electronics

MDPI

Article

Interactivity Recognition Graph Neural Network (IR-GNN) Model for Improving Human–Object Interaction Detection

Jiali Zhang *, Zuriahati Mohd Yunos * and Habibollah Haron

Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia
* Correspondence: jiali-20@graduate.utm.my (J.Z.); zuriahati@utm.my (Z.M.Y.)

Abstract: Human–object interaction (HOI) detection is important for promoting the development of many fields such as human–computer interactions, service robotics, and video security surveillance. A high percentage of human–object pairs with invalid interactions are discovered in the object detection phase of conventional human–object interaction detection algorithms, resulting in inaccurate interaction detection. To recognize invalid human–object interaction pairs, this paper proposes a model structure, the interactivity recognition graph neural network (IR-GNN) model, which can directly infer the probability of human–object interactions from a graph model architecture. The model consists of three modules: The first one is the human posture feature module, which uses key points of the human body to construct relative spatial pose features and further facilitates the discrimination of human–object interactivity through human pose information. Second, a human–object interactivity graph module is proposed. The spatial relationship of human–object distance is used as the initialization weight of edges, and the graph is updated by combining the message passing of attention mechanism so that edges with interacting node pairs obtain higher weights. Thirdly, the classification module is proposed; by finally using a fully connected neural network, the interactivity of human–object pairs is binarily classified. These three modules work in collaboration to enable the effective inference of interactive possibilities. On the datasets HICO-DET and V-COCO, comparative and ablation experiments are carried out. It has been proved that our technology can improve the detection of human–object interactions.

Keywords: human–object interaction; interactivity recognition; graph neural network

Citation: Zhang, J.; Mohd Yunos, Z.; Haron, H. Interactivity Recognition Graph Neural Network (IR-GNN) Model for Improving Human–Object Interaction Detection. *Electronics* **2023**, *12*, 470. https://doi.org/10.3390/electronics12020470

Academic Editors: Namgi Kim and Hyunsoo Yoon

Received: 19 December 2022
Revised: 8 January 2023
Accepted: 10 January 2023
Published: 16 January 2023

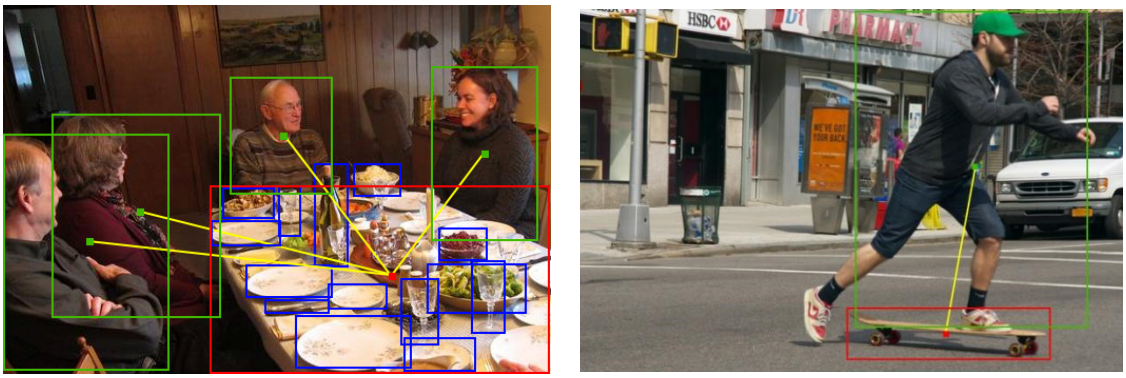
1. Introduction

A major research area in computer vision, human–object interaction detection is essential for robots to obtain a more holistic understanding of the physical environment [1]. Unlike object detection, human pose detection, and scene segmentation, which are vision tasks that only detect and segment objects in the scene independently, human–object interaction detection is performed to further infer the possible interaction between a person and an object in a scene, specifically to locate the person and object in the interaction relationship while inferring their interaction action category [2]. Human–computer interactions, service robots, and video security monitoring are just a few of the numerous areas where human–object interaction detection could have a significant impact on future research [3–5]. Improving the precision with which human–object interaction is detected is becoming an increasingly critical concern in these areas.

In recent years, this research has tended to introduce more and more features (e.g., visual appearance features, spatial features, human pose features, etc.) for the input of neural networks to facilitate the inference of human–object interactions, and accordingly, many neural network architectures have been explored to solve the HOI detection problem [6–26]. In general, to solve such problems, human–object interaction detection is transformed into an interaction classification problem. First, an object detector is trained

to obtain the localization of people and objects, and then the contextual information contained in the image is combined to classify the interaction between humans and objects, and the extraction of common contextual information in images is mainly divided into three kinds: the pose information of the human body, the detail information of human body parts, and the relative position information of humans and objects [6–10]. The most common method is based on the detection of a human and an object by an object detection network, followed by the fusion of one or more pieces of contextual information to assess the interaction between the human and the object. For example, Gkioxari et al. developed InteractNet [6], a human-centric model that extends the Faster R-CNN model with an extra branch to categorize actions at the target object position and action-specific probability density estimations to recognize human–object interactions; Gao et al. used an instance-centric attention module to extract contextual features that are complimentary to the appearance features of the local region (human/object frame) to improve HOI identification using an Instance-Centric Attention Network for Human–Object Interaction Detection (ICAN) [7]; Fang et al. presented a new paired body part attention model [8] to learn to attend to important parts and their connections for HOI detection; Li et al. created TIN [9], which uses interaction networks to gain general interaction knowledge from different HOI datasets and performs noninteraction suppression prior to HOI classification during inference; and Wan et al. introduced the Pose-aware Multilevel Feature Network (PMFNet) in light of the wide variations in human–object appearance and spatial arrangements, in addition to the subtle variances in similarity relations [10]. However, most methods based on human–object detection combine all detected humans and objects in the image sequentially when constructing a human–object pair, while in practice, a person generally interacts with only individual objects in the scene. This has certain pitfalls for most methods.

As seen in Figure 1a, the object detection network detects all persons and objects within the image, and then only a real interaction exists between the person and the table. However, since the existing methods combine humans with all objects, the combination yields <human, plate>, <human, cup>, <human, wine item>, etc. Nevertheless, these combinations are without any interaction information, and the number of these noninteractive combinations is more than the number of truly interactive <human, table> combinations. Due to the large number of noninteractive human–object pairs as negative samples do not provide useful learning information for training, the sample imbalance problem is often encountered in the training process, which makes adequate model training difficult; additionally, the network gradient decreases more slowly, and the optimization direction of the model is not as expected and may not be optimized the best, which eventually leads to inaccurate detection results and other problems.



(**a**) (**b**)

Figure 1. The figure on the left (**a**) shows existing methods to detect all persons and objects in an image, using yellow line connections with valid human–object interaction pairs. The figure on the right (**b**) shows the high probability of a 'ride' or 'jump' interaction indicated by the human bounding box above the skateboard, which will be described in Section 2.3.

In order to improve the accuracy of detecting interactions between persons and objects, we introduce a novel interactivity recognition graph neural network. The strategy described in this study is divided into three major parts, which are the human posture feature module, human–object pair graph module, and classification output module, in which the human body feature module uses a pretrained model to extract human key point information [10,22,27] and calculates relative spatial pose features, which are mapped to higher dimensional features using a connectivity layer. The person–object pair graph model employs a pretrained model to extract the coordinates and confidence of all persons and objects in an image [6,7,9,16,17,19]. Using the coordinates of persons and objects, instance features are extracted by ROI [9,10,12,20] cropping of persons and objects. Human and object instance features are used as nodes to construct a graph neural network (GNN) centered on humans [14,28,29]. The model uses the spatial relationship of distances between nodes as initialization weights for edges, guiding message passing to learn contextual information between related nodes. We consider that since the number of connections for invalid interactions is greater than the number of connections for valid interactions, an attention mechanism is used to reduce the impact of invalid interactions. Finally, the person–object pairs and their respective node features are output through a pairing operation. The advantages indicated in this study have been tested on HICO-Det [30] and V-COCO [31] through a number of experimental comparisons.

