AI

MDPI

Article

Integrating Pose Features and Cross-Relationship Learning for Human–Object Interaction Detection

Lang Wu [1], Jie Li [2,*], Shuqin Li [3], Yu Ding [2], Meng Zhou [1] and Yuntao Shi [1]

[1] School of Electrical and Control Engineering, North China University of Technology, Beijing 100144, China; lw12samuel@mail.ncut.edu.cn (L.W.); zhoumeng@ncut.edu.cn (M.Z.); shiyuntao@ncut.edu.cn (Y.S.)
[2] School of Information, North China University of Technology, Beijing 100144, China; dingyu@ncut.edu.cn
[3] Institute of Science and Technology, North China University of Technology, Beijing 100144, China; lsq@ncut.edu.cn
* Correspondence: lijie1986@ncut.edu.cn

Abstract: Background: The main challenge in human–object interaction detection (HOI) is how to accurately reason about ambiguous, complex, and difficult to recognize interactions. The model structure of the existing methods is relatively single, and the image input may be occluded and cannot be accurately recognized. Methods: In this paper, we design a Pose-Aware Interaction Network (PAIN) based on transformer architecture and human posture to address these issues through two innovations: A new feature fusion method is proposed, which fuses human pose features and image features early before the encoder to improve the feature expression ability, and the individual motion-related features are additionally strengthened by adding to the human branch; the Cross-Attention Relationship fusion Module (CARM) better fuses the three-branch output and captures the detailed relationship information of HOI. Results: The proposed method achieves 64.51%$AP^{\#1}_{role}$, 66.42%$AP^{\#2}_{role}$ on the public dataset V-COCO and 30.83% AP on HICO-DET, which can recognize HOI instances more accurately.

Keywords: human–object interaction detection; human posture; transformer decoder; cross-attention relationship fusion

check for updates

Received: 27 December 2024
Revised: 22 February 2025
Accepted: 4 March 2025
Published: 12 March 2025

Citation: Wu, L.; Li, J.; Li, S.; Ding, Y.; Zhou, M.; Shi, Y. Integrating Pose Features and Cross-Relationship Learning for Human–Object Interaction Detection. *AI* **2025**, *6*, 55. https://doi.org/10.3390/ai6030055

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

HOI detection is a visual relation detection task. It aims to associate the interaction between people and objects by using the image features of the human body, object and person pairs, in which human pose also plays a crucial role, and then realizes image action classification (Figure 1). This task requires not only accurate localization of people and objects in the image but also identification and classification of object categories and discrimination of the influence of human pose on behavior so as to accurately infer interactive behavior.

With the development of society and the acceleration of urbanization, the types of public places, such as communities, fairs, and shopping malls are gradually increasing, and the composition of residents is becoming more complex. As a result, preventing dangerous events in these places has become increasingly important [1,2]. For example, open robbery (such as attacks with knives or other dangerous tools), theft (such as pickpocketing of pedestrians), and violent conflicts caused by drinking (such as fighting with wine bottles) may all occur in the community. With the aim of more accurately assessing HOI instances and effectively gauging the relationships between people and objects and providing technical support for the judgment of abnormal behavior, an advanced HOI detection method

is needed. This approach should not only focus on fine-grained human-centered action interactions [3–8] (armed threats or imminent attacks), but should also consider multiple complex actions that occur simultaneously [9–12] (e.g., a bicycle hitting a pedestrian). By accurately identifying and analyzing these interactions, we can fill in the gaps in the ability of community surveillance to identify, analyze, and prevent dangerous behaviors so as to effectively prevent them in advance and accurately trace them after the fact.

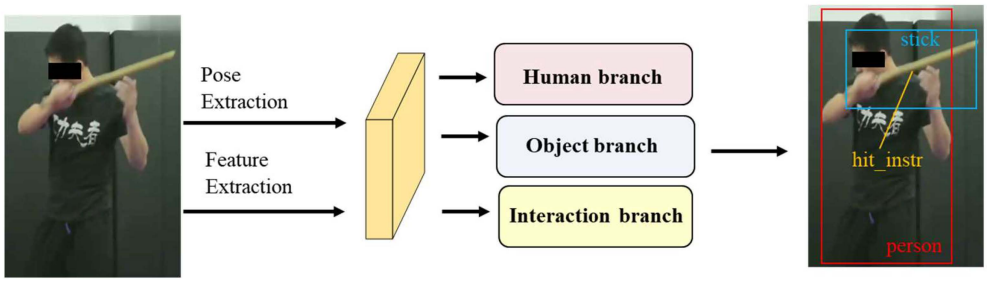


Figure 1. Human–object interaction detection process.

In view of the problems existing in the existing research, in order to enhance the feeling of the interaction relationship, the character interaction area is highlighted. In this paper, we designs a Pose-Aware Interaction Network (PAIN) based on transformer architecture and human posture (see Figure 2), which effectively integrates human pose information and HOI features to improve detection accuracy. Specifically, the human posture detection method is used to obtain the coordinates of two-dimensional skeleton key points, and the adaptive graph convolutional neural network is used to extract the posture features and fuse them with the image features, so that the model combines the visual information in the image and the spatial relationship provided by the posture to form a comprehensive feature representation. The three-branch decoder is used to replace the traditional single transformer decoder, and the posture features are deeply fused in the human detection branch so as to better obtain the potential action features of the human body. In addition, the Cross-Attention Relationship fusion Module (CARM) is designed to unify the information from all three branches, delivering more elaborate interaction details and improving the grasp of interaction relationships. The main contributions of this paper are as follows:

- We design a pose-aware human interaction network to provide a more comprehensive interpretation of human interaction behaviors, including the influence of human pose on the determination of interactive actions, so as to realize effective human interaction behavior detection.
- In PAIN, we propose a novel feature fusion method to address the limitations of existing HOI methods by early fusing 2D human pose features and image features before the encoder to improve feature expressivity and creatively adding individual motion-related features to the human branch, a practice that has not been explicitly explored in existing HOI methods.
- In PAIN, we develop the Cross-Attention Relation Fusion Module (CARM), which uses the cross-attention mechanism to interact with information from multiple inputs; captures the detailed relationship between the three branches of human, object and interaction; and improves the prediction accuracy.
- We conduct experiments to quantitatively and qualitatively verify that our proposed PAIN achieves 64.51%$AP^{\#1}_{role}$, 66.42%$AP^{\#2}_{role}$ on the public dataset V-COCO and 30.83% AP on HICO-DET.

