IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 November 2023, accepted 15 November 2023, date of publication 20 November 2023,
date of current version 29 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3335193

RESEARCH ARTICLE

Parallel Multi-Head Graph Attention Network (PMGAT) Model for Human-Object Interaction Detection

JIALI ZHANG, ZURIAHATI MOHD YUNOS, AND HABIBOLLAH HARON, (Senior Member, IEEE)
Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia
Corresponding authors: Jiali Zhang (jiali-20@graduate.utm.my) and Zuriahati Mohd Yunos (zuriahati@utm.my)

ABSTRACT Human-object interaction (HOI) detection is an advanced task in the field of computer vision and is crucial for deep scene understanding. However, current HOI detection models face serious challenges in the following aspects: first, they overly rely on appearance features and neglect the local details of human-object interactions; second, the training cost of the existing detection model is quite high. To overcome these challenges, this study proposes a Parallel Multi-Head Graph Attention Network (PMGAT) model for detecting human-object interaction correlations. First, the close relationship between facial landmarks and body keypoints with objects is recognized, thereby introducing a local feature module to construct a relational graph model between facial keypoints, body keypoints, and objects. A multi-head graph attention network was utilized to accurately capture the interaction correlations between keypoints, addressing the issue of neglecting local details. Furthermore, the global feature module is designed to extract absolute spatial pose features and relative spatial pose features based on the positions of human keypoints relative to objects, enabling a more in-depth extraction of interactions between humans and objects. To reduce the training cost of the model, it adopts a multi-branch parallel structure and employs a multi-threaded multi-GPU scheme for parallel training acceleration. The empirical results demonstrate that the PMGAT model outperforms the current state-of-the-art ViPLO method in terms of mAP on the V-COCO and HICO-DET datasets. On V-COCO, it exhibits a notable improvement of up to 0.8% mAP over ViPLO, while on the more demanding HICO-DET, the improvement reaches up to 1.47% mAP. Furthermore, PMGAT stands out for its minimal training time compared to existing approaches. Overall, these results corroborate the dual augmentation of PMGAT in accuracy and training efficiency.

INDEX TERMS Human–object interaction, graph attention network, local feature, deep learning.

I. INTRODUCTION

Research has shown that people often rely on analyzing interactions involving humans and objects in an image to interpret the connotations of the image, and that such interactions include both interpersonal interactions and interactions between objects [1], [2], [3]. The recognition of these interaction relationships is of great significance for advancing the computer-based understanding of image and video content [4], generating descriptions of image and video scenes automatically [5], and enabling automatic questioning of image and video scenes [6].

In earlier research, the fusion of multiple features for human-object interaction (HOI) detection, including visual and spatial features, was a major trend [7], [8]. These studies are significant for the development of HOI detection. However, as research has progressed, more emphasis on interaction details has become important. To address this, various methods have been introduced, such as attention mechanisms [9], [10], [11], [12], context information [13], [14], graph convolutional neural networks [15], [16], [17], [18], [19], body parts, and poses [20], [21], [22], [23], [24] to enhance the focus on local details within images. Specifically,

The associate editor coordinating the review of this manuscript and approving it for publication was Amjad Ali.

the construction of context appearance features has become crucial, in addition to the visual and spatial features of humans and objects. These methods construct attention maps that concentrate on regions relevant to interactions between people and objects, and regions that may contain information about interactions [19], [25], [26].

Moreover, earlier research has often treated the human body as a whole and allocated the same attention to the entire body region. However, this overlooks the fact that humans typically use only a portion of their bodies to interact with objects. Therefore, some studies introduced information about human body parts and poses to improve the accuracy of HOI detection [11], [20], [23]. Furthermore, graph models and graph convolutional neural networks have become essential approaches for addressing HOI detection problems. Previous research combined these network structures with graphical models [27], [28], achieving significant progress in applications such as scene understanding [29], [30], [31], object detection and parsing [32], [33], and Visual Question Answering (VQA) [15]. When conducting human–object interaction detection (HOI detection), the central concept of a graph model is to utilize nodes to represent humans and objects while using edges to denote the interactive relationships between humans and objects. By introducing feature processing networks with attention mechanisms, contextual information from images is integrated into the feature representations of the graph nodes. When there is a stronger level of interaction between humans and objects, the associated weights for the corresponding edges increase [18], [34].