Overall, the contributions of this paper include the following:

(1) We present interactivity identification graph neural networks that identify valid interacting human–object pairings to increase the accuracy of human–object interaction detection. Before HOI model inference, removing invalid interactive human–object pairs and carrying only valid interactive human–object pairs for HOI inference helps to improve HOI inference performance.
(2) The advantage of our proposed model on the V-COCO and HICO-DET datasets is proved through experimental and comparative validation.

2. Related Works

2.1. Object Detection

One of the core goals of computer vision scene understanding is object detection, which tries to localize and identify the types of objects in the scene [32]. Numerous exceptional and mature object recognition algorithms, such as Faster R-CNN [33], SSD [34], YOLO [35], and Feature Pyramid Network [36] are now capable of detecting multiscale objects in images due to the rapid growth of computer technology and deep learning. The human–object interaction detection technique typically begins with an object detector to identify potential scene items and then proceeds to infer HOIs based on the resulting data. Based on previous work, we chose to use a pretrained object detector so that we could focus on the second half of the HOI inference network research design.

2.2. Human–Object Interaction Detection

Deep learning has greatly improved the performance of computer vision, and one can now extract features from large scale datasets rather than being limited to manually extracted features. Combined with the emergence of datasets dedicated to HOI detection, HOI detection tasks have entered a new phase of development. For example, Chao et al. proposed the Human–Object Region-based Convolutional Neural Network (HO-RCNN) [30], which is of great importance for the study of HOI detection. Researchers have attempted to model a structured output with attention mechanisms, and Gao et al. [7] proposed an instance-centric attention module to extract contextual features that are complementary to the appearance features of local regions (human/object frames) to improve HOI detection using HO-RCNN, their proposed Instance-Centric Attention Network for Human–Object Interaction Detection (ICAN). Graph models or graph convolution are being used by researchers to tackle the HOI detection challenge. Liang et al. built a Visual–Semantic Graph Attention Network (VS-GATS) [16], which is a dual graph attention network that aggregates

visual spatial and semantic information in parallel while providing robust disambiguation. However, the above two-stage method exhausts the combinations of people and objects in an image and then makes inference judgments pair by pair, which puts a huge burden on computational resources and has a high error rate. Our key idea is to identify human–object combinations that are valid interactions and remove the invalid interaction combinations as much as possible before performing HOI relationship identification.

2.3. Relative Posture Spatial Detection

Relative posture spatial features can provide important information when inferring interactive actions. For example, in Figure 1b, the human bounding box above the skateboard indicates a high probability of "ride" or "jump" interactions. Two methods are commonly used to encode spatial information between objects: One is by Chao et al. to acquire relative spatial information implicitly, build a "interactive pattern" (interactive template feature), and feed it into a convolutional neural network. This "interactive template feature" is a feature map where the pixels inside the object bounding box are 1 and the remainder are 0 [30]. The other is by Gupta et al. to use object bounding box coordinates explicitly to build relative or absolute posture spatial information [37]. This study also adopts the same relative spatial feature encoding as Gupta et al. [37]. Specifically, the relative distance feature from each joint point of the human body to the center of the object is measured so that a more detailed spatial feature can be constructed from the human body joint point coordinates.

2.4. Graph Neural Network

Graph neural networks (GNNs) have lately emerged as a scientific and technical hotspot in the field of computer vision, while convolutional neural networks (CNNs) [38] are among the most widely utilized neural networks in the area of computer vision. There has been some work integrating network structures with graphical models [38,39], and good results have been obtained in applications such as scene understanding [40,41], object detection and parsing [6,42], and Visual Question Answering (VQA) [43]. In the human–object interaction detection task, in order to address the HOI detection problem, it is crucial to take use of the notion of employing a graph model or graph convolution, as there is an unavoidable interaction between humans and objects, which builds a connected network. The objectives of the graph model in HOI detection are to represent people and things as nodes and interactions between people and items as edges, with the strength of the edges increasing as the relationship between persons and objects becomes more relevant. Qi et al. first integrated graph models and neural networks to implement HOI recognition, and they proposed a Graph Parsing Neural Network (GPNN) [13], which is a generalization of the Message Passing Neural Network (MPNN), which inherits the learning ability of neural networks and the representation ability of graphical models, but the representation of people and objects with the same type of nodes in a GPNN is not perfect. Therefore, Wang et al. proposed a contextually heterogeneous graph network [14] where persons and objects are represented by different nodes while the spatial relationship between persons and objects is the basic information for recognizing interactions, so it is encoded into the edges connecting the heterogeneous nodes. To further enhance the visual features, VSGNet [17] expands on the GPNN by using the spatial layout of human–object pairings and graph convolution branches. Our approach differs from them in that, first, human pose features are introduced as external complementary information for graph neural networks, and the updated graph nodes and edges are uniformly encoded with human posture features for the inference of interactivity. Second, human-centered graph neural networks are designed to learn about human–object interactions.

2.5. Transformer-Based HOI Methods

Transformers have made great breakthroughs in vision, and in recent years, Transformers have been widely used in HOI. Zou et al. [44] proposed the HOI Transformer,

which effectively inferred object–human associations from the global image context and directly predicted HOI instances. This method improves accuracy and has a low conceptual threshold. Kim et al. [45] presented HOTR, a Transformer encoder–decoder architecture that can immediately measure a collection of (person, object, interaction) triples and then effectively exploit the semantic connections in the pictures via ensemble prediction without requiring costly postprocessing. Zhang et al. [46] proposed Structure-aware Transformer over Interaction Proposals (STIP), designed to perform interaction proposal generation and structure-aware Transformers. HOI prediction is enhanced by encoding the overall semantic structure between interaction suggestions as well as the local spatial structure of people/objects in each interaction suggestion. Although Transformers enable end-to-end training to improve recognition speed, our approach is to determine the interactivity of human–object pairs, so the use of graph models can be sufficient.

3. Methodology

We refer to the proposed novel interactivity recognition graph neural network as IR-GNN, and the overall architecture diagram is shown in Figure 2, which contains three modules that work together to achieve accurate human–object pair interactivity detection. IR-GNN is mainly divided into the human posture feature module, human–object pair graph module, and classification output module.

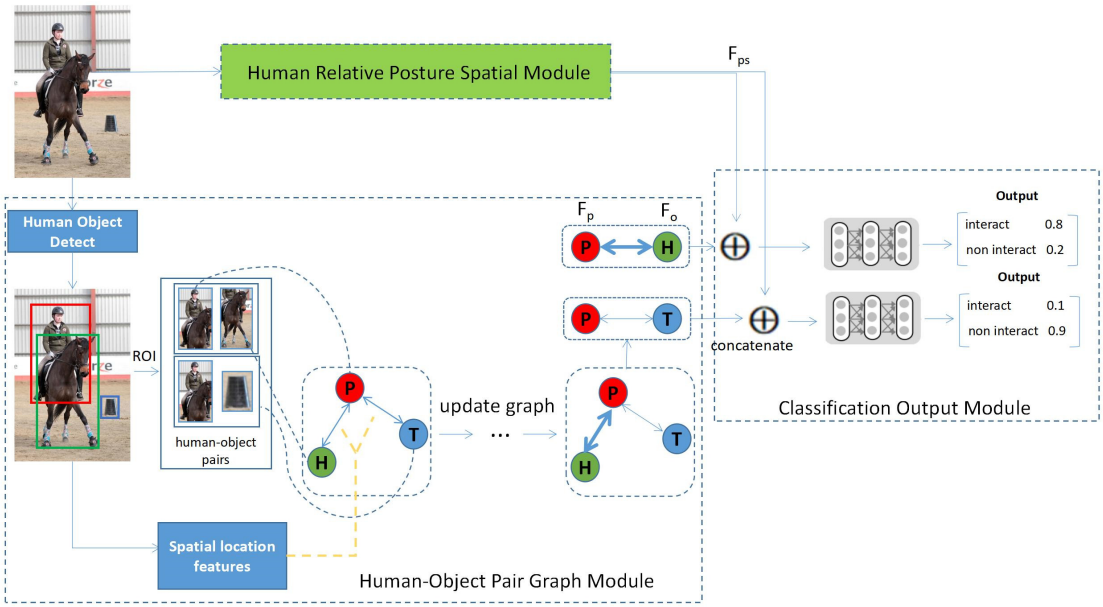


Figure 2. Architecture of the IR-GNN (interactivity recognition graph neural network), where P, H, and T represent the instance features of a person and two objects, respectively, and F_p and F_o represent the node information after graph update.