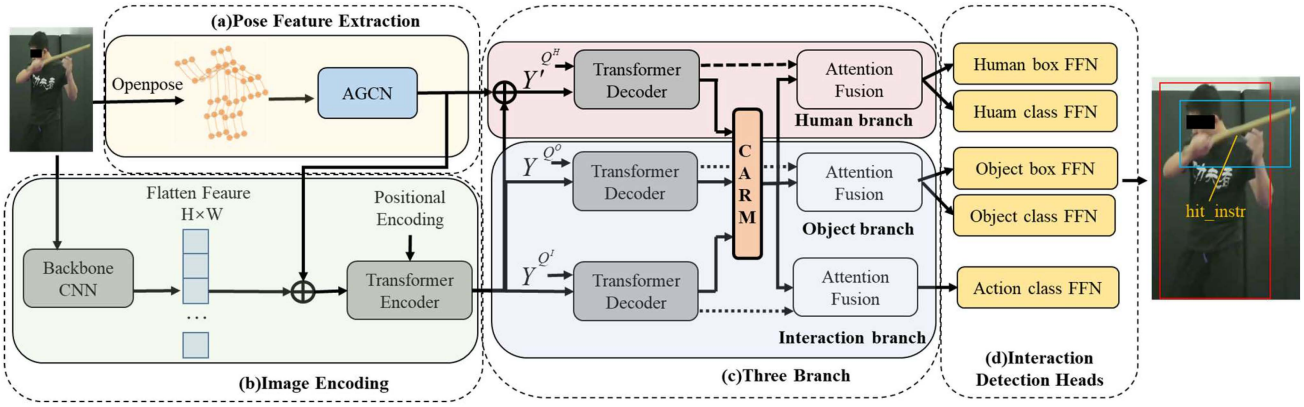


Figure 2. Network architecture of our proposed method.

2. Related Work

At present, the main HOI detection techniques mainly include the traditional one-stage HOI detection, two-stage HOI detection, and transformer-based HOI detection. The two-stage HOI detection methods [3–7,13,14] usually decompose the detection task into an object detection task and an interaction classification task. First, it uses a trained fine-tuned object detector to generate bounding boxes and categories for people and objects. Then, all detected pairs of people and objects are paired, and all pairs are passed to a separate neural network for training and interactive classification. Kim et al. [15] proposed multiple relational networks to perform rich context exchange in three decoder branches. Gao et al. [14] first proposed a two-channel binary image representation to encode spatial relations. Chao et al. [16] proposed the widely used dataset HICO-DET and adopted a two-stage method to propose the HO-RCNN model, which integrated the spatial location information of human–object for the first time to improve the detection effect. Zhou et al. [17] decouple triplet prediction from human–object pair detection and interaction classification through an instance encoder–decoder stream and an interactive encoder–decoder stream. In addition to spatial relations, graph neural networks [3,5] have been proposed to explicitly model the interaction between people and objects, which indeed improves the representation power of the model. Park et al. [18] proposed a novel feature extraction method, overlapping region masking, combined with a pose-conditioned self-loop structure to effectively solve the quantization problem in vision transformers.

Compared with the two-stage methods, the time complexity of the one-stage methods [19–22] is greatly reduced because all pairs of human combinations do not need to be trained, but they still require complex post-processing to combine object detection results and interaction predictions. In these methods, designing a reasonable matching pattern is the key to match the object detection and interaction detection results. Liao et al. [20] and Wang et al. [21] consider HOI as a point detection task, enabling the direct identification of interactions in a single-stage process by introducing a new definition of interaction points. Kim et al. [19] proposed a one-stage anchor-based interaction detection framework, in which the network directly captures the interaction region to detect the interaction, without the matching phase. The joint-level detection framework is used to directly capture the interaction region, and the instance-level detector is used to perform object detection and action classification. Fang et al. [23] proposed a dense interaction region selection framework (DIRV), which focuses on the interaction regions of human–object pairs.

At present, some transformer-based end-to-end HOI detection algorithms [9–12,15,24,25] regard HOI detection as a set prediction problem. Aaron et al. [26] decode N objects in parallel by replacing LSTM with a transformer and using transformer parallel decoding. Cheng et al. [11] proposed an HOI transformer to handle HOI detection in an end-to-end

manner. It simplifies the HOI pipeline and eliminates the need for many hand-designed components. The HOI transformer reasons about object and human relationships from the global image context, introduces a quintuple matching loss to enforce HOI prediction in a unified manner, and directly predicts HOI instances in parallel. Chen et al. [27] improve upon the transformer-based approach by utilizing query-based anchors to derive HOI embeddings and forecast HOI instances. In this method, interactive queries of random parameters (learnable position-embedding sequence, learnable query mechanism) are fed into the transformer decoder to directly map a set of HOI predictions, which makes it tend to detect high-confidence target regions and ignore regions where interactive actions occur. At the same time, such interactive queries are usually unknown at the beginning of prediction, lacking intuitive interaction relationships and human poses with strong topology. Ma et al. [28] introduced a novel staged training strategy to reduce the training pressure caused by the complexity of the task and proposed the HOI-SDC dataset to deal with the challenges in HOI detection. Chen et al. [29] proposed an uncertainty-aware robust HCI learning method, which aims to refine detection and interaction prediction by estimating prediction uncertainty during training.

Many one-stage HOI detection methods [10,30–32] tend to divide the original task into multiple subtasks by using a two-branch or three-branch network structure. Fang et al. [30] separate the detection process of people and objects and emphasize the importance of human features in interaction by introducing a human-guided link method. At the same time, they adopt a stopping gradient mechanism to manage the influence of interaction on detection so as to optimize the detection effect of people and objects. Chan et al. [31] proposed a one-stage three-branch parallel HOI detection method to mitigate the noise interference generated during the fusion process by fusing a noise suppression module. Wu et al. [32] integrate two key branches: time-enhanced recurrent graph network (TRGN) and parallel transformer encoder (PTE), aiming to extract rich hierarchical temporal features from video data. However, not enough attention has been paid to the information exchange between these branches, and a fully parallel approach would destroy the association between different pieces of information.

3. Method

This study aims to improve the diversity of different individual interactions in HOI detection. By strengthening the network structure and optimizing the feature extraction method, we aim to improve the attention of the algorithm to the human pose in the HOI instances so as to further improve the detection accuracy. In this paper, we designs a Pose-Aware Interaction Network (PAIN) based on transformer architecture and human posture, and the network structure is shown in Figure 2. The architecture of PAIN is divided into four components: OpenPose obtains 2D skeleton key point coordinates and uses AGCN to extract pose features. Image features are fused with pose features, and deep feature learning is performed by transformer encoder. Three-branch transformer decoder of human, object and interaction, and pose features are added to the human branch. The CARM generates cross-relation contexts for relation inference. The attention fusion module passes the output of CARM to each subtask for context exchange. The quintuple HOI detection head directly outputs HOI instances.