The field of HOI detection has been further enriched by recent advances in the neural heuristic analysis of video content, with the application of graph neural networks becoming increasingly compelling, from the initial generic approach evolved to a nuanced analysis that takes into account specific body parts and poses, while utilizing graph models and attentional mechanisms to improve accuracy. In recent years, the research community has introduced innovative frameworks to address multifaceted challenges that address the problem of modeling interactions at different levels of granularity [35], [36], overcoming the challenges posed by label skew [37], and the complexities of effectively integrating multimodal data [38]. These innovations have significantly advanced application areas such as scene understanding and VQA. These advances mark a crucial role for graph neural networks in HOI detection in video, opening up new possibilities for more accurate and dynamic interpretation of human-object interactions in video.

Although existing HOI detection methods perform well in most cases, the performance of graph-based HOI detection methods may not be satisfactory in certain action contexts. To gain a deeper understanding of this issue, we centered our study around action categories, encompassing a total of 117 different action categories. The sourced images were selected from the HICO-DET dataset as our test dataset [39]. State-of-the-art and widely accepted HOI detection techniques were evaluated under identical hardware conditions. The performances of different action categories were compared using the recall rate of true-positive samples and visualizing these results, as shown in Figure 1. Various methods, including Instance-centric Attention Network (iCAN) [10], Transferable Interactiveness Knowledge (TIN) [21], Dual Relation Graph (DRG) [19], Spatially Conditioned Graphs (SCG) [40], Translational Model for Human-Object Interaction Detection (TMHOI) [41], and Vision Transformer based Pose-Conditioned Self-Loop Graph (ViPLO) [42], have been applied to HOI detection on the HICO-DET dataset, and the recorded recall rates have been documented.

A common problem has detection is insufficient attention to local information, leading to a decrease in detection accuracy. To gain a deeper understanding of this issue, three action categories (watch, inspect, and ride) with the lowest recall rates were focused explicitly on, and representative misclassified samples for observation were selected, as shown in Figure 2. In Figure 2, the red text represents incorrect interactions, while the black text represents correct interaction results.

For example, in Figure 2 (a-b), it can be observed that the human in the red bounding box partially overlaps with the bicycle in the green bounding box. Because the red bounding box was situated above the green bounding box, it was misclassified as the ''ride'' action. In addition, in Figure 2 (c), when the human in the red bounding box was positioned at the front of the bicycle, it was also misclassified as the ''Hold'' action. These misclassification cases highlight the importance of considering human body posture and body part spatial relationships in reasoning about human-object interactions. In Figures 2 (d-g), despite the detection of the presence of a human and a television, they were misclassified as the ''watch'' action. In Figure 2 (h), although three people stood around a bicycle, there was no physical interaction between humans and the object. However, it was still misclassified as having no interactive action. These misclassifications were due to the neglect of local facial details, which could indicate possible interactions with objects.

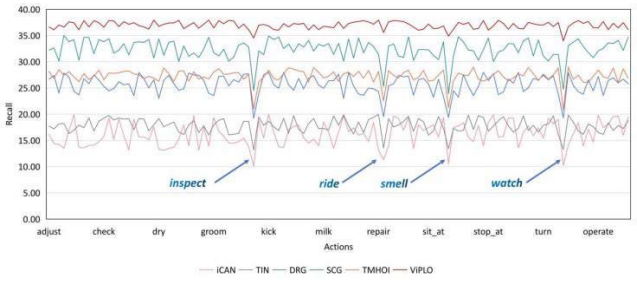


FIGURE 1. Recall rates for 117 different actions.