3.1. Human Relative Posture Spatial Module

In order to detect the human joint points (17 key points), we used the existing human key point pretraining detection model. In this study, the relative spatial posture feature was used, and the relative spatial posture feature is the relative distance feature to the center of the object, as shown in Figure 3. If the coordinates of the ith joint point of the human body are defined as (x_i, y_i), then its relative spatial feature f_{rp}^i is as follows:

$$f_{rp}^i := (x_i', y_i') = \left(\frac{x_i - x_c^o}{W}, \frac{y_i - y_c^o}{H}\right) \tag{1}$$

where (x_c^o, y_c^o) is the center coordinate of the object's bounding box and (W, H) is the image's size. All nodes of relative spatial posture attributes are specified as $f_{rp} \in \Re^{17\times 2}$.

Figure 3. Relative posture spatial features.

Human relative posture spatial module is shown in Figure 4. The f_{rp} is mapped to higher dimensional features through two fully connected layers, and then the obtained features are processed to batch normalization and dropout procedures, where the activation function is ReLU, the process of which is described by the following equation:

$$h_{pose} = \mathrm{Re}LU(\mathrm{Re}LU(f_{rp}W_0)W_1) \tag{2}$$

where $W_0 \in \Re^{2\times128}$ and $W_1 \in \Re^{128\times64}$ are trainable parameters.

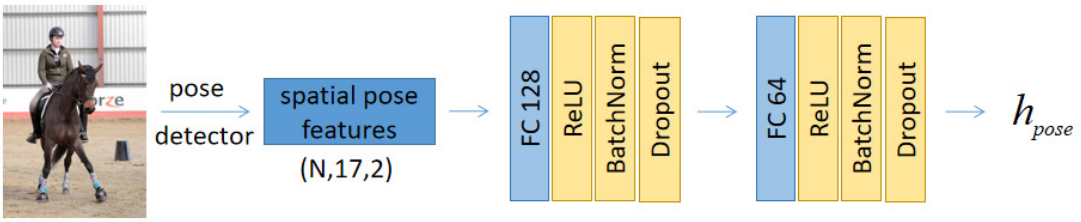


Figure 4. Human Relative Posture Spatial Module.

3.2. Human–Object Pair Graph Module

We suggest learning interaction knowledge in the connection graph [9], where nodes represent all people and objects, edges indicate the relationship between two nodes, and the spatial distance between persons and objects is constructed as the initial weight of the edges. Interactive inference is then performed under the supervision of the GNN. Thus, interactivity recognition may be viewed as a binary categorization of pairs of nodes. Higher values, on the other hand, suggest a stronger pair interaction and greater confidence in the edge weights between these two nodes.

3.2.1. Instance Feature Extraction of Persons and Objects

First, based on previous work, a pretrained model for object detection is taken to extract the original image features F_{org} from the image. Then, according to the bounding boxes $L_h = (x_h, y_h, w_h, h_h)$ and $L_o = (x_o, y_o, w_o, h_o)$ of the person and object obtained through the target detector detection and the confidence level S_h of the person and the confidence level S_o of the object, (x_h, y_h) and (x_o, y_o) are the center coordinates of the human body and the object bounding box, respectively; w_h and h_h are the width and height of the human body bounding box, respectively; and w_o and h_o are the width and height of the object bounding box, respectively. Next, the RoI pooling [3] operation is used to

obtain the instance features F^h_{inst} and F^o_{inst} of the person and object, respectively, as shown in Equations (3) and (4).

$$F^h_{inst} = RoiPooling(F^h_{org}, L_h) \tag{3}$$

$$F^o_{inst} = RoiPooling(F^o_{org}, L_o) \tag{4}$$

where the instance features F^h_{inst} and F^o_{inst} of persons and objects are denoted as v_i, and i denotes the entity feature of the ith person or object.

3.2.2. Spatial Location Features

Among the existing HOI relevant research work [7,11,13,16,17], spatial vision alone is not sufficient to determine the classes of interaction actions, but it has a strong relevance in the recognition of valid or invalid interactions. For example, a person may interact with a chair because they overlap spatially; if there is no person near the object, the probability of that object interacting with a person is low. Considering that the spatial relationship between individual nodes has an influence on whether they interact or not and is, therefore, centered on people [16], the distance space is modeled as the weight of the edges of the connectivity graph. The definition is as follows:

$$h_{f_{ij}} = \frac{F_{dist}(i,j)}{\sum_{j=1}^{M} F_{dist}(i,j)} \tag{5}$$

$$F_{dist}(i,j) = \frac{1}{D(b_i, b_j)} \tag{6}$$

where $F_{dist}(i,j)$ represents the spatial connection of two instances. $D(b_i, b_j)$ indicates the distance determined from the two instances' box coordinates.

3.2.3. Graph Model

In order to identify which objects a person is validly interacting with, a connectivity graph is constructed of people and objects, which can be represented by an adjacency matrix of physically stored node features: $X_v \in R^{n \times d}$. The edge feature adjacency matrix is represented by $X_e \in R^{m \times c}$, where n is the number of feature instances of persons and objects, m is the number of edges, d is the length of the node feature, and c is the length of the edge feature. f_i is the features of instances of persons and objects, and m is the number of relationships constituted by persons and objects. The point set V consists of person and object instances as nodes, and the edge set E consists of individual person and object instance relationships as edges. V and E consist of the graph $G = (V, E)$, where $v_i \in V$ is the ith node and $e_{ij} = (v_i, v_j) \in E$ is the directed edge from node v_i to node v_j, where the identity of node v_i is denoted $h_{v_i} \in R^d$ and the identity of edge e_{ij} is $h_{e_{ij}} \in R^c$. The graph-based approach focuses mainly on designing various subgraphs, in which the nodes of persons connect only the nodes of objects. In fact, in the real world, persons and objects are often interconnected, and there is no interactivity of objects that exist separately from persons; therefore, the graph model is constructed with persons in the center. All the nodes in our graph are considered unions based on the same code space, and the relationships among these nodes are also learned and contribute to the interactivity prediction.

To infer the graph model as a GNN (graph neural network), first use the edge function $f_{edge}(\cdot)$ to encode the features of the relationship between two connected nodes:

$$h_{e_{ij}} = f_{edge}([h_{v_i}, h_{fij}, h_{v_j}]) \tag{7}$$

where $[\cdot, \cdot]$ denotes the concatenation operation.

In the connection graph, which contains human–object connections with and without interactions, if Equation (7) is used directly when integrating the features of neighboring nodes, the number of connections with invalid interactions is greater than the number of

connections with valid interactions [47], i.e., the number of noises is too large; thus, this algorithm introduces an attention mechanism to reduce the interference of invalid interactions:

$$a_{ij} = softmax_j(f_{attn}(X_e^j)) = \frac{\exp(f_{attn}(h_{e_{ij}}))}{\sum_{v_o \in N_i} \exp(f_{attn}(h_{e_{io}}))} \tag{8}$$

where X_e^j denotes the feature matrix of all edges starting from node v_j, function $f_{attn}(\cdot)$ is used to map the internode relationship features (Equation (7)) to another implicit space, and the a_{ij} weight value indicates the importance or relevance of node v_j to node v_i.

Integrate the neighboring node features and the relationship features between the two nodes using the weight values weighted as described above, and update the features of each node using the update function $f_{update}(\cdot)$:

$$z_{v_i} = \sum_{v_o \in N_i} a_{io}(h_{v_o} + h_{e_{io}}) \tag{9}$$

$$\widetilde{h}_{vi} = f_{update}([h_{v_i}, z_{v_i}]) \tag{10}$$

After $f_{update}(\cdot)$, the updated visual graph of node features shown in Figure 2 is obtained, where the edges with different thicknesses indicate the magnitude of the edge weights. This algorithm implements the above attention function $f_{attn}(\cdot)$, edge function $f_{edge}(\cdot)$, and node update function $f_{update}(\cdot)$ in experiments using a single fully connected layer with neuron nodes of 1, 1024, and 1024, respectively.

