)$ via CR | 74.1 | **35.28** | **31.91** | **36.29** |
| PID | w/o aux loss | 74.1 | 32.74 | 28.5 | 34.01 |
| | w/ query embedding | 74.1 | 34.34 | 30.73 | 35.42 |
| | ours | 74.1 | **35.28** | **31.91** | **36.29** |
| HOF | w/o cascade layer | 74.1 | 34.24 | 30.16 | 35.46 |
| | w/o $\mathbf{E_{spat}}$ | 74.1 | 34.51 | 31.05 | 35.54 |
| | $g(.)$ via softmax | 74.1 | 34.93 | 31.25 | 36.03 |
| | ours(via sigmoid) | 74.1 | **35.28** | **31.91** | **36.29** |

**Table 5**
Selecting all layers for training and comparing the predict result of the number of stack layers in inference. The Layer Num. represents the number of layers in the stack from the last layer. The best and the second best results are **highlighted** and underlined.

| Layer Num. | Full mAP | Rare mAP | Non-Rare mAP | Full recall | Rare recall | Non-Rare recall |
|---|---|---|---|---|---|---|
| 1 | 34.77 | 31.17 | 35.84 | 74.85 | 73.42 | 75.28 |
| 2 | 34.99 | 31.20 | 36.12 | 75.11 | 73.73 | 75.52 |
| 3 | 35.02 | 31.18 | 36.16 | 75.64 | 74.34 | 76.03 |
| 4 | 35.21 | **31.94** | 36.18 | 75.66 | 74.03 | 76.15 |
| 5 | 35.27 | 31.75 | **36.32** | 75.95 | 74.75 | 76.30 |
| 6 | **35.28** | 31.91 | 36.29 | **76.05** | **74.86** | **76.40** |

**Table 6**
Selecting the last layer for training and comparing the predict result of the number of stack layers in inference. The Layer Num. represents the number of layers in the stack from the last layer. The best and the second best results are **highlighted** and underlined.

| Layer Num. | Full mAP | Rare mAP | Non-Rare mAP | Full recall | Rare recall | Non-Rare recall |
|---|---|---|---|---|---|---|
| 1 | 34.52 | 30.93 | 35.59 | 74.85 | 73.42 | 75.28 |
| 2 | **34.67** | 30.90 | 35.80 | 75.11 | 73.73 | 75.52 |
| 3 | 34.65 | 30.71 | **35.82** | 75.64 | 74.34 | 76.03 |
| 4 | 34.49 | 30.60 | 35.65 | 75.66 | 74.03 | 76.15 |
| 5 | 34.53 | 30.77 | 35.66 | 75.95 | 74.75 | 76.30 |
| 6 | 34.54 | **31.13** | 35.56 | **76.05** | **74.86** | **76.40** |

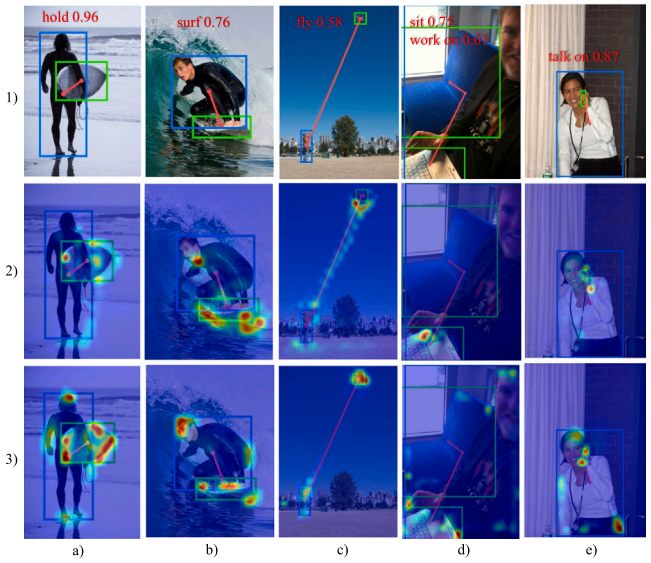


**Fig. 4.** Visualization of the GCE-PID module. (1) shows the image prediction results, (2) visualizes feature maps of GCE-PID, and (3) visualizes feature maps of the DETR encoder–decoder. (a) and (b) confirm that our method can extract features from the environmental information that contribute to interaction inference. (c) demonstrates that the features we infer are not limited to the interior of the object localization but the actual interaction region. (d) and (e) illustrate that the method can mine the interaction region features and discard the useless features.