To address these issues, we believe that integrating human body posture and body part spatial information, as well as local facial details, into embedding a graph provides more interpretable information. While previous graph-based

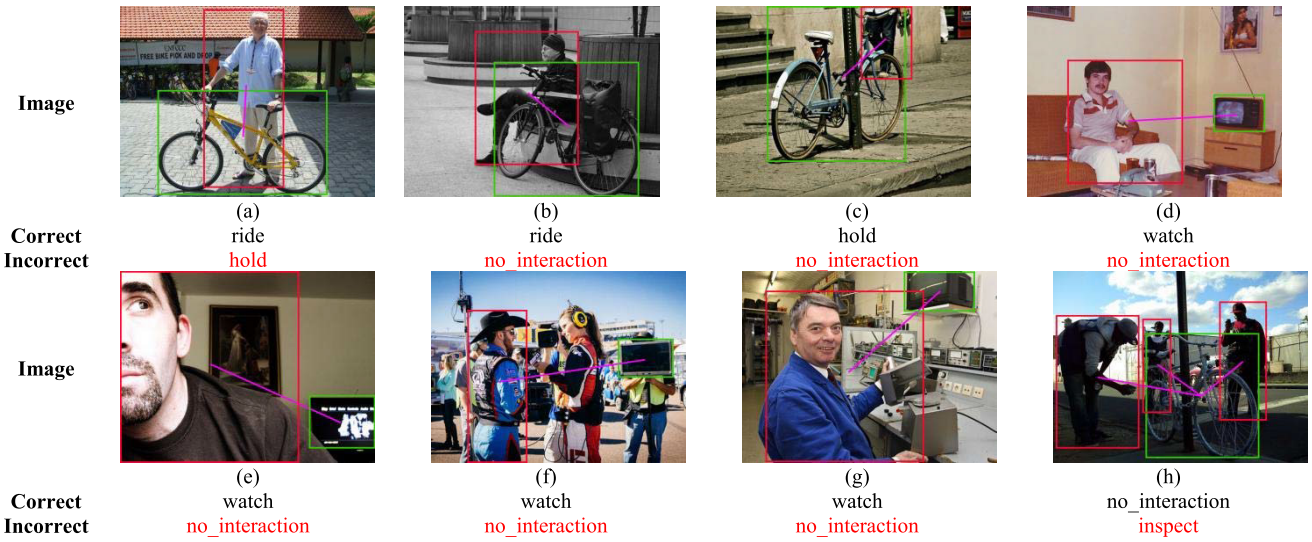


FIGURE 2. Misclassified samples.

methods have attempted to encode human body posture and body part spatial information into node embeddings [14], [18], [43], [44], [45], insufficient attention has been paid to body parts and posture features for interacting objects. In addition, facial part information was not encoded into the nodes of the graph.

Considering the constraints mentioned earlier, this study presents a model for HOI detection utilizing graph structures called the parallel multi-head graph attention network (PMGAT) model for human–object interaction detection. The model architecture, as depicted in Figure 3, primarily consists of five modules: the human–object interaction proposal module, feature extraction module, local feature module, global feature module, and semantic feature module.

Module 1: Human-Object Interaction Proposal Module: This module's main task is to sift out potential human-object pairs that might interact with the images. Its output consists of a set of possible human-object interaction proposals, laying the foundation for subsequent feature extraction and graph modeling.

Module 2: Feature Extraction Module: This module extracts the appearance features of humans and objects, facial part features, and keypoint features of the human body from the images. These features are crucial for identifying interactions between humans and objects. They are passed on to the subsequent graph model to construct a relationship graph between humans and objects.

Module 3: Local Feature Module: In this module Unlike previous methods of feature fusion in graph-based models, our approach integrates fine-grained facial details into graph embeddings. This aids the model in understanding interactions related to faces. Simultaneously, human body part features are embedded into the graph nodes, and a multi-head attention network is employed to update the relationship weights between nodes and aggregate features from neighboring nodes. This directs the model's attention more towards important local details that influence interaction detection.

Module 4: Global Feature Module: In the global feature component Relative spatial poses and absolute spatial poses of human subjects were employed as contextual global features to enhance the semantic understanding of the image.

Module 5: Semantic Feature Module: Finally, semantic features are introduced to further strengthen the semantic comprehension of images.