3.3. Classification Output Module

With the graph neural network described above, each node feature already encodes rich information about the relationship of a scene (the relationship between a node and other nodes). In this study, we propose that all persons are paired with all objects one by one to form a specific "person-object pair" and then they are combined with features of the person's pose to infer whether there is interaction between them.

Following the pairing process, the paired person–object pair features are combined with the human posture features obtained from the human posture module to create the vector feature $F_{ij} = [h_{v_i}, h_{pose}, h_{v_j}]$ which is used to determine if there is an interaction between that human and object pair. Finally, the probability of the existence of interaction is calculated after the fully connected layer FC and the sigmoid activation function. The equation is as follows:

$$S^a = sigmoid(FC(F_{ij})) \tag{11}$$

In calculating the probability of the existence of human–object pair interactions (triples (humans, predicates, objects)), as with some existing algorithms [6,7,10,15–17], to take into consideration the confidence S_h of humans from the object detector output as well as the confidence S_o of the object, the formula is calculated as follows:

$$S = S_h \times S_o \times S^a \tag{12}$$

This formula determines whether a human–object pair interacts falls under the category of a binary classification issue; hence, in this technique, the interaction category of each human–object combination is determined using the binary cross-entropy loss function $BCE(\cdot)$, and the loss function is as follows:

$$Loss_1 = \frac{1}{N \times 2} \sum_{i=1}^{N} \sum_{j=1}^{2} BCE(S_{ij}, y_{ij}^{label}) \tag{13}$$

where N is the number of human–object pairs in the data, S_{ij} is the probability that a person–object pair interacts with each other, and y_{ij}^{label} is the corresponding true label.

Because the set of all possible human–object pairs is denoted as $\rho = \{P = (h, o) \in H \times O\}$, the detection results of the pretrained target detection network are H and O, representing the set of humans and the set of objects, respectively. ρ can be further divided into two subsets: $\rho = \widehat{\rho} \cup \breve{\rho}$, where $\widehat{\rho}$ and $\breve{\rho}$ denote the set of annotated and unannotated human–object pairs, respectively. Here, the ranking score of the set of annotated human–object pairs is higher than that of the unannotated human–object pairs, so $g(\widehat{\rho})$ is much larger than $g(\breve{\rho})$. Therefore, the loss function is designed as follows:

$$Loss_2 = \sum_{\rho \in \widehat{\rho}} \sum_{\rho \in \breve{\rho}} \max(0, g(\breve{\rho}) - g(\widehat{\rho}) + \alpha) \tag{14}$$

In summary, the loss function is integrated, and the formula is as follows:

$$Loss = Loss_1 + Loss_2 \tag{15}$$

4. Experiments

4.1. Experimental Configuration

4.1.1. Experimental Dataset

Using the 80 object categories in MS COCO (humans as a class among objects) and the 116 frequently used verbs, Chao et al. created an image classification dataset of human–common object interactions, HICO [17]. This dataset was based on the MS COCO [5] dataset, which is frequently used for target detection. Each object has a "no-interaction" action, with 600 human–object interactions, over 250,000 labeled individual human–object instances, and over 150,000 labeled instances of human interactions. The dataset contains more than 40,000 images, with at least six images per person–interaction category, and each category is represented by at least one image in the test set. HICO does not offer instance-level annotation for each person–interaction pair that appears in each image, and images with many people present are not fully labeled. Based on this, the authors of HICO created the HICO-Det dataset with detection [9], containing categories and bounding boxes for each human–object pairing and categories of human–object interactions.

A common dataset for person–interaction detection is V-COCO (Verbs in COCO) [8]. Such as HICO, object categories are obtained from the COCO dataset; however, unlike HICO, V-COCO uses images from COCO and existing person–object categories and bounding box labels to design and label 26 common interaction categories. The V-COCO dataset is divided into a training validation set (trainval set) and a test set (test set). The training validation set contains 5400 images and 8431 human instances, and the test set contains 4946 images and 7768 human instances. Each image has an average of 1.57 individuals, and each individual has an average of 2.87 behaviors.

4.1.2. Evaluation Metrics

For V-COCO, this paper uses the mean average precision of roles (mAP role) to assess the accuracy. For HICO-DET, this paper uses a generic evaluation setting with three categories: full (full, 600 HOIs), rare (rare, 138 HOIs), and nonrare (nonrare, 462 HOIs), where the rare category is defined as the number of instances of a certain type of human–object interaction in the training set that is less than 10. It is also divided into unknown (default) and known object (known default) classes, where the unknown class means that for each interaction class, the entire test set is detected and evaluated, and the detection is performed uniformly regardless of whether there are object instances of the corresponding class in the image, and the known object class means that only the images with object instances of the corresponding class are detected. The proposed method in this paper will be experimentally validated on the HICO-DET dataset and the V-COCO dataset.

4.1.3. Experimental Setting Parameters

This experiment was built based on the Pytorch deep learning library as well as the DGL library [48], where all network layers were built using fully connected layers. For the

object detector, this algorithm directly selected the Faster R-CNN model with ResNet-50-FPN as the backbone network that was trained on the COCO dataset in the Pytorch library, referring to previous work [15], and the detection results were composed of a set of human instances for targets identified as "human" with a confidence level greater than 0.8, and a set of object instances for targets not identified as "human" with a confidence level greater than 0.4. For the human keypoint detector, the torchvision library of Pytorch was used with the model keypointrcnn_resnet50_fpn for keypoint detection, which can detect 17 human keypoints. During training, the object detector and the human keypoint detector were frozen (no parameter update). We used a minibatch that was size 32, and the dropout rate was set to 0.3. For the activation function, the ReLU activation function was used except for the LeakyReLU activation function (parameter set to 0.1) in the attention network layer, and the optimizer was chosen from Adam [49], with the initial learning rate set to 1×10^{-5} and other parameters defaulted. The activation function after each GNN layer in the graph model was ReLU, and the dropout rate was 0.5. The epoch of training was set to 200.

In this experiment, the type of graphics card used in training and testing was NVIDIA RTX 3060, programmed in Python 3.6 under the Ubuntu 16.04 operating system platform, and the deep learning framework used was Pytorch 1.1.0.

5. Result and Discussion

5.1. Comparison with Other Methods

In this experiment, three representative models were selected, all of which shared the common inference feature of exhaustively enumerating all persons and object combinations in the image before making inference judgments pair by pair, which affected the final model performance due to the extremely large negative samples. Changing the original model in the object detection stage to the method IR-GNN in this paper made the model reason from the valid human–object pair information only, and we compared the new model with the original model. The selected representative models are presented as follows:

1. ICAN [7]. Using an instance-centered attention module, in order to extract contextual features that are complementary to the appearance features of local regions (person and object frames) to improve HOI detection.
2. DRG [19]. The contextual information of the aggregated scenes, one of which is human centric and one object centric, was used to refine the prediction by exploiting the relationship between different HOIs. The model effectively captured distinguishing cues from the scenes to resolve ambiguities in local prediction.
3. RPNN [20]. Detailed body part features were introduced, and the model incorporated a graph structure for feature refinement, and then the learnable graph model was extended from human and object appearance features to obtain a robust representation.

To verify the validity, we compared the model with three representative models on two datasets and three new models generated after adding our method of IR-GNN. The HOI detection results were evaluated according to the evaluation metrics of V-COCO and HICODET. For a fair comparison, object detectors for all methods were pretrained on the COCO dataset only.

The results obtained by applying our method to an existing model were more competitive than the original model. As can be seen in Figure 5, ICAN + IR-GNN achieved 17.32%, 13.18%, and 20.59% on the HICO-DET dataset for mAP full, rare, and nonrare categories, which was a 2.48%, 2.73%, and 4.44% improvement over the original ICAN, respectively. In Figure 6, it can be seen that the application of our method on ICAN resulted in the highest performance improvement because although ICAN uses a multistream structure and adds attention mechanisms to visual and spatial information, it is not designed to specifically deal with human–object noninteraction judgments; in this case, ICAN combined with our method made up for the deficiencies of ICAN and improved the inference performance of ICAN. Similarly, DRG and RPNN improved by 0.91%, 1.21%, and 0.92% and 2.16%, 2.04%, and 1.08% for the mAP full, rare, and nonrare categories, respectively, on the HICO-DET dataset after applying the methods in this paper. As can be seen in Figure 6, ICAN had the

highest improvement on the V-COCO dataset with 3.61%, followed by RPNN and DRG with 1.59% and 2.64% increases, respectively. RPNN focused on modeling the relationship between object–body part pairs and human–body part pairs. Although it helps in the determination of interactions, the features of human–object pairs were not comprehensive enough, resulting in the inability to better model the subtle interactions between body parts and objects; therefore, the inclusion of our method IR-GNN helped to improve the inference performance. DRG constructed two different graph models centered on the person and object to analyze this problem together and it had a good performance on the interactivity judgment, so DRG also had a performance improvement, but the improvement was not large. We found that the mAPs of rare categories were improved as seen from the data results, proving that our method is helpful in dealing with the long tail of HOI and providing clues for future research.