3.1. Pose Feature Extraction

To achieve a more accurate extraction of posture features, OpenPose is used to generate the two-dimensional key point coordinates of the human body [33]. The generated data format follows the COCO format. For each human body, 25 key points are provided, and each key point is composed (x, y) [34] of two-dimensional coordinates. A spatial

graph (Figure 3) is constructed. In the image, each significant point of the human body is represented as a node on a graph, while the lines connecting these points are treated as edges. By employing this approach, multiple connected graphs can be derived, thereby leveraging the interdependencies among all nodes, as illustrated in Figure 3.

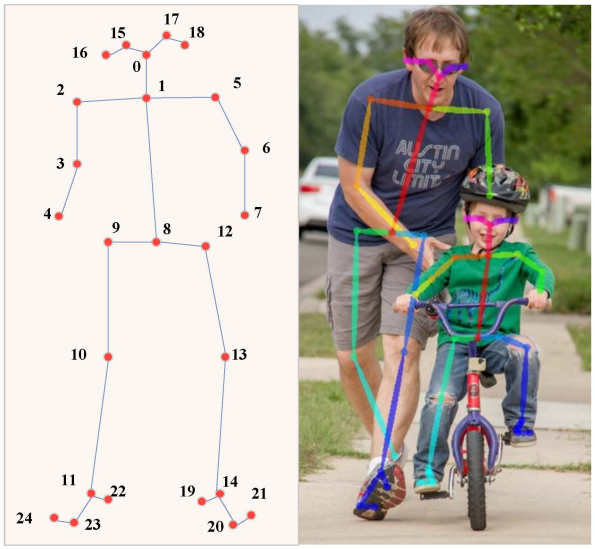


Figure 3. Spatial diagram of bone keys.

The adaptive graph convolutional network is used to extract the posture features and obtain the position coordinates of all key points of the human body. The bone point data form is generally expressed as a second-order tensor of $C \times J$, where C and J are the two-dimensional coordinates of the key points and the total number of all key points, respectively. The coordinates and the number of key points are taken as the data input $X_{in} \in \mathbb{R}^{C \times J}$ to encode and embed the skeleton points. The ReLU activation function and the FC layer are used to obtain the graph convolution input, which is denoted by f_{in}:

$$f_{in} = \delta(\lambda(\delta(\lambda(x)))) \tag{1}$$

where $f_{in} \in \mathbb{R}^{C_{in} \times J}$, c_{in} indicates the number of input channels utilized by the graph convolutional network, whereas δ represents the ReLU activation function, and λ represents the FC.

Since the number of human bodies in each HOI sample is different and the size of J is different, the calculation method [13] of the graph convolution module is similar to 2s-AGCN, but the partition set of spatial skeleton key points is set to 1, the data and graph convolution network are used to automatically learn to generate the adjacency matrix, and then the graph convolution module is passed through three layers in order. The formula for each layer is as follows:

$$f_{out}^{l+1} = f_{in}^{l} + (A + B) f_{in}^{l} W \tag{2}$$

where $W \in \mathbb{R}^{C_{in} \times C_{out}}$ is the parameter matrix of graph convolution and weight vector obtained by 1×1 convolution operation, $(A + B) \in \mathbb{R}^{J \times J}$ is the adjacency matrix, in which B is completely to learn, the initial value is 1, and adaptive adjustment can be used in training and is used to represent the attention of the connection strength between two nodes.

The A matrix is an adjacency matrix related to the data, which is represented by the similarity between nodes and is used to determine whether there is a connection between

two nodes and their connection strength. The similarity between nodes is calculated by using the normalized Gaussian function:

$$A = softmax(f_{in}^T W_\theta^T W_\phi f_{in}) \tag{3}$$

where the normalized Gaussian function calculation is equivalent to the matrix operation of the SoftMax function, and W_θ and W_ϕ are the parameters associated with the Gaussian embedding function. After conducting three graph convolution operations, the result is the pose feature embedding P.

3.2. Image Encoding

The CNN backbone network ResNet was used to extract the visual features [35], and the 3-channel RGB image is input into the backbone network to generate the feature map of the shape $x \in R^{C \times H \times W}$, where H and W represent the height and width of the image, respectively, and C is the feature dimension. To reduce the feature dimension from $C = 2048$ to $d = 256$, a 1×1 projection convolutional layer is used, and the feature map of shape $x_d \in R^{d \times H \times W}$ is generated, which has d channels and spatial dimensions $H \times W$. In order to meet the input requirements of the subsequent transformer encoder for the feature sequence, we apply the flattening operator to collapse the spatial dimension into a single dimension, denoted as the flattening feature $x_S \in R^{1*d \times H \times W}$.

The encoder layer is built on top of the standard transformer architecture [17], and to enable it to distinguish relative positions in a sequence, positional encodings are added to the input of each attention layer. The sum of the flattened image features and pose features embedded in P fusion and position encoding are fed into the transformer encoder to summarize the global information, generating image tokens $Y \in \mathbb{R}^{T \times d}$, T for the subsequent network.

3.3. Inference on Three-Branch HOI Decoding

3.3.1. HOI Decoder

The original single decoder can only deal with a single feature of the input but cannot capture the complex interaction between people and objects at the same time and cannot effectively use the interaction information between people and objects. We adopt a three-branch architecture [15] to replace the original architecture, which is responsible for human detection, object detection, and interaction classification, respectively. Different from the previous three-branch architecture, additional pose features are added to the human detection branch to better understand the influence of human pose on instance. The human detection branch τ has l layers, where the input is query vector $Q^H \in \mathbb{R}^{N \times d}$, the fused feature Y' of the image feature Y, and the pose feature embedding P. In each layer, the transformer decoder updates the input query vector through attention, thereby generating an output D_l^H, which contains contextual information for the subtask responsible for the prediction branch:

$$D_l^H = \left\{ k_{l,i}^H = Decoder_l\left(Q = Q^H, K = V = Y'\right) \right\}_{i=1}^N \tag{4}$$

The object detection and interaction classification branches $v \in (O, I)$ also have l layers, whose inputs are the query vector $Q^O, Q^I \in \mathbb{R}^{N \times d}$ and image features Y. In each layer, the transformer decoder updates the input query vector Q^v with attention Y to produce outputs D_l^v that contain contextual information for predicting the subtask responsible for the branch:

$$D_l^v = \left\{ k_{l,i}^v = Decoder_l(Q = Q^v, K = V = Y), v \in O, I \right\}_{i=1}^N \tag{5}$$

3.3.2. HOI Relationship Fusion

In order to fuse the output information generated by the three branches for different tasks and combine the relationship information carried by them, we propose the CARM, as shown in Figure 4. In this module, the pairwise relations are formed by concatenating the separately generated unary relations, and the cross-attention mechanism is used to perform relation reasoning for unary, pairwise, and ternary relation groups so as to fully mine the useful information in the context. The unary relations generated by the three branches are concatenated in pairs, and the generated unary, binary, and ternary relation groups are respectively reasoned with the cross-attention mechanism so as to make full use of the useful information in the context.