$\mathbf{M_G}$, according to Table 3, shows that the number of its layers does not significantly affect the experimental results, so it can be confirmed that the GCE module is essential. In addition, global context memory plays the leading role. Moreover, the crossover-residual mechanism promotes mAP with a gain of 0.68 versus the concatenate operation.

For the PID module, we studied whether it is necessary to utilize the multi-layer auxiliary loss and query embedding. Adding auxiliary loss boosts the performance (11.54 mAP) because ignoring it will lose the advantage of refining the middle layer. Meanwhile, training with the learnable query embedding will reduce 0.94 of mAP, which illustrates that the extra query embedding is unnecessary for GFIN.

The multi-layer feature extraction of the first stage for increasing proposals is also a significant innovation of the HOF module. The mAP decreased by 1.04 in the case of using single-layer feature extraction. Also, we explore the necessity of positional encoding as a query sequence and whether it can be normalized using softmax. To improve the variability, we introduce positional encoding for our interaction pairs. The experimental results prove that positional encoding yields a gain of 0.77 mAP. The result of softmax is slightly worse than sigmoid.

We evaluate the effectiveness of concatenating the predicted results through multiple layers of DETR. Tables 5 and 6 show the comparison of the predicted result of stacking different number of layers in inference when selecting all the layers and only the last layer for training. Table 5 shows that increasing the number of layers improves the recall and results in the case of multi-layer evaluation. The results are better even with the last layer of inference than with single-layer training, so we use this approach to improve the experimental results without reducing the efficiency of inference. In contrast, as shown in Table 6, the single-layer training multi-layer inference does not improve the results, so this method is the training method that does not work directly on the inference.

### 4.5. Qualitative results

This section reveals the characteristics of GFIN qualitatively and the outstanding advantages it outperforms existing methods. We justify the model inference by visualizing attention feature maps of each part.

Fig. 4 shows the cross-attention maps of GCE-PID and the decoder of the first stage. It shows that the proposal GCE-PID can determine whether a person is surfing or holding a skateboard based on information about scenes like the sea or the beach that may not be included in the traditional interaction region. Also, the GCE-PID can determine the interaction between a person and a kite based on the kite string, which is outside the bounding box. Therefore, we conclude that the GCE-PID module can absorb the global contextual information for inference. When an image is working with computers, the GCE-PID pays more attention to the hand area, which is not what the object

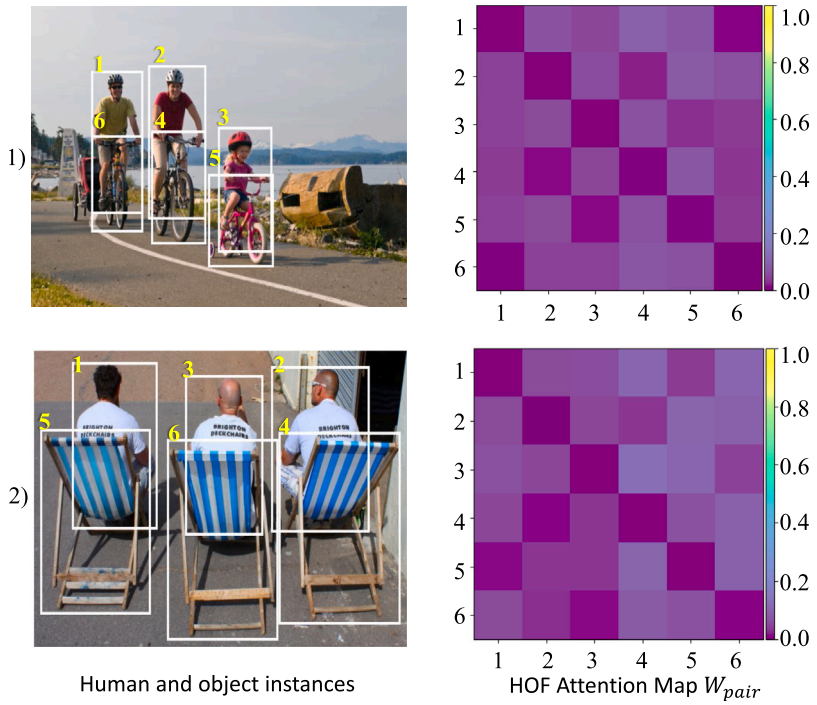


**Fig. 5.** Pairwise attention maps in the HOF module. (1) and (2) can be seen that the HOF module focuses more on pairs of the human–object with similar semantics, although without interactions. In the early fusion, our method prioritizes selecting to fuse instance features with similar interactions to assist in inference because premature fusion can lead to less feature variability between interaction pairs in the later fusion.