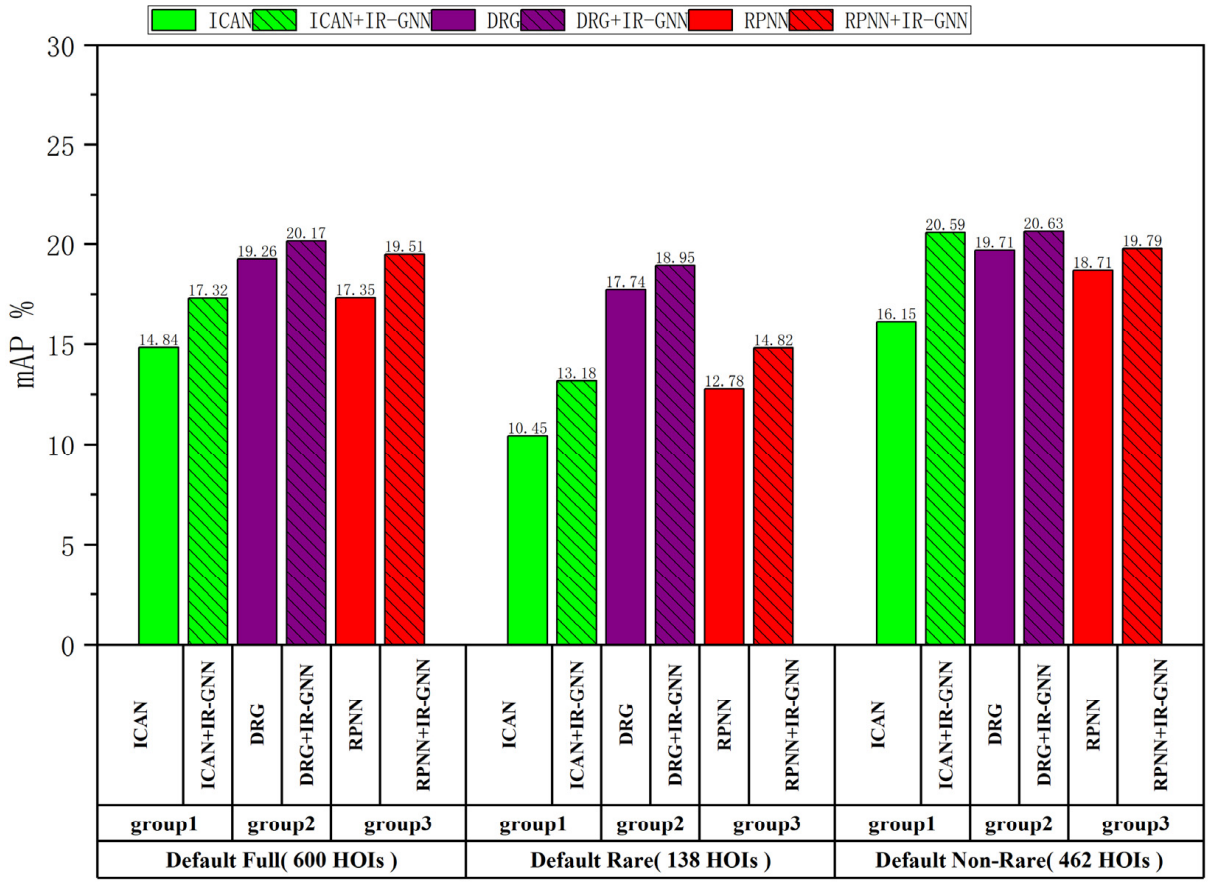


Figure 5. Comparison of mAP with other state-of-the-art methods on the HICO-DET dataset.

Combining the above comparative analysis, removing the invalid interacting human–object pairs before HOI model inference, and inferring only the valid interacting human–object pairs for HOI helped to improve the HOI inference performance.

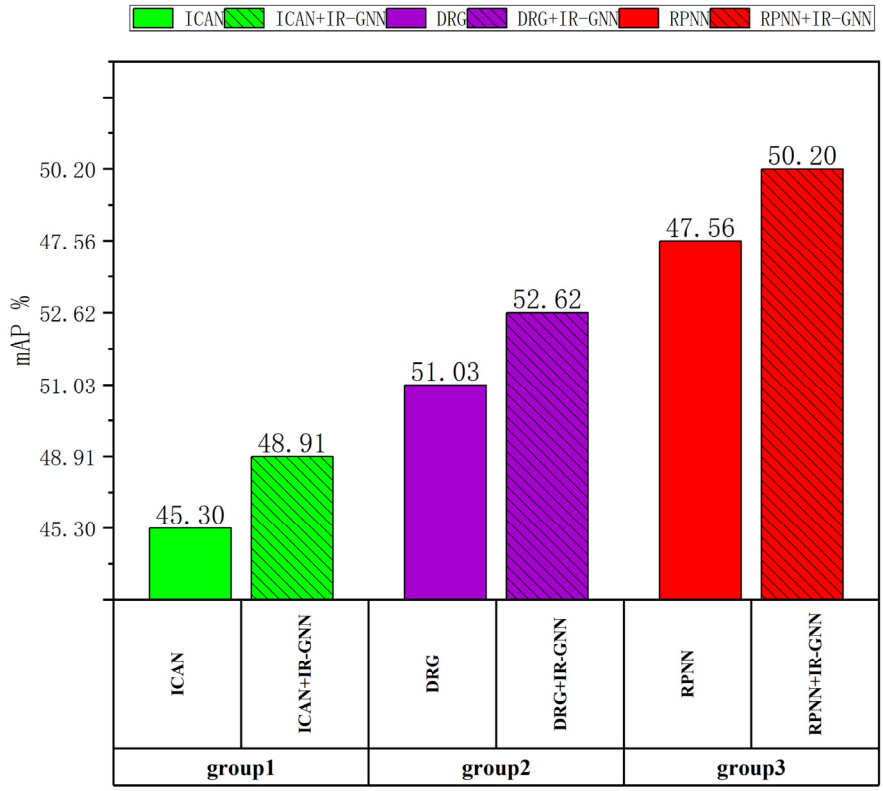


Figure 6. Comparison of mAP with other state-of-the-art methods on the V-COCO datasets.

5.2. Ablation Experiment

We took the ICAN+IR-GNN with the highest improvement in each performance on the two datasets among the previously applied models and used it as the benchmark model, which is noted as Experiment 0. To examine the impact played by each component of our approach, in Table 1 we evaluate their performance on the V-COCO and HICO-DET test sets where, in the HICO-DET dataset, we selected the full data. In Experiment 1, we removed the human posture information and the performance decreased by about 1.08% and 0.78% on the two datasets, respectively; therefore, the human posture information facilitated the inference of the whole model. In Experiment 2, after removing the attention mechanism, the model achieved 47.06% and 15.67% mAP detection results, which showed that the application of the attention mechanism can improve the detection accuracy of the model to a certain extent. In Experiment 3, we sought to determine if using the human-centered person-to-object distance as the initial edge weight contributes to the effectiveness of messaging between nodes, and to verify this, we set the initial weight of all edges to one. In this case, the model achieved 46.19% and 15.12% mAP and a decrease of 2.72% and 2.20%, which were significant decreases; additionally, the removal of the initial edge weight value caused the node relationship to lack the initiation direction, and this experiment again validated the conclusions of [16]. In Experiment 4, we only retained the single-layer graph convolution and removed the nodal feature update operation, and the experiments only achieved 45.89% and 15.08% of the detection results. Lacking the nodal update, the relationship between the nodal features could not be effectively transmitted; therefore, the nodal update operation played an important role in the sparsity of graph relationships, which had a greater impact on the subsequent interaction inference.