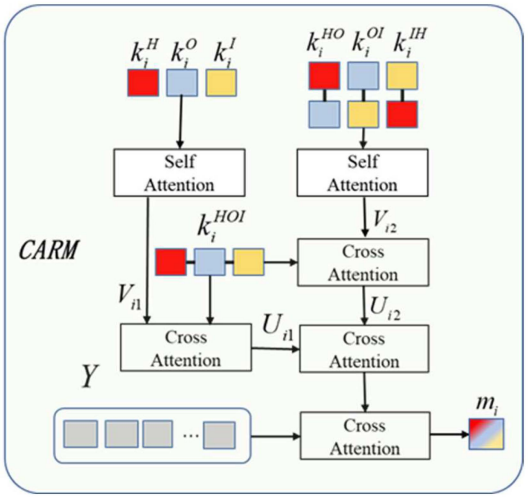


Figure 4. The architecture of the Cross-Attention Relation fusion Module (CARM) takes image features and branch-specific features as input, and the ternary relation information is embedded into the unary and binary relation information, respectively.

Specifically, each output k_i^H, k_i^O, k_i^I of the decoder is concatenated and then passed through the MLP layer to generate the binary relation group and ternary relation group of the i-th HOI instance:

$$\begin{aligned} k_i^{HO} &= MLP\left[k_i^H : k_i^O\right] \\ k_i^{OI} &= MLP\left[k_i^O : k_i^I\right] \\ k_i^{HI} &= MLP\left[k_i^H : k_i^I\right] \\ k_i^{HOI} &= MLP\left[k_i^H : k_i^O : k_i^I\right] \end{aligned} \tag{6}$$

where [:] represents the join operation. In order to exploit the detailed understanding of subtasks by unary relations, the joint attention of tuples to two subtasks and the overall understanding of HOI tasks by triples. Self-attention is used on unary relations and tuples to explore the internal information of their sequences:

$$\begin{aligned} V_{i1} &= SelfAttention\left\{\left[k_i^H, k_i^O, k_i^I\right]\right\} \\ V_{i1} &= SelfAttention\left\{\left[k_i^{HO}, k_i^{OI}, k_i^{HI}\right]\right\} \end{aligned} \tag{7}$$

The cross-attention mechanism is used to fuse the unary relation and binary relation with the ternary relation context, respectively:

$$\begin{aligned} U_{i1} &= CrossAttention\left\{q = k_i^{HOI}, k = v = V_{i1}\right\} \\ U_{i2} &= CrossAttention\left\{q = k_i^{HOI}, k = v = V_{i2}\right\} \end{aligned} \tag{8}$$

The same cross-attention mechanism is used to fuse U_{i1} and U_{i2} to produce $k_i^{HOI\prime}$. Finally, by focusing on the image feature Y, the transformed $k_i^{HOI\prime}$ is used to produce a joint context output m_i:

$$k_i^{HOI\prime} = CrossAttention\{q = U_{i1}, k = v = U_{i2}\} \tag{9}$$

$$m_i = CrossAttention\left\{q = k_i^{HOI\prime}, k = v = Y\right\} \tag{10}$$

3.4. HOI Inference

MLP is used to combine the features of each specific task with the joint context output, so that the corresponding context information can be propagated according to the requirements of each subtask. We make use of the channel attention to the necessary context information selected for each subtask. Then, the detailed features are generated by propagating the necessary context information to the task-specific token $k_{l,i}^H, k_{l,i}^v$. The channel attention α_1, α_2 and detailed features β_l^H, β_l^v generated by the human detection branch, the object detection branch, and the interaction relationship branch are shown in the following equation:

$$\begin{aligned} \alpha_1 &= \sigma\left(MLP\left[k_{l,i}^H : m_{l,i}\right]\right) \\ \alpha_2 &= \sigma\left(MLP\left[k_{l,i}^v : m_{l,i}\right]\right) \end{aligned} \tag{11}$$

$$\begin{aligned} \beta_{l,i}^H &= k_{l,i}^H + \alpha_1 \otimes MLP\left[k_{l,i}^H : m_{l,i}\right] \\ \beta_{l,i}^v &= k_{l,i}^v + \alpha_2 \otimes MLP\left[k_{l,i}^v : m_{l,i}\right] \end{aligned} \tag{12}$$

where $\otimes$ denotes element-by-element multiplication, σ denotes the sigmoid function, α denotes channel attention, and $\beta_{l,i}$ denotes refined tokens. $k_{l,i}^H$,$k_{l,i}^v$ comes from formulas (4) and (5).

We define each HOI instance as a quintuple of (human, interaction, object, human box, object box) [11]. The last output β_l^H and β_l^v of each branch are decoding predicted as HOI instances. Each of the five FFNS predicts HOI quintuples, where the input of human confidence and human frame is β_l^H, the input of object confidence and object frame is β_l^O, and the input of interaction classification confidence is β_l^I.

3.5. Loss Functions

Set the HOI instance as a quintuple of $(c_h, c_o, c_i, b_h, b_0)$ [11], where (c_h, c_o, c_i) represents the confidence of the human, interaction, and object classes, and (b_h, b_0) is the bounding box for the human and object. The interaction is approximated with probability p in a given data set.

We denote the predicted HOIs as $H = p_i, i = 1, 2, \ldots, N$ and the ground true HOIs as $G = g_i, i = 1, 2, \ldots, M, \theta, \ldots, \theta$. M denotes the number of actual interactions in the image. The length of the two sets can be made equal by filling in the background truth set.

In each step of training, the best one-to-one match between the set of true values and the current set of predictions should be found first. The following matching cost is designed for HOI:

$$L_{match}\left(g_i, p_{\omega(i)}\right) = \beta_1 \sum_{j \in h,o,r} \alpha_j L_{cls}^j + \beta_2 \sum_{k \in h,o} L_{box}^k \tag{13}$$

where $L_{cls}^j = L_{cls}\left(g_i^j, p_{\omega(i)}^j\right), j \in h, o, r$ denotes the loss between humans, object and interaction are calculated using standard SoftMax cross-entropy loss, L_{box}^k is the regression loss of human and object frames and is the weighted sum of GIOU loss and L_1 loss. α and β are the hyperparameters of the loss weight. β_1 and β_2 dominate the classification weight and localization weight, respectively. In the matching process, the classification plays a

more important role than the localization with $\beta_1 = 2, \beta_2 = 1$. In HOI, human appears in every instance of HOI, assuming that person classification is the simplest, set $\alpha_h = 1$, and interaction is more important than object objects, $\alpha_o = 1, \alpha_i = 2$.