The key innovation lies in the fusion of facial fine-grained details and human body part features into the node encoding process of the graph, coupled with a multi-head attention network to provide supplementary insights to the graph model and enhance the focus on local details. Regarding the model architecture, we designed the local, global, and semantic features as three independent parallel modules. By leveraging this structural characteristic, a parallel training method was adopted to enhance the training efficiency of the model. In summary, our contributions encompass three main aspects: incorporation of facial and body part details, parallel graph attention networks, and experimental validation.

(1) Incorporated facial and body part details into the graph node encoding process, enhancing the focus of the model on local features.

(2) Developed a multi-branch parallel structure that employs multi-threading and multi-GPU techniques to significantly accelerate the training process.

(3) Conducted experimentation on the V-COCO and HICO-DET datasets, validating the superior performance of the model over existing methods with significant enhancements in detection accuracy and training efficiency.

II. RELATED WORK

The primary objective is to localize both human subjects and objects while simultaneously recognizing their interaction relationships during Human-Object Interaction

(HOI) detection. Over the past few years, there has been a significant proliferation of deep learning architectures in the field of HOI, with numerous HOI detection models currently dependent on neural network frameworks. For instance, Chao et al. [7] introduced HO-RCNN, which is a multi-stream network structure comprising three streams: one for humans, one for objects, and one for pairs. The human and object streams encode the visual attributes of humans and objects, whereas the pair stream represents the spatial connections between humans and objects. This classic multi-stream network structure serves as a benchmark and source of inspiration for subsequent research.

A. ATTENTION MECHANISM

Building upon the HO-RCNN, Gao et al. [10] introduced a model named ICAN. ICAN integrates an attention module focused on individual instances to extract contextual attributes that complement the appearance features within local regions (i.e., bounding boxes for humans and objects). This integration augments the effectiveness of HOI detection. The instance-centric attention map provides the model with greater flexibility, as it allows for focus on different regions of the image based on different object instances. This innovation improves the model's understanding of human-object interactions in complex scenes.

B. MULTI-FEATURE FUSION

However, ICAN relied solely on the visual and spatial features of humans and objects for inference without incorporating additional information, leaving significant room for improvement in terms of accuracy. To address this challenge, Li et al. [21] introduced the TIN model, which incorporates spatial positioning and human pose information to enhance HOI reasoning. TIN is independent of the HOI classification task, has excellent generalization properties, can be transferred across datasets, and has flexible portability. Nevertheless, the model overlooks the substantial differences between human and object appearances and spatial positions, as well as the subtle distinctions between similar relationships. These differences may have a significant impact on the accuracy of reasoning in certain scenarios.

C. TRANSFORMER

Owing to the growing concern over the homogenization of external features, researchers have started to pay more attention to subtle differences among local features. To address this challenge, deep learning models that leverage self-attention mechanisms, known as transformer technologies, have emerged. Initially, Transformer technology achieved tremendous success in natural language processing tasks. However, in recent years, it has been successfully applied to the field of HOI detection [46], [47]. Transformer models have demonstrated remarkable performance in HOI detection, leading to improved comprehension of human-object interactions within images. Consequently, this advancement has enhanced the accuracy and efficiency of the HOI detection. The introduction of this technology provides a new approach for addressing the homogenization of external features in HOI detection.

D. GRAPH MODELS

While the Transformer has achieved good results in HOI detection, it requires significant computational resources, posing challenges for subsequent inference tasks. However, alternatives to the transformer, such as graph-based techniques, have been suggested to tackle its limitations and can be considered a viable approach [19], [40], [41], [42].

The core idea of graph models is to construct a graph in which humans and objects in the image are represented as nodes and connected by edges based on their interaction relationships. Each node contains feature information regarding the corresponding human or object, and the edges represent the associations between humans and objects. Using graph algorithms, graph models leverage this information from nodes and edges to detect and classify interactions between humans and objects.

In recent years, graph model-based HOI detection methods have been developed. There are four popular models based on this graph: DRG, SCG, TMHOI, and ViPLO, as described below.

DRG: Gao et al. [19] first adopted an abstract spatial semantic representation to characterize every individual person-object pair. Subsequently, they introduced DRG to consolidate contextual details derived from the surrounding environment. The DRG model consists of two parts, one focusing on the person and the other on the object. It effectively captures diverse information from a scene to address ambiguities in local predictions. Unlike other methods, DRG leverages the relationships between different HOIs to refine predictions.