Table 1. Comparison of mAP for ablation experiments on V-COCO and HICO-DET (full) test sets.

Experiment No.	Methods	V-COCO	HICO-DET
0	ICAN + IR-GNN	48.91%	17.32%
1	w/o human pose stream	47.83%	16.54%
2	w/o attention	47.06%	15.67%
3	w/o distance space	46.19%	15.12%
4	w/o f_{update} in graph	45.89%	15.08%

5.3. Comparison with TIN

Li et al. [9] proposed an interaction recognition method TIN (Transferable Interactiveness Knowledge Network) to explicitly distinguish noninteractive pairs and suppress them before HOI classification, thus reducing the interference caused by too many noninteractive candidate pairs. In both datasets, our method and TIN were validated against each other in two ways, the recall of false positive samples (invalid human–object interaction pairs) and the mAP of HOI classification. For a fair comparison with TIN, ResNet-50 was uniformly selected as the backbone network of the target detector, with the aim that the object and combined human–object pairs detected by the two models were the same in the object detection phase. During the interaction detection experiments(as shown in Table 2), with the V-COCO and HICO-DET datasets, our method IR-GNN achieved a recall of 70.42% and 69.83%, respectively, for false positive samples, while the TIN method reached only 65.98% and 64.76%, respectively. Therefore, our method performed better in removing the false interaction pairs. For the human–object interaction test, we modified TIN to replace the Noninteraction Suppression (NIS) part with our method IR-GNN and left the rest of the structure unchanged. From Table 3, it can be seen that TIN's accuracy improved by 1.15% and 2.29% on both datasets after adapting our method. It was demonstrated that the higher the rate of removing false positive samples, the higher the accuracy of human–object interaction detection.

Table 2. Comparison of recall with TIN for false positive samples on both datasets.

	V-COCO (Recall)	HICO-DET (Recall)
TIN [9]	65.98%	64.76%
IR-GNN	70.42%	69.83%

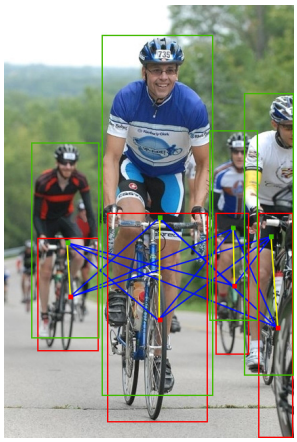
Table 3. Comparison of mAP with TIN for human–object interaction category identification on two datasets.

	V-COCO (mAP)	HICO-DET (mAP)
TIN [9]	48.70%	17.22%
TIN + IR-GNN	49.85%	19.51%

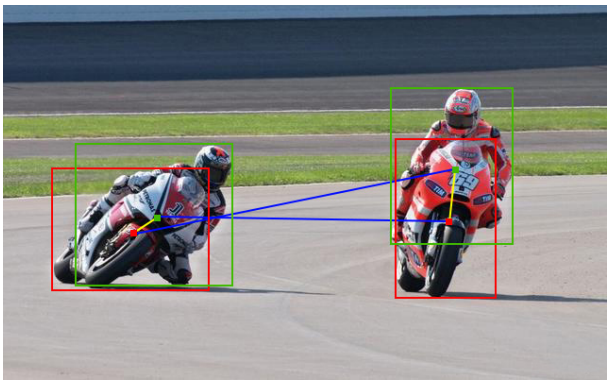
5.4. Qualitative Results

In Figure 7, we show the visualization results of the interactivity recognition. To test the results of our method for detecting human–object interactivity in images, we deliberately chose humans and objects in complex scenes, which in general may have had noninteractive pairs. In the image in Figure 7a, despite the same pose of the person and the same features of the bicycle, the spatial relationship between the instances as the edge relationship of the graph model made no interactive pairing of the person with other bicycles, indicating the great stability of our method in highly correlated scenarios. The same principle was applied to the image in Figure 7b. When analyzing the Figure 7e image, due to the interaction between the spatial relationship between the human pose features and the instances as the edge relationship of the graph model, the model could correctly determine the existence of an interaction between two persons and the corresponding

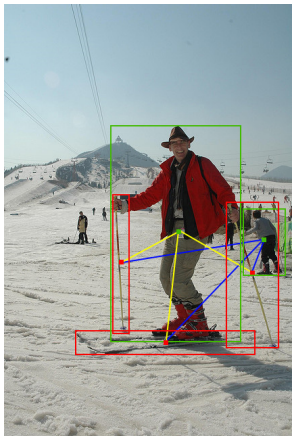
bicycle and exclude the third person. In the Figure 7d image, a person is holding a cell phone and is talking on the phone; however, our method failed to recognize the interactivity of the person and the phone, and this deficiency area can be solved from the aspect of object detection.