Denote the match as an injective function, $\omega_{G \to H}$, where $\omega(i)$ is the index of the predicted HOI assigned to the truth. The matching cost function is defined as follows:

$$L_{cost} = \sum_{i}^{N} L_{match}\left(g_i, p_{\omega(i)}\right) \tag{14}$$

where $L_{match}\left(g_i, p_{\omega(i)}\right)$ is the matching cost between the true label and the prediction. Finally, the Hungarian algorithm [36] is used to find the binary match.

4. Experiments

4.1. Dataset and Evaluation Metrics

Experiments are carried out on HICO-DET [16] and V-COCO [37], two datasets that are widely used in HOI detection. V-COCO is a subset of MSCOCO and consists of 5400 images in the original dataset and 4946 images in the test set. We annotate binary labels for twenty-nine different action categories (five of which do not involve the associated object) and also contain 80 object categories. The proposed method is evaluated on scenario 1 and Scenario 2 according to the evaluation settings, and the average precision of roles in both scenarios ($AP^{\#1}_{role}$ in scenario 1 and $AP^{\#2}_{role}$ in scenario 2) is reported. In scenario 1, the model should compute the bounding box of the occluded object and ignore the predicted bounding box of the occluded object in Scenario 2. HICO-DET contains 47,776 images (38,118 for training and 9658 for testing) and includes more than 150 K human–object pairs. The full, rare, and non-rare splits contain all 600 HOI classes, 138 HOI classes, and contain 80 object categories, 117 action categories, and 600 HOI triples, of which 138 are rare categories (i.e., less than 10 training instances), and the remaining 462 categories are non-rare categories.

In accordance with standard evaluation guidelines, we leverage the commonly utilized role mean average precision (mAP) to analyze the performance of the model. According to the P–R curve, we can obtain a numerical metric by averaging the precision values corresponding to each recall value: Average Precision (AP) is a measure of how well the trained model detects the class of interest. In the human–object interaction task, HOI detection results are considered as correct positive samples only if the Intersection Over Union (IOU) ratio between the predicted bounding box and the corresponding real bounding box is greater than 0.5 and both the object class and the action class are predicted correctly.

4.2. Implementation Details

The experiments in PAIN are carried out on the backbone network ResNet-50. The model is trained with AdamW [35], with a transformer learning rate established at 1e–4, a backbone learning rate of 1e–5, and weight decay set to 1e–4. The number of encoding layers and decoding layers are both set to six, and the number of queries N for HICO-DET is set to 64, and the number of queries for V-COCO is set to 100. The model is trained for 150 epochs and the learning rate decays at 90 epochs. The DETR [38] model pre-trained by MS-COCO is used to initialize the weights of the backbone and transformer codec. The batch size is set to four.

4.3. Results Analysis

In the experimental part of this paper, we conduct a systematic evaluation on two important benchmark datasets, V-COCO and HICO-DET, to verify the effectiveness and superiority of the proposed method. The experimental results on the V-COCO dataset are

shown in Table 1, where the comparison results with the baseline method HOITrans [11] (the first end-to-end human–object interaction model) and other recent methods are listed in detail. Similarly, Table 2 lists the experimental results on the HICO-DET dataset. To ensure the comparability of the experiments, we all adopted ResNet-50 as the backbone network. For the HICO-DET dataset, the object detector is pre-trained on MS-COCO.

Table 1. Performance comparison on the VCOCO dataset. The best score is highlighted in bold.

	Method	Backbone	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
Two-stage methods	iCAN [14]	ResNet-50	45.3	52.4
	RPNN [3]	ResNet-50	47.5	-
	TIN [13]	ResNet-50	48.7	54.2
	UnionDet [19]	ResNet-50-FPN	47.5	56.2
	IDN [4]	ResNet-50	53.3	60.3
	ViPLO [18]	ViT-B/16	60.9	**66.6**
One-stage methods	QPIC [9]	Resnet-50	58.8	61.0
	HOI-Trans [11]	Resnet-50	51.15	-
	HOTR [10]	Resnet-50	55.2	64.4
	FGAHOI [28]	Swin-Tiny	60.5	61.2
	MSTR [12]	Resnet-50	62.0	65.2
	CDN-B [24]	Resnet-50	62.3	64.4
	UAHOI [29]	Resnet-50	62.6	66.1
	Ours	Resnet-50	**64.51**	66.42

Table 2. Performance comparison on the HICO-DET dataset. For the detector, COCO means that the detector is trained on COCO, while HICO-DET means that the detector is first trained on COCO and then fine-tuned on HICO-DET. The best score is highlighted in bold.

	Method	Backbone	Detector	Default		
				Full	Rare	Non-Rare
Two-stage methods	iCAN [14]	ResNet-50	COCO	14.84	10.45	16.15
	GPNN [5]	ResNet-101	COCO	13.11	9.41	14.24
	TIN [13]	ResNet-50-FPN	COCO	17.22	13.51	18.32
	DRG [6]	ResNet-50-FPN	COCO	24.53	19.47	26.04
	IDN [4]	Resnet-50	COCO	26.29	22.61	27.39
	GGNet [7]	HG104	HICO-DET	28.83	22.13	30.84
	UnionDet [19]	Resnet-50	HICO-DET	17.58	11.52	19.33
	PPDM [20]	HG104	HICO-DET	21.10	14.46	23.09
One-stage methods	HOTR [10]	Resnet-50	COCO	23.46	16.21	25.62
	HOI-Trans [11]	Resnet-50	COCO	23.46	16.91	25.41
	QPIC [9]	Resnet-50	COCO	29.07	21.85	31.23
	QAHOI [27]	Swin-Base	COCO	29.47	22.24	31.63
	FGAHOI [28]	Swin-Tiny	COCO	29.94	22.24	32.24
	Ours	Resnet-50	COCO	**30.83**	**25.52**	**32.41**

Through the integration of posture and the introduction of the three-branch structure of human, object, and interaction, the model can better capture the subtle changes related to interaction and more complex interaction patterns. The proposed method uses ResNet-50 as the backbone benchmark on the V-COCO dataset, and compared with the single-branch decoder method HOI-Trans, AP^1_{role} increases by 15.54 percentage points (51.15→64.51), and AP^2_{role} reaches 66.42%. It increases 7.37 percentage points (23.46→30.83) on HICO-DET; in particular, there is a significant improvement in the "rare" category, which indicates the potential advantage of the model when dealing with sparse data, as different actions may have similar pose or visual features, and in some scenes, the human body is occluded. Combining pose features improves the robustness of the model in these cases. It shows that the method in this paper has made substantial improvements on the two benchmarks, confirming the model's effectiveness.