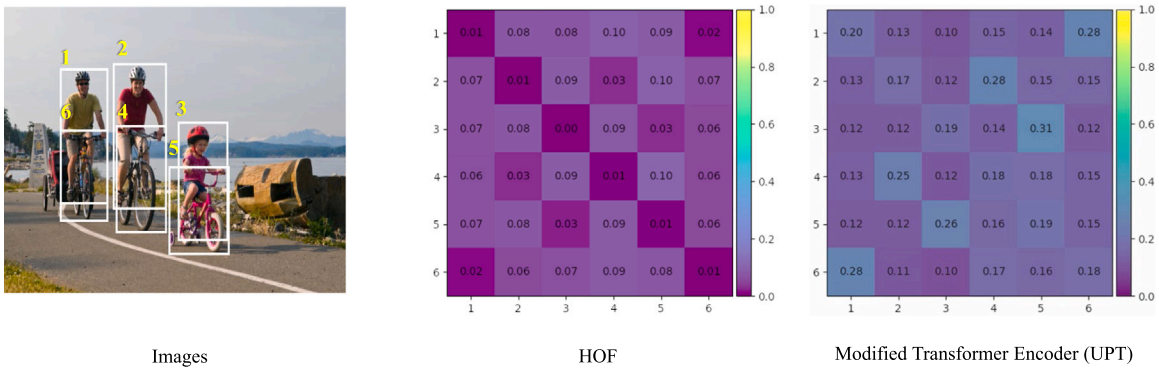


**Fig. 6.** The attention graph visualization of our HOF, and the modified transformer encoder layer in UPT.

detector is concerned about. When an image is talking on the phone, the GCE-PID cuts the area of attention to the area around the phone. So, it can selectively discard a large number of features that help in object localization but do not facilitate interaction, which meets the theoretical expectation.

Fig. 5 shows pairwise-level attention maps of HOF. The difficulty lies in accurately matching the relationships between humans and objects. We can see that the proposed pairwise horizontal attention mechanism pays attention to the surrounding related objects, enabling more valuable features to assist in human–object matching and interaction recognition. For example, although the person in the first bounding box and the bicycle in the fifth bounding box do not interact, we can assist in reasoning about the interaction object and the interaction action based on the interaction between the person in the third bounding box and the bicycle in the fifth bounding box because of the similar semantics. In summary, each module of GFIN can capture rich interaction semantic information from the contextual features for inference.

Fig. 6 shows the attention graph visualization of our HOF and the modified transformer encoder layer in UPT. A clear difference is that UPT focuses more on matching pairs with interactions, while HOF focuses on pairs without interactions. We have also investigated this phenomenon through experiments, and the main reason is the modification of softmax to the sigmoid. In the early fusion stage, we employ the concatenate operation to aggregate features in pairs. The HOF and the modified transformer encoder layer in UPT operations are capable of enhancing the features of individual instances. Only relying on the early feature fusion to extract the instance-level features, these interacting and non-interacting instances are connected, introducing unnecessary interaction information and leading to error detection. The early fusion stage should enhance the variability of instance features and obtain contextual information from semantic-like features to improve the results. Therefore the feature map of HOF is more reasonable.

Moreover, excluding the two-stage transformer model already compared in the paper and comparing it to the iCAN (Gao et al., 2018), TIN (Li et al., 2019) model, we have found that TIN operates with multi-stream instance-level attention similar to iCAN. We reproduced the iCAN model to verify the significant differences between the proposed methods. We similarly explore the differences between instance-level