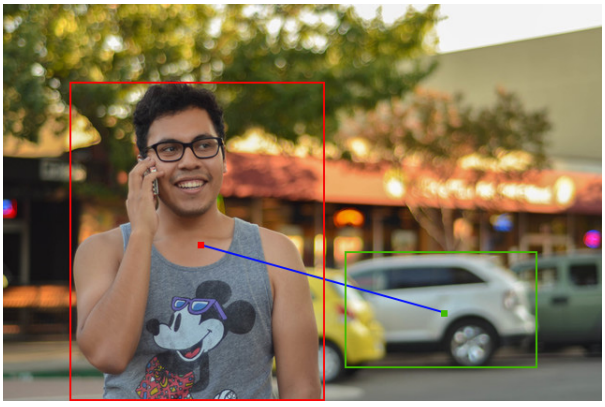
(**a**) ride a bike



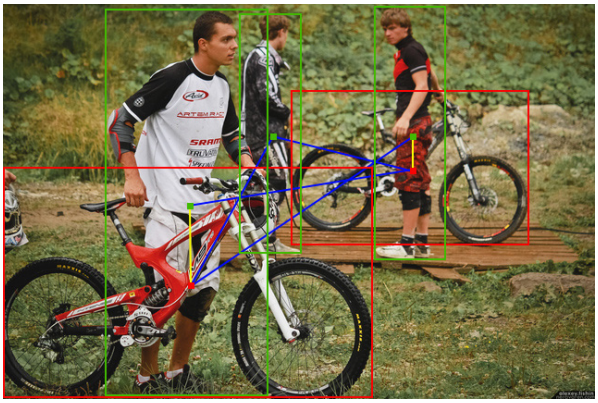
(**b**) ride a motorbike



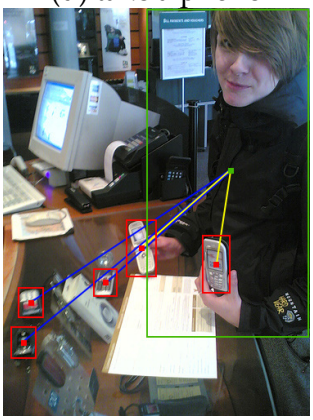
(**c**) stand on skis, hold snow stick



(**d**) take a phone

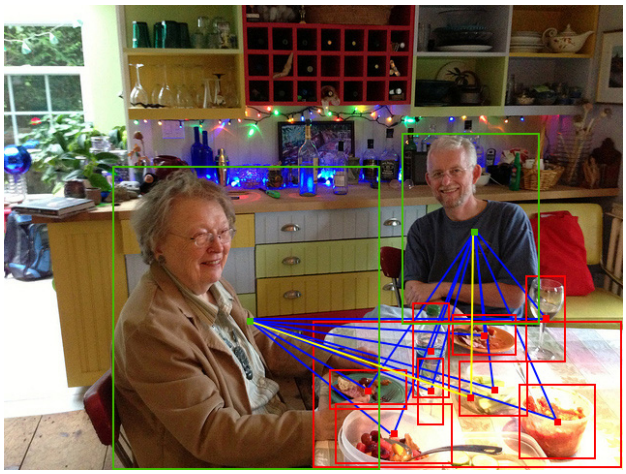


(**e**) hold a bike

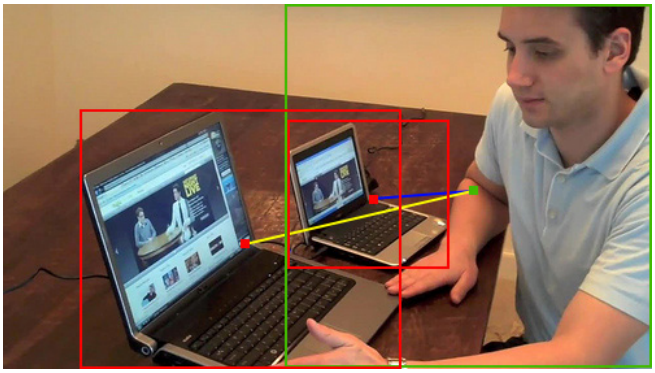


(**f**) hold a phone

Figure 7. *Cont.*



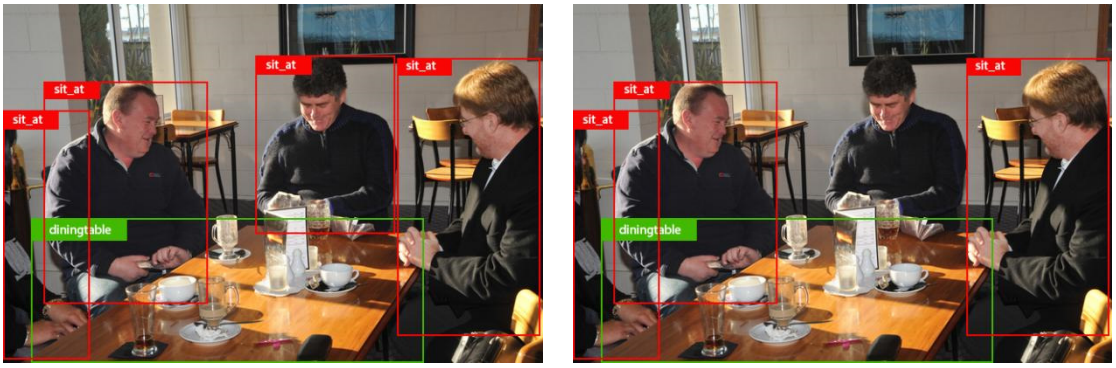
(**g**) sitting at a dining table



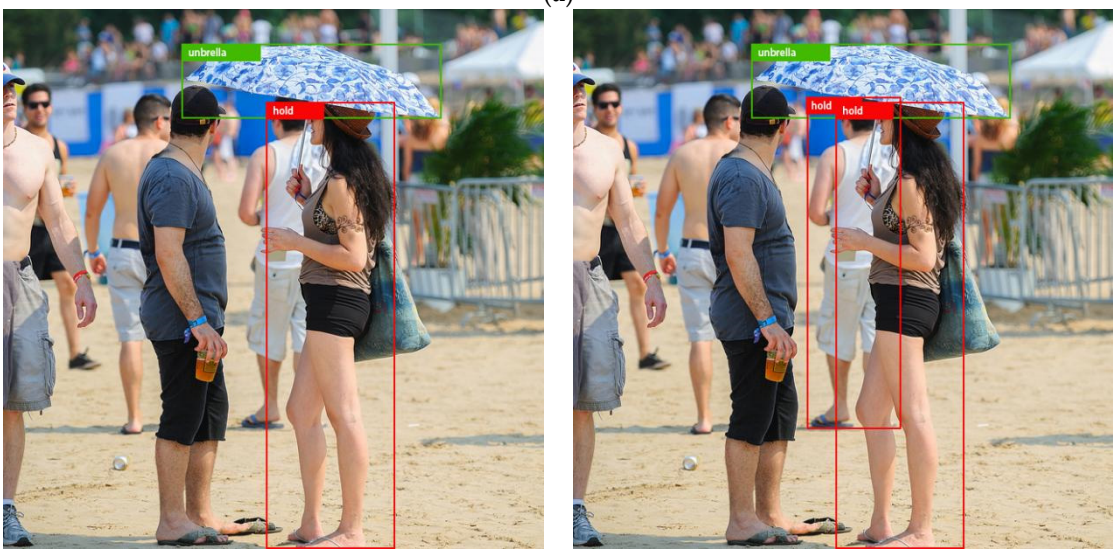
(**h**) watch a laptop

Figure 7. Shows the interaction recognition results on the test sets in the HICO-DET dataset and the V-COCO dataset. The object detection results use green rectangular boxes for persons and red rectangular boxes for objects, with the green solid rectangle being the center coordinates of the person target box and the red rectangle being the center coordinates of the object target box. The interaction detection results use yellow connecting lines to connect the true positive predictions for the person corresponding to the interacting object and blue connecting lines to connect the noninteracting false positive predictions for the person and object.

In addition to this, the visualization result graphs of DRG+IR-GNN and DRG [19] were compared, thus illustrating more clearly the advantages of our method. Figure 8 shows the visualization results of the detection of DRG+IR-GNN and DRG on the V-COCO dataset. As shown in Figure 8, the method in this paper utilized the human pose, the correlation between human–object pairs and scene information, and identified the human–object interaction in advance, which could effectively avoid the problem of false detection and missed detection in DRG [19]. In Figure 8a, people are sitting around a dining table. Using the scene information and the correlation between human–object pairs, the method in this paper could detect four people having an interaction with the table, enabling DRG to detect <sit at, dining table> and avoiding the problem of missed detection. In Figure 8b, the body posture feature was used, and the hand position of the misidentified person was very far away from the umbrella. The method in this paper removed the misidentified person, which could avoid the <hold, umbrella> misidentification problem.



(**a**)



(**b**)

Figure 8. (**a**) People are sitting around a dining table; (**b**) many people under umbrellas; in each subfigure, the left figure shows the DRG + IR-GNN method, the right figure uses the DRG method.

6. Conclusions and Future Work

In this paper, we propose a novel interactivity recognition graph neural network around a large number of noninteractive human–object pairs as negative samples that severely affect the performance of HOI detection and identify valid interactive human–object pairs for increasing the detection accuracy of HOI. We designed the human posture feature module to enhance the fine spatial features of interactive actions, which helped to combine contextual information to predict interactivity. A human–object pair graph model is proposed to construct a graph model with human and object features as nodes and human–object relationships as edges, with the spatial relationship of human–object distances as the initialization weights of edges, and the graph is updated by combining the message passing of the attention mechanism so that the edges of nodes with interactions obtain higher weights. Finally, the node and edge information of the human–object pairs are concatenated with the human posture features, and the score of the HOI or not is derived by the classification model.

This paper draws the following conclusions:

1. Eliminating invalidly interacting human–object pairs before HOI model inference and subjecting only validly interacting human–object pairs to HOI inference helps to improve HOI inference performance. We used our method to improve on existing state-of-the-art methods and conducted comparative experiments on the state-of-the-art methods. As can be seen from Figures 5 and 6, the improved method was significantly better than the original method.

2. The human posture information, attention mechanism, human-to-object distance spatial features as initial edge weights, and graph update operations in this paper's approach all had an enhancing effect on the performance of the model. As can be seen from Table 1, significant results were obtained from ablation experiments in the HICO-DET and V-COCO datasets.
3. The higher the rate of excluding false positive samples, the higher the accuracy of human–object interaction detection will be. As can be obtained from Tables 2 and 3, our method performed better on excluding false interaction pairs, and on subsequent interaction detection experiments, our method removed more invalid interaction pairs, resulting in better results on the accuracy of human–object interaction detection.

Overall, we have made the following contributions in this study.

We propose a novel neural network for interactivity recognition graphs to improve the accuracy of human–object interaction detection. Through experimental and comparative validation, it was demonstrated that our proposed model performed superiorly on the V-COCO and HICO-DET datasets.

In future work, we expect to design a two-stage human–object interaction recognition task based on this approach.