SCG: Zhang et al. [40] proposed an SCG and creatively applied spatial relations to information propagation between pairs of nodes. Through a multi-branch fusion mechanism, SCG utilizes the spatial arrangement of person-object pairs to adjust appearance features, improve edge computations, and thereby enhance the quality of HOI detection.

TMHOI: Zhu et al. [41] proposed TMHOI, a method that utilizes a knowledge graph embedding model as a translation model. The purpose is to incorporate relationship features into node embeddings through embedding and integration. This approach not only enhances the representation of nodes but also improves the consistency between node embeddings and edge embeddings, thus enhancing the detection performance.

ViPLO: Park et al. [42] introduced ViPLO, which combines a novel feature extraction method with a two-stage HOI detector assisted by pose-conditioned graphs. ViPLO has the benefits of minimal complexity and ease of application in practical scenarios.

The advantages of graph models in effectively extracting inter-node relationship features have led to significant advancements in human-object interaction (HOI) detection, thereby improving the HOI accuracy. However, existing methods are limited to the visual features of humans and objects, along with human body parts or pose features, without providing fine-grained feature information. In response to this challenge, we propose a parallel multi-head graph attention network (PMGAT) for human-object interaction detection. This model starts by refining the local features, focusing primarily on the local features of faces and human bodies, and encoding them into a graph model for fine-grained graph computations. These further extract subtle features, aiding in mapping the associations between human-object interactions. With the introduction of PMGAT, we can gain a more comprehensive understanding of the interactions between humans and objects, thereby achieving more precise and efficient HOI detection.

III. THE PROPOSED PMGAT MODEL

In this study, a Parallel Multi-head Graph Attention Network (PMGAT) is employed for detecting human-object interaction relationships. The model structure of PMGAT is primarily divided into the following modules: human and object interaction proposal module (HOIPM), feature extraction module (FEM), local features module (LFM), global features module (GFM), and semantic features module (SFM). The five modules of the proposed PMGAT model are shown in Figure 3.

The operation of PMGAT is as follows: First, an input image is provided, and as the backbone network, we utilize the ResNet-50 architecture to extract the initial image features denoted as F_{org}^{im} from the image. Here, ($batch$, $w/16$, $h/16$, d) denotes the scale size, where ''batch'' refers to the batch size for each iteration, (w, h) represents the width and height of the input image, 16 indicates the stride of ResNet-50, and d represents the depth of the feature map. Next, the Human and Object Interaction Proposal Module is utilized to detect effective human-object interaction pairs, simultaneously providing the positional information L^{obj} and L^{h} for the individuals and interacting objects, as well as the features of the interacting object F_{inst}^{obj}. Subsequently, L^{obj}, L^{h} and F_{org}^{im} are passed into the Feature Extraction Module to extract facial keypoint features F_{inst}^{fp} and the location coordinate L^{fp}, as well as human body part features F_{inst}^{hp} and location coordinate L^{hp}. Subsequently, these features were separately fed into the Local Feature Module (LFM) and Global Feature Module (GFM) for the extraction of local and global features in parallel. By utilizing fully connected neural network (FCN) layers, we obtained interaction scores S^{fp} based on local facial features, interaction scores S^{hp} based on local human body part features, and interaction scores S^{p} based on global features related to human body spatial poses. The word pair <person, book> is processed in the Semantic Feature Module (SFM) to extract semantic features, and the interaction scores $S^{semantic}$ are obtained through FCN layers. Finally, the interaction scores from multiple modules are fused to calculate the interaction score, S_{inter} between the person and object.