The comparison between the results of the proposed model and the baseline model tested on the VCOCO validation set is shown in Figures 5 and 6, and the results prove that in the model training stage, compared with the original algorithm, the total training loss, GIOU loss, bounding box loss, and behavioral cross-entropy loss of the proposed model converge faster than the original model. The proposed method is more efficient in learning and can quickly adapt to the characteristics of the data. In the inference stage of the model, four images are randomly selected for interactive action detection after the improvement. Through the comparison of the bar charts, it can be seen that the confidence indicators of the four images in recognizing the action are greatly improved, and the randomly selected samples show a significant improvement in the ability of action recognition, which indicates that the proposed method has better performance in dealing with complex HOI inference.

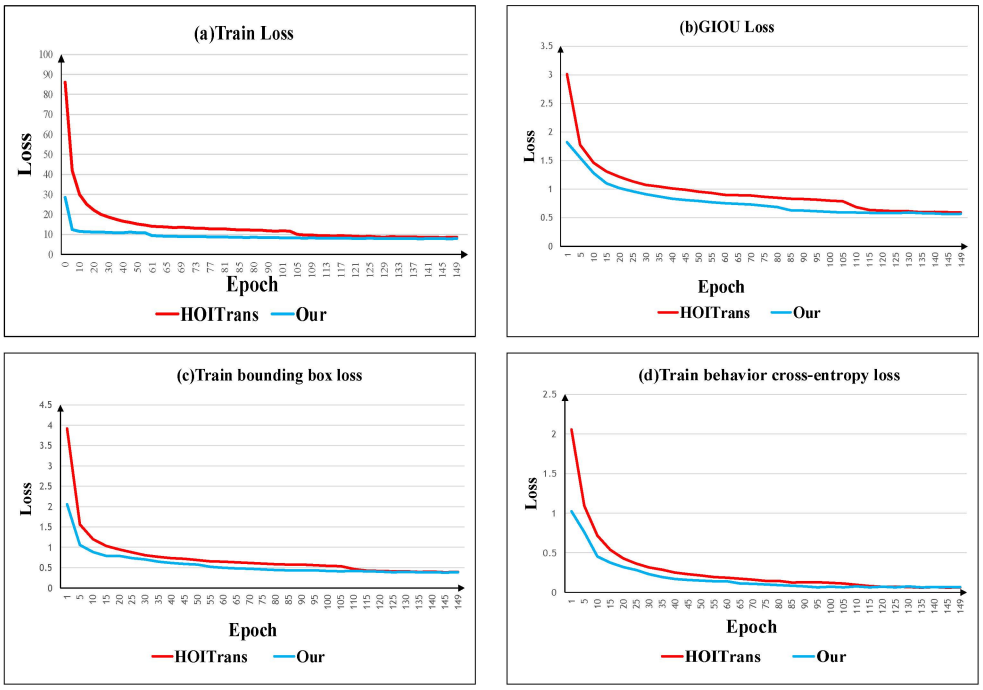


Figure 5. Comparison of model loss with baseline model during training, including total training loss, Generalized Intersection Over Union (GIOU) loss, bounding box loss, and behavior cross-entropy loss.

4.4. Ablation Study

We conducted ablation experiments on the V-COCO dataset to evaluate the performance differences of the model under different feature fusion strategies, verify the effectiveness of early fusion, and remove or replace key components in feature fusion one by one to analyze their impact on the final performance of the model.

4.4.1. Feature Fusion Verification

In order to verify the effectiveness of different fusion methods of posture features in the experiment, posture features are added at different positions to observe their influence on the results. The results in Table 3 demonstrate that the method represented in Figure 7f is the most effective. For HOI detection, real-time use of pose information is crucial to understand complex scenes and action sequences. Passing pose features to the encoder in advance enables the self-attention mechanism to conduct more accurate context modeling based on these features. Adding posture features to the human branch instead of other branches can enhance the expression of human-related features and reduce redundant features, thereby improving the overall performance of the model.

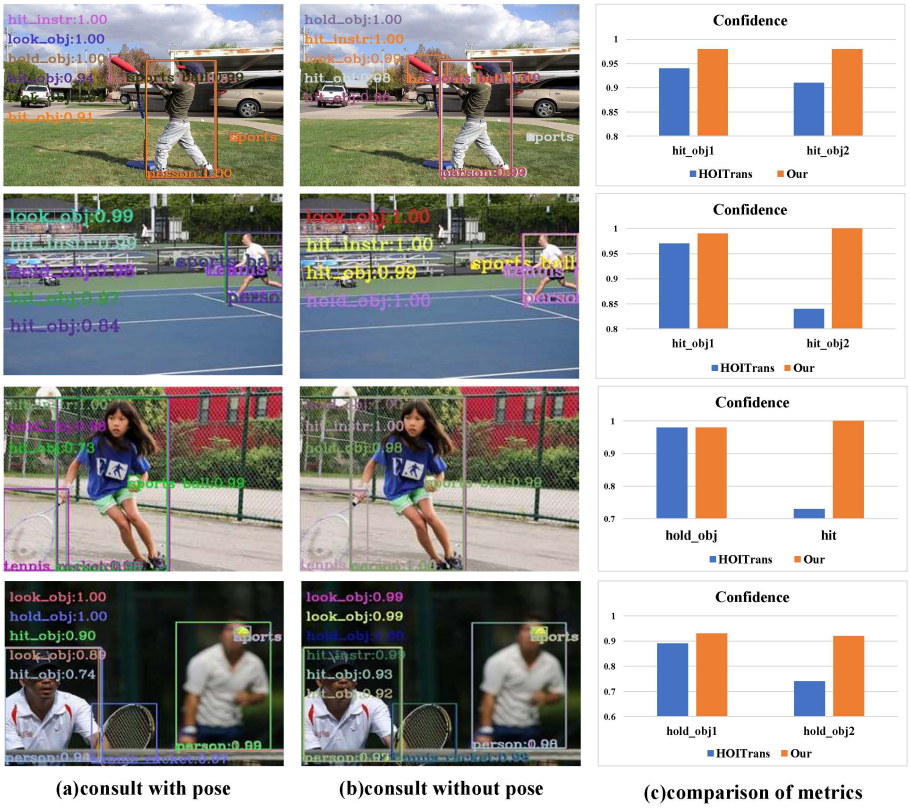


Figure 6. Influence of pose features on model inference results: (**a**) result graph with pose features, (**b**) result graph without pose features, (**c**) bar graph of confidence comparison for each HOI action.