Author Contributions: Conceptualization, J.Z.; methodology, J.Z.; software, J.Z.; validation, J.Z.; formal analysis, J.Z.; investigation, J.Z.; resources, J.Z.; data curation, J.Z.; writing—original draft preparation, J.Z.; writing—review and editing, J.Z., H.H. and Z.M.Y.; visualization, J.Z.; supervision, H.H. and Z.M.Y.; project administration, H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sunaina, S.; Kaur, R.; Sharma, D. A Review of Vision-Based Techniques Applied to Detecting Human-Object Interactions in Still Images. *J. Comput. Sci. Eng.* **2021**, *15*, 18–33. [CrossRef]
2. Khaire, P.; Kumar, P. Deep learning and RGB-D based human action, human–human and human–object interaction recognition: A survey. *J. Vis. Commun. Image Represent.* **2022**, *86*, 103531. [CrossRef]
3. Li, Y.-L.; Liu, X.; Wu, X.; Li, Y.; Qiu, Z.; Xu, L.; Xu, Y.; Fang, H.-S.; Lu, C. HAKE: A Knowledge Engine Foundation for Human Activity Understanding. *arXiv* **2022**, arXiv:2202.06851. [CrossRef]
4. Ashraf, A.H.; Alsufyani, A.; Almutiry, O.; Mahmood, A.; Attique, M.; Habib, M. Weapons detection for security and video surveillance using cnn and YOLO-v5s. *CMC-Comput. Mater. Contin.* **2022**, *70*, 2761–2775. [CrossRef]
5. Wu, B.; Zhong, J.; Yang, C. A visual-based gesture prediction framework applied in social robots. *IEEE/CAA J. Autom. Sin.* **2021**, *9*, 510–519. [CrossRef]
6. Gkioxari, G.; Girshick, R.; Dollár, P.; He, K. Detecting and recognizing human-object interactions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
7. Gao, C.; Zou, Y.; Huang, J.-B. ican: Instance-centric attention network for human-object interaction detection. *arXiv* **2018**, arXiv:1808.10437.
8. Fang, H.-S.; Cao, J.; Tai, Y.-W.; Lu, C. Pairwise body-part attention for recognizing human-object interactions. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
9. Li, Y.-L.; Zhou, S.; Huang, X.; Xu, L.; Ma, Z.; Fang, H.-S.; Wang, Y.; Lu, C. Transferable interactiveness knowledge for human-object interaction detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
10. Wan, B.; Zhou, D.; Liu, Y.; Li, R.; He, X. Pose-aware multi-level feature network for human object interaction detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019.
11. Kolesnikov, A.; Kuznetsova, A.; Lampert, C.; Ferrari, V. Detecting visual relationships using box attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.

12. Wang, T.; Anwer, R.M.; Khan, M.H.; Khan, F.S.; Pang, Y.; Shao, L.; Laaksonen, J. Deep contextual attention for human-object interaction detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019.
13. Qi, S.; Wang, W.; Jia, B.; Shen, J.; Zhu, S.C. Learning human-object interactions by graph parsing neural networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
14. Wang, H.; Zheng, W.-S.; Yingbiao, L. Contextual heterogeneous graph network for human-object interaction detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
15. Xia, L.-M.; Wu, W. Graph-based method for human-object interactions detection. *J. Cent. South Univ.* **2021**, *28*, 205–218. [CrossRef]
16. Liang, Z.; Liu, J.; Guan, Y.; Rojas, J. Visual-semantic graph attention networks for human-object interaction detection. In Proceedings of the 2021 IEEE International Conference on Robotics and Biomimetics (ROBIO), Sanya, China, 6–9 December 2021.
17. Ulutan, O.; Iftekhar, A.; Manjunath, B. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
18. Zhang, F.Z.; Campbell, D.; Gould, S. Spatio-attentive Graphs for Human-Object Interaction Detection. *arXiv* **2020**, arXiv:2012.06060.
19. Gao, C.; Xu, J.; Zou, Y.; Huang, J.-B. Drg: Dual relation graph for human-object interaction detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
20. Zhou, P.; Chi, M. Relation parsing neural network for human-object interaction detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019.
21. Liu, H.; Mu, T.-J.; Huang, X. Detecting human—Object interaction with multi-level pairwise feature network. *Comput. Vis. Media* **2021**, *7*, 229–239. [CrossRef]
22. Liang, Z.; Liu, J.; Guan, Y.; Rojas, J. Pose-based modular network for human-object interaction detection. *arXiv* **2020**, arXiv:2008.02042.
23. Sun, X.; Hu, X.; Ren, T.; Wu, G. Human object interaction detection via multi-level conditioned network. In Proceedings of the 2020 International Conference on Multimedia Retrieval, Dublin, Ireland, 8–11 June 2020.
24. Liao, Y.; Liu, S.; Wang, F.; Chen, Y.; Qian, C.; Feng, J. 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, Seattle, WA, USA, 13–19 June 2020.
25. Wang, T.; Yang, T.; Danelljan, M.; Khan, F.S.; Zhang, X.; Sun, J. Learning human-object interaction detection using interaction points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
26. Kim, B.; Choi, T.; Kang, J.; Kim, H.J. Uniondet: Union-level detector towards real-time human-object interaction detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
27. Chéron, G.; Laptev, I.; Schmid, C. P-cnn: Pose-based cnn features for action recognition. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2015.
28. Scarselli, F.; Gori, M.; Tsoi, A.C.; Hagenbuchner, M.; Monfardini, G. The graph neural network model. *IEEE Trans. Neural Netw.* **2008**, *20*, 61–80. [CrossRef] [PubMed]
29. Zhou, H.; Zhou, H.; Ren, D.; Xia, H.; Fan, M.; Yang, X.; Huang, H. AST-GNN: An attention-based spatio-temporal graph neural network for Interaction-aware pedestrian trajectory prediction. *Neurocomputing* **2021**, *445*, 298–308. [CrossRef]
30. Chao, Y.-W.; Chao, Y.W.; Liu, Y.; Liu, X.; Zeng, H.; Deng, J. Learning to detect human-object interactions. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (wacv), Lake Tahoe, NV, USA, 12–15 March 2018.
31. Gupta, S.; Malik, J. Visual semantic role labeling. *arXiv* **2015**, arXiv:1505.04474.
32. Zhao, Z.-Q.; Zheng, P.; Xu, S.T.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef]
33. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1137–1149. [CrossRef]
34. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.
35. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
36. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
37. Gupta, T.; Schwing, A.; Hoiem, D. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019.
38. Bobick, A.F.; Davis, J.W. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 257–267. [CrossRef]
39. Weinland, D.; Ronfard, R.; Boyer, E. Free viewpoint action recognition using motion history volumes. *Comput. Vis. Image Underst.* **2006**, *104*, 249–257. [CrossRef]

40. Gorelick, L.; Blank, M.; Shechtman, E.; Irani, M.; Basri, R. Actions as space-time shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 2247–2253. [CrossRef]
41. Liu, J.; Kuipers, B.; Savarese, S. Recognizing human actions by attributes. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011.
42. Marszalek, M.; Laptev, I.; Schmid, C. Actions in context. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
43. Wu, Q.; Teney, D.; Wang, P.; Shen, C.; Dick, A.; Hengel, A.V.D. Visual question answering: A survey of methods and datasets. *Comput. Vis. Image Underst.* **2017**, *163*, 21–40. [CrossRef]
44. Zou, C.; Wang, B.; Hu, Y.; Liu, J.; Wu, Q.; Zhao, Y.; Li, B.; Zhang, C.; Zhang, C.; Wei, Y.; et al. End-to-end human object interaction detection with hoi transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11820–11829.
45. Kim, B.; Lee, J.; Kang, J.; Kim, E.-S.; Kim, H.J. Hotr: End-to-end human-object interaction detection with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 74–83.
46. Zhang, Y.; Pan, Y.; Yao, T.; Huang, R.; Mei, T.; Chen, C.-W. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 19548–19557.
47. Wang, H.; Jiao, L.; Liu, F.; Li, L.; Liu, X.; Ji, D.; Gan, W. IPGN: Interactiveness Proposal Graph Network for Human-Object Interaction Detection. *IEEE Trans. Image Process.* **2021**, *30*, 6583–6593. [CrossRef]
48. Wang, M.; Zheng, D.; Ye, Z.; Gan, Q.; Li, M.; Song, X.; Zhou, J.; Ma, C.; Yu, L.; Gai, Y.; et al. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv* **2019**, arXiv:1909.01315.
49. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.