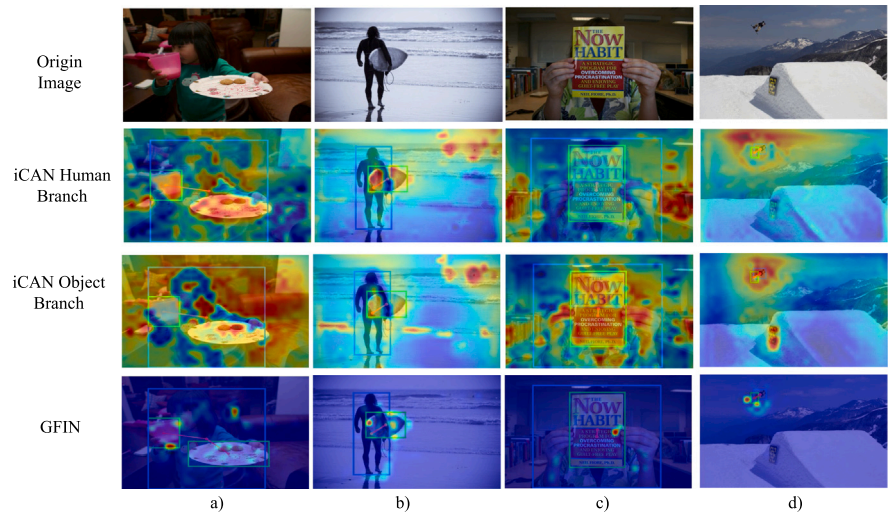


**Fig. 7.** The heat map of global context feature extraction for iCAN and GFIN. It shows that the heat map of GFIN is more rational and concentrated in the interaction region, while the heat map of iCAN is distributed in all areas of the image.


**Fig. 8.** Some failure examples. (a) Although the interaction between the person and object is correctly identified, the wrong action is caused by the fact that the child is less likely to hold a knife and the blowing of a candle is more common. (b) The overlap between the person and the object causes that it is difficult to distinguish precisely which person is interacting with the object. (c) The interaction in the mirror is not identified.

**Table 7**
The difference of iCAN and GFIN. The iCAN is our replication of the DETR-based.

| Method | Full mAP | Rare mAP | Non-Rare mAP |
|---|---|---|---|
| iCAN | 21.07 | 15.57 | 22.71 |
| GFIN | **35.28** | **31.91** | **36.29** |

and pairwise-level attention mechanisms. It can be seen that iCAN is a two-stage approach based on the Faster RCNN object detector, which has a performance gap with existing DETR object detectors. Our reproduced method uses the iCAN instance-level attention mechanism with DETR, and the experimental results are shown in Table 7. Our GFIN has a significant improvement over the iCAN. Fig. 7 shows the visualization of our heat map. It can be seen that the heat map of our approach is more rational and concentrated in the interaction region, while the heat map of iCAN is distributed in all areas of the image.

Fig. 8 shows some failure examples. (a) Although the interaction between the person and object is correctly identified, the wrong action is caused by the fact that the child is less likely to hold a knife and the blowing of a candle is more common. (b) The overlap between the person and the object causes that it is difficult to distinguish precisely which person is interacting with the object. (c) The interaction in the mirror is not identified.

## 5. Conclusion and limitation

In this paper, we present a two-stage Global Context and Pairwise-level Fusion Features Integration Network (GFIN) for HOI detection. GCE-PID achieves an efficient fusion of global context and local instance memory. Moreover, the pairwise-level attention mechanism enhances multi-layer instance feature fusion and adaptively learns relationships between humans and objects. Extensive experiments and analysis show that our proposed GFIN outperforms the current state-of-the-art HOI detectors on both benchmarks. Our method fills the gap of the two-stage HOI detector utilizing global contextual information and catalyzes development. In the future, we will consider incorporating multi-scale or linguistic features.

Our method could be improved. In the future, there are two points that could be further investigated. Firstly, as a two-stage method also

uses post-processing such as NMS, but the inference speed may be low, which can be partially alleviated by fewer stacking layers or by using a more lightweight network. Secondly, our method does not introduce multi-modal methods to enrich the input information. The failure examples tell us that the long-tail problem is a major problem of the present approach and how to solve the long-tail problem by introducing multi-modal method is an optional solution.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (No. 2021ZD0112400), the NSFC - Liaoning Province United Foundation under Grant U1908214, National Natural Science Foundation of China under Grant 61906032, the Fundamental Research Funds for the Central Universities, China under grant DUT21TD107, and the LiaoNing Revitalization Talents Program, No. XLYC2008017.

## References

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

Chao, Y.-W., Liu, Y., Liu, X., Zeng, H., & Deng, J. (2018). Learning to detect human-object interactions. In *2018 Ieee winter conference on applications of computer vision (Wacv)* (pp. 381–389). IEEE.

Chen, M., Liao, Y., Liu, S., Chen, Z., Wang, F., & Qian, C. (2021). Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9004–9013).

Dai, Z., Cai, B., Lin, Y., & Chen, J. (2021). Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1601–1610).

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

Fang, H.-S., Xie, Y., Shao, D., & Lu, C. (2021). Dirv: Dense interaction region voting for end-to-end human-object interaction detection. *Vol. 35*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 1291–1299).