Table 3. Results of adding poses at different positions on the V-COCO dataset.

Method	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
a	63.74	65.64
b	64.16	65.97
c	63.13	64.91
d	63.67	65.54
e	64.48	66.34
f(ours)	64.51	66.42

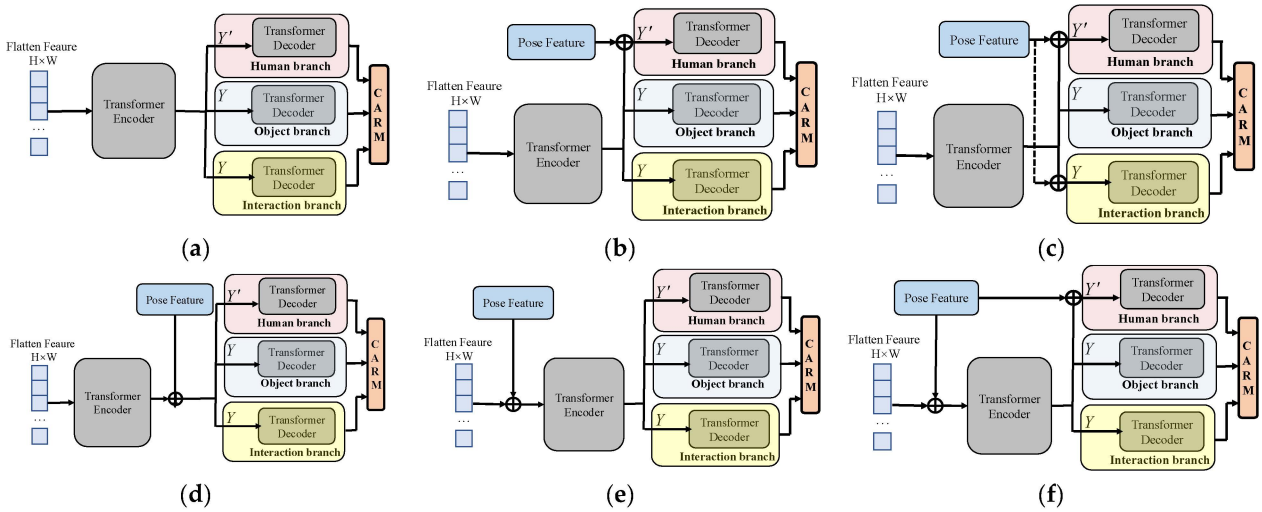


Figure 7. Illustration of adding pose features at different locations: (**a**) no pose features are added; (**b**) add c in the human branch; (**c**) added in the human and interaction branch; (**d**) added on all three branches; (**e**) added to the input of the encoder; (**f**) added in the input and human branches of the encoder.

4.4.2. Component Impact Analysis

In our ablation experiments, the impact of adding different components separately on the accuracy of the model was first evaluated. The experimental results are shown in Table 4, and the results show that the accuracy of the model is improved by 0.87% after only adding posture features. Adding only the CARM module increases the model's accuracy by 1.8%. Finally, when adding pose features and the CARM module at the same time, the accuracy of the model is improved by 2.57% in total.

Table 4. Ablation tests for each component on the V-COCO dataset.

Pose	CARM	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
-	-	61.94	63.88
√	-	62.81	64.50
-	√	63.74	65.64
√	√	**64.51**	**66.42**

These results show that pose features make a significant contribution in capturing action details, and the CARM module further enhances the feature representation ability of the model. Combining the capabilities of the two modules, the overall performance of the model is significantly improved. This further verifies the effectiveness and importance of the feature fusion strategy in improving complex action recognition tasks.

5. Conclusions

This paper designs a pose-aware interaction network (PAIN) to improve HOI detection. Some previous methods tend to detect target regions with high confidence and ignore regions where interactive actions occur. At the same time, it lacks the interaction relationship that conforms to intuitive feeling and the human posture with strong topology. In order to solve these problems, the pose features were extracted and fused with image features. The three-branch decoder was used to obtain richer relationship contexts, and the posture features were integrated into the human detection branch to better obtain the potential action features of the human body. CARM is designed to fuse the information of the three branches to provide more detailed HOI information. Experimental results on V-COCO and HICO-DET datasets show that PAIN is superior to all two-stage methods and common one-stage methods, which achieves 63.74% $AP^{\#1}_{role}$, 65.64% $AP^{\#2}_{role}$ on the public dataset V-COCO and 30.83% AP on HICO-DET.

HOI detection can be very computationally demanding, especially when processing high-definition video streams from multiple cameras, and with edge-computing strategies, preprocessing and preliminary processing can be performed near the data source, spreading the computational load and minimizing network latency. In future work, we plan to further deepen the proposed method for HOI detection and explore the application of 3D human pose information. 3D pose information can provide a richer context to help the model better understand and recognize complex interactive behaviors. It is a technical difficulty to effectively fuse 2D image features with 3D human pose information, and it is necessary to ensure the relevance between them. At the same time, the introduction of 3D pose data will increase the computational burden, especially in real-time applications, where the accuracy and processing speed need to be balanced. Edge-computing strategies can be used to distribute the computational load and minimize network latency by performing pre-processing and preliminary processing near the data source. Through the enhancement of this technology, we hope to reconstruct dynamic scenes and provide a richer environmental context for the community so as to realize more comprehensive instances analysis and understanding.

Author Contributions: Conceptualization, L.W.; formal analysis, L.W. and J.L.; methodology, L.W.; investigation, M.Z. and Y.D.; resources, J.L. and S.L.; data curation, S.L.; writing—original draft preparation, L.W. and J.L.; visualization, L.W. and J.L.; supervision, Y.D. and Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key R&D Program of China, Research on risk early warning and precise policy technology of community caregivers based on collaborative perception of multi-source heterogeneous Internet of things devices (2023YFC3304604).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets used in this research are available as open sources at the following links: V-COCO dataset at https://cocodataset.org/#download (accessed on 20 December 2024); HICO-DET dataset at https://umich-ywchao-hico.github.io/ (accessed on 20 December 2024).

Conflicts of Interest: The authors declare no competing interests.