A. MODULE 1: HUMAN AND OBJECT INTERACTION PROPOSAL MODULE (HOIPM)

In the HOI task, there are a large number of negative samples, indicating combinations of individuals and objects where no interaction occurs. This can lead to significant computational resources and time consumption when filtering for valid interacting objects. The main reason for this is the substantial quantity of negative samples, which far exceeds the number of samples where actual interactions occur [21]. This can adversely affect the precision and effectiveness of HOI detection [34]. Detecting the object with which a person effectively interacts is a formidable challenge in the field of computer vision. The primary goal of HOIPM is to provide boundary box information for individuals and interacting objects, as it plays a crucial role in detecting a limited number of valid interacting objects within the object regions in the image. The boundary box information for individuals and interacting objects is defined as L_h^i and L_{obj}^j, as shown in Equations (1) and (2), respectively.

$$L_h^i = (x_h^i, y_h^i, w_h^i, h_h^i) \tag{1}$$

$$L_{obj}^j = (x_{obj}^j, y_{obj}^j, w_{obj}^j, h_{obj}^j) \tag{2}$$

In this context, i represents the $i-$th person, j represents the $j-$th interacting object, x_h^i, and y_h^i represent the central coordinates of the person's location, w_h^i and h_h^i represent the width and height values of the person's bounding box, x_{obj}^i and y_{obj}^j represent the central coordinates of the interacting object's location, and w_{obj}^j and h_{obj}^j represent the width and height values of the interacting object bounding box.

In the HOIPM, we employed IR-GNN [48] to perform this task. The IR-GNN model employs a graph-based structure, enabling the effective estimation of interaction probabilities between humans and a multitude of objects. This helps filter out the objects that are most likely to be involved in the interaction. Additionally, IR-GNN utilizes a versatile architecture that seamlessly integrates various network configurations. This allows the detection of relevant interacting objects, which are then passed on to the next stage of the network for interaction classification. Performance evaluation and comparison were conducted using the experimental results on the HICO-DET [39] and V-COCO [49] datasets. The results indicated a substantial enhancement in the detection of interactions between humans and objects [48]. In this process, the IR-GNN identifies the positions L_h^i and h_h^i of the human and interacting object in the image, and the features F_{inst}^h and F_{inst}^{obj} of the human and interacting objects. This position information and features are subsequently fed into the next stages: the Local Feature Module (LFM) and the Global Feature Module (GFM).