Gao, C., Xu, J., Zou, Y., & Huang, J.-B. (2020). Drg: Dual relation graph for human-object interaction detection. In *European conference on computer vision* (pp. 696–712). Springer.

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

Gkioxari, G., Girshick, R., Dollár, P., & He, K. (2018). Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8359–8367).

Gupta, S., & Malik, J. (2015). Visual semantic role labeling. arXiv preprint arXiv:1505.04474.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hou, Z., Yu, B., Qiao, Y., Peng, X., & Tao, D. (2021). Detecting human-object interaction via fabricated compositional learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14646–14655).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Iftekhar, A., Chen, H., Kundu, K., Li, X., Tighe, J., & Modolo, D. (2022). What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5353–5363).

Kim, B., Choi, T., Kang, J., & Kim, H. J. (2020). Uniondet: Union-level detector towards real-time human-object interaction detection. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, Part XV 16* (pp. 498–514). Springer.

Kim, B., Lee, J., Kang, J., Kim, E.-S., & Kim, H. J. (2021). Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 74–83).

Kim, B., Mun, J., On, K.-W., Shin, M., Lee, J., & Kim, E.-S. (2022). MSTR: Multi-scale transformer for end-to-end human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 19578–19587).

Lee, D., Cheon, Y., & Han, W.-S. (2021). Regularizing attention networks for anomaly detection in visual question answering. *Vol. 35*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 1845–1853).

Li, Y.-L., Zhou, S., Huang, X., Xu, L., Ma, Z., Fang, H.-S., et al. (2019). Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3585–3594).

Liao, Y., Liu, S., Wang, F., Chen, Y., Qian, C., & Feng, J. (2020). Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 482–490).

Liao, Y., Zhang, A., Lu, M., Wang, Y., Li, X., & Liu, S. (2022). GEN-VLKT: Simplify association and enhance interaction understanding for HOI detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20123–20132).

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

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755). Springer.

Liu, X., Li, Y.-L., Wu, X., Tai, Y.-W., Lu, C., & Tang, C.-K. (2022). Interactiveness field in human-object interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20113–20122).

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., et al. (2021). Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3651–3660).

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

Qu, X., Ding, C., Li, X., Zhong, X., & Tao, D. (2022). Distillation using oracle queries for transformer-based human-object interaction detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19558–19567).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, *28*.

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., & Liu, J. (2022). Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tamura, M., Ohashi, H., & Yoshinaga, T. (2021). Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10410–10419).

Tan, M., Pang, R., & Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10781–10790).

Wang, T., Yang, T., Danelljan, M., Khan, F. S., Zhang, X., & Sun, J. (2020). Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4116–4125).

Wang, Y., Zhang, X., Yang, T., & Sun, J. (2022). Anchor detr: Query design for transformer-based detector. *Vol. 36*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2567–2575).

Xu, K., Ye, F., Zhong, Q., & Xie, D. (2022). Topology-aware convolutional neural network for efficient skeleton-based action recognition. *Vol. 36*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2866–2874).

Yuan, H., Wang, M., Ni, D., & Xu, L. (2022). Detecting human-object interactions with object-guided cross-modal calibrated semantics. arXiv preprint arXiv:2202.00259.

Zhang, F. Z., Campbell, D., & Gould, S. (2021). Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 13319–13327).

Zhang, F. Z., Campbell, D., & Gould, S. (2022). Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20104–20112).

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

Zhang, Y., Pan, Y., Yao, T., Huang, R., Mei, T., & Chen, C.-W. (2022). Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19548–19557).

Zheng, W., Yin, L., Chen, X., Ma, Z., Liu, S., & Yang, B. (2021). Knowledge base graph embedding module design for visual question answering model. *Pattern Recognition*, *120*, Article 108153.

Zhong, X., Ding, C., Li, Z., & Huang, S. (2022). Towards hard-positive query mining for DETR-based human-object interaction detection. In *Computer vision–ECCV 2022: 17th European conference, Tel Aviv, Israel, October 23–27, 2022, proceedings, Part XXVII* (pp. 444–460). Springer.

Zhong, X., Qu, X., Ding, C., & Tao, D. (2021). Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13234–13243).

Zhou, P., & Chi, M. (2019). Relation parsing neural network for human-object interaction detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 843–851).

Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., et al. (2021). Deepvit: Towards deeper vision transformer. arXiv preprint arXiv:2103.11886.

Zhou, D., Liu, Z., Wang, J., Wang, L., Hu, T., Ding, E., et al. (2022). Human-object interaction detection via disentangled transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19568–19577).

Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., et al. (2021). End-to-end human object interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11825–11834).