References

1. Tripathi, V.; Mittal, A.; Gangodkar, D.; Kanth, V. Real Time Security Framework for Detecting Abnormal Events at ATM Installations. *J. Real-Time Image Process.* **2019**, *16*, 535–545. [CrossRef]
2. Hatirnaz, E.; Sah, M.; Direkoglu, C. A Novel Framework and Concept-Based Semantic Search Interface for Abnormal Crowd Behaviour Analysis in Surveillance Videos. *Multimed. Tools Appl.* **2020**, *79*, 17579–17617. [CrossRef]
3. Zhou, P.; Chi, M. Relation Parsing Neural Network for Human-Object Interaction Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 November 2019; pp. 843–851.
4. Li, Y.-L.; Liu, X.; Wu, X.; Li, Y.; Lu, C. HOI Analysis: Integrating and Decomposing Human-Object Interaction. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 5011–5022.
5. 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.
6. 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*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020.
7. Zhong, X.; Qu, X.; Ding, C.; Tao, D. Glance and Gaze: Inferring Action-Aware Points for One-Stage Human-Object Interaction Detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13229–13238.
8. Wang, H.; Zheng, W.; Yingbiao, L. Contextual Heterogeneous Graph Network for Human-Object Interaction Detection. In *Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XVII; Springer: Berlin/Heidelberg, Germany, 2020; pp. 248–264.
9. Tamura, M.; Ohashi, H.; Yoshinaga, T. QPIC: Query-Based Pairwise Human-Object Interaction Detection with Image-Wide Contextual Information. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 10405–10414.
10. 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 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 74–83.
11. 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 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11820–11829.
12. Kim, B.; Mun, J.; On, K.-W.; Shin, M.; Lee, J.; Kim, E.-S. MSTR: Multi-Scale Transformer for End-to-End Human-Object Interaction Detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2022; pp. 19556–19565.
13. Li, Y.-L.; Liu, X.; Wu, X.; Huang, X.; Xu, L.; Lu, C. Transferable Interactiveness Knowledge for Human-Object Interaction Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3870–3882. [CrossRef] [PubMed]
14. Gao, C.; Zou, Y.; Huang, J.-B. iCAN: Instance-Centric Attention Network for Human-Object Interaction Detection. *arXiv* **2018**, arXiv:1808.10437.

15. Kim, S.; Jung, D.; Cho, M. Relational Context Learning for Human-Object Interaction Detection. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 2925–2934.
16. 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; pp. 381–389.
17. Zhou, D.; Liu, Z.; Wang, J.; Wang, L.; Hu, T.; Ding, E.; Wang, J. Human-Object Interaction Detection via Disentangled Transformer. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 19546–19555.
18. Park, J.; Park, J.-W.; Lee, J.-S. ViPLO: Vision Transformer Based Pose-Conditioned Self-Loop Graph for Human-Object Interaction Detection. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 17152–17162.
19. Kim, B.; Choi, T.; Kang, J.; Kim, H.J. UnionDet: Union-Level Detector Towards Real-Time Human-Object Interaction Detection. In *Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XV; Springer: Berlin/Heidelberg, Germany, 2020; pp. 498–514.
20. 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 (CVPR), Seattle, WA, USA, 13–19 June 2020.
21. Wang, T.; Yang, T.; Danelljan, M.; Khan, F.S.; Zhang, X.; Sun, J. Learning Human-Object Interaction Detection Using Interaction Points. Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
22. Zhang, F.Z.; Campbell, D.; Gould, S. Efficient Two-Stage Detection of Human-Object Interactions with a Novel Unary-Pairwise Transformer. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2021; pp. 20072–20080.
23. Fang, H.-S.; Xie, Y.; Shao, D.; Lu, C. DIRV: Dense Interaction Region Voting for End-to-End Human-Object Interaction Detection. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 1291–1299. [CrossRef]
24. Zhang, A.; Liao, Y.; Liu, S.; Lu, M.; Wang, Y.; Gao, C.; Li, X. Mining the Benefits of Two-Stage and One-Stage HOI Detection. In *Proceedings of the 35th International Conference on Neural Information Processing Systems, Online, 6–14 December 2021*; Curran Associates Inc.: Red Hook, NY, USA, 2021.
25. Zhang, Y.; Pan, Y.; Yao, T.; Huang, R.; Mei, T.; Chen, C. Exploring Structure-Aware Transformer over Interaction Proposals for Human-Object Interaction Detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 19526–19535.
26. van den Oord, A.; Li, Y.; Babuschkin, I.; Simonyan, K.; Vinyals, O.; Kavukcuoglu, K.; van den Driessche, G.; Lockhart, E.; Cobo, L.; Stimberg, F.; et al. Parallel WaveNet: Fast High-Fidelity Speech Synthesis. In Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 3918–3926.
27. Chen, J.; Yanai, K. QAHOI: Query-Based Anchors for Human-Object Interaction Detection. In Proceedings of the 2023 18th International Conference on Machine Vision and Applications (MVA), Hamamatsu, Japan, 23–25 July 2023; pp. 1–5.
28. Ma, S.; Wang, Y.; Wang, S.; Wei, Y. FGAHOI: Fine-Grained Anchors for Human-Object Interaction Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 2415–2429. [CrossRef] [PubMed]
29. Chen, M.; Chen, M.; Yang, Y. UAHOI: Uncertainty-Aware Robust Interaction Learning for HOI Detection. *Comput. Vis. Image Underst.* **2024**, *247*, 104091. [CrossRef]
30. Fang, S.; Lin, Z.; Yan, K.; Li, J.; Lin, X.; Ji, R. HODN: Disentangling Human-Object Feature for HOI Detection. *IEEE Trans. Multimed.* **2024**, *26*, 3125–3136. [CrossRef]
31. Chan, S.; Zeng, X.; Wang, X.; Hu, J.; Bai, C. Auxiliary Feature Fusion and Noise Suppression for HOI Detection. *ACM Trans. Multimed. Comput. Commun. Appl.* **2024**, *20*, 1–8. [CrossRef]
32. Wu, J.; Zhang, Y.; Kampffmeyer, M.; Pan, Y.; Zhang, C.; Sun, S.; Chang, H.; Zhao, X. HierGAT: Hierarchical Spatial-Temporal Network with Graph and Transformer for Video HOI Detection. *Multimed. Syst.* **2024**, *31*, 13. [CrossRef]
33. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 172–186. [CrossRef] [PubMed]
34. Lin, T.-Y.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
36. Kuhn, H.W. The Hungarian Method for the Assignment Problem. *Nav. Res. Logist.* **1955**, *2*, 83–97. [CrossRef]

37. Gupta, S.; Malik, J. Visual Semantic Role Labeling. *arXiv* **2015**, arXiv:1505.04474.
38. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part I; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.