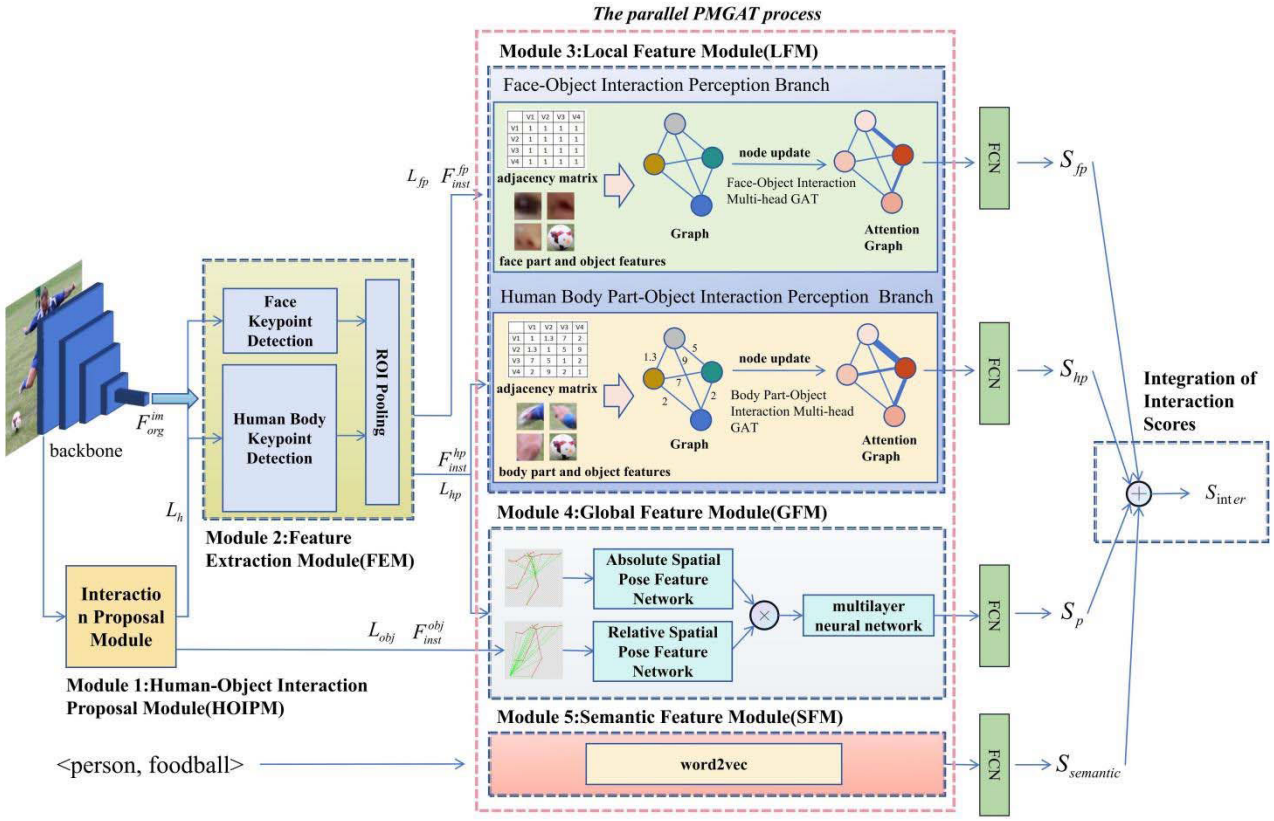


FIGURE 3. Structure of the proposed PMGAT model.

B. MODULE 2: FEATURE EXTRACTION MODULE (FEM)

To accurately extract human body part and facial keypoints, we utilized open-source pre-trained models to extract relevant keypoint information. In this study, we employed the pre-trained models openpose_body_estimation [50] and face_landmark_localization [51] to extract the human body part keypoints (18 keypoints) and facial keypoints (68 keypoints), denoted as Num_{hp} for the number of human body part keypoints and Num_{fp} for the number of facial keypoints. The sets of human body parts and facial keypoints are represented as $N_{hp}=\{n_i^{hp}\}, i\in 0,1,\cdots,17$ and $N_{fp}=\{n_j^{fp}\}, j\in 0,1,\cdots,67$, respectively. We define the set of interacting objects as $N_{obj}=\{n_k^{obj}\}, k\in N$. The positions of the human body parts and facial keypoints are denoted as L_{hp} and L_{fp}, respectively, and are defined in Equations (3) and (4).

$$L_{hp} = \{l_{hp}^i = (x_{hp}^i, y_{hp}^i)\}, i \in 0, 1, \cdots, Num_{hp} - 1 \quad (3)$$

$$L_{fp} = \{l_{fp}^j = (x_{fp}^j, y_{fp}^j)\}, j \in 0, 1, \cdots, Num_{fp} - 1 \quad (4)$$

where (x_{hp}^i, y_{hp}^i) represents the position coordinates of the $i-$th human body part keypoints, and (x_{fp}^j, y_{fp}^j) represents the position coordinates of the $j-$th facial keypoints.

Next, we computed the dimensions of both the bounding box for human body parts and the bounding box for facial features. These coefficients are denoted by α and β. Construct the bounding box around the human body part keypoints as B^{hp} and the bounding box around the facial keypoints as B^{fp}, as defined by Equations (5) and (6):

$$B_{hp} = \alpha L_h^i \quad (5)$$

$$B_{fp} = \beta L_h^i \quad (6)$$

The influence of the values of the coefficients, denoted as α and β, on the performance of the HOI detection model is explored in later sections.

Bounding boxes B^{hp} and B^{fp} are then constructed based on the center coordinates of the body part keypoints and facial keypoints, respectively. Then, the RoI pooling [52] operation is used to obtain the instance features for the body and facial parts, denoted as F_{inst}^{hp} and F_{inst}^{fp}, respectively. The size of the feature maps is $(batch, w, h, d)$, where $batch$ refers to the batch size for each iteration; w and h represent the width and height of the feature maps, respectively; and d represents the depth of the feature maps.

C. MODULE 3: LOCAL FEATURE MODULE (LFM)

To enhance the attention on the features within the region where persons and objects interact, a local feature module that incorporates facial key features and local body part features was designed. The graph is constructed by utilizing image feature data from interacting objects in conjunction with bounding boxes outlining human body parts and facial keypoints, all of which serve as nodes. The relationships

between nodes and interaction objects are considered edges, and these relationships are stored in an adjacency matrix to construct a graph model. Additionally, an attention mechanism was introduced to prioritize the local body parts involved in effective human-object interactions. To achieve this, the local feature module consisted of two parallel graph attention network branches. These two branches focus on the facial regions and human body parts, respectively. Therefore, the face–object interaction perception branch and the human body part–object interaction perception branch were designed.

1) FACE-OBJECT INTERACTION PERCEPTION BRANCH (FOI)

Facial features are effective in revealing a person's emotions and intentions [53]. In particular, there is a close relationship between the interaction behavior categories related to the face and facial keypoints. To delve deeper into this relationship, this study proposes a face-object interaction multi-head GAT. In this research, facial keypoints are first detected, and the image features within the bounding box of the facial keypoints and the object bounding box are employed as nodes in the graph neural network. Then, by connecting the facial keypoints to the object center, edges are formed in the graph. In this way, we constructed a multi-head graph attention network to describe the interactions between the face and objects.

Inspired by the literature [54], Graph Attention Networks (GAT) [55] have been effective in learning information between facial keypoints and objects. The GAT updates node representations using an attention mechanism. In each layer's computation, the GAT dynamically assigns different attention weights to nodes and their neighboring nodes based on their relationships, allowing it to focus more on important neighbor nodes. This mechanism enables the GAT to concentrate on neighbor nodes with higher relevance to a given node and effectively utilize information within the graph structure.

Given a graph $G_{fpo}=(V_{fpo},E_{fpo})$, where $V_{fpo}=V_{fp}\cup V_{obj}$ is the set of nodes representing facial keypoints and object features, E_{fpo} is the set of edges representing the relationships between nodes of image features, $v_i^{fpo}\in V_{fpo}$ is the $i-$th node, $\varepsilon_{ij}^{fpo}=(v_i^{fpo},v_j^{fpo})\in E_{fpo}$ is the edge connecting nodes v_i^{fpo} and v_j^{fpo}, and graph G_{fpo} is a complete graph. The features of v_i^{fpo} and v_j^{fpo} are denoted as f_i^{fpo} and f_j^{fpo}, where f_i^{fpo} and f_j^{fpo} are flattened feature vectors of F_{inst}^{fp} and F_{inst}^{obj} with dimensions of $m\times 1$. First, the features of nodes v_i^{fpo} and v_j^{fpo} in layer l, denoted as $f_i^{fpo(l)}$ and $f_j^{fpo(l)}$, are enhanced by a shared linear transformation, resulting in embedded vectors $z^{fpo(\ell)i}$ and $z_j^{fpo(\ell)}$, as obtained from Equation (9) and initialized with a weight matrix W. The attention coefficients $\varepsilon_{ij}^{fpo(l)}$ between nodes v_i^{fpo} and v_j^{fpo} in layer l are computed using Equation (7), as shown in Figure 4.

$$\varepsilon_{ij}^{fpo(l)} = Leaky\,\mathrm{Re}\,Lu(\varphi^{(l)^T}(z_i^{fpo(l)}||z_j^{fpo(l)})) \quad (7)$$

$$Leaky\,\mathrm{Re}\,Lu(x) = \begin{cases} x, & x > 0 \\ \lambda x, & x \leqslant 0 \end{cases} \quad (8)$$

$$z_i^{fpo(l)} = W^{(l)} f_i^{fpo(l)} \quad (9)$$

$W\in R^{d(M)\times d(m)}$, where $d(M)$ represents the dimensionality of output features. $\varphi^{(l)}(\cdot)$ is a single-layer feedforward neural network represented by a weight vector that maps concatenated high-dimensional features to a real number. T represents the transpose, and $LeakyReLu(\cdot)$ [56] in Equation (8) is the activation function, where $\|$ represents the concatenation operation and λ is equal to 0.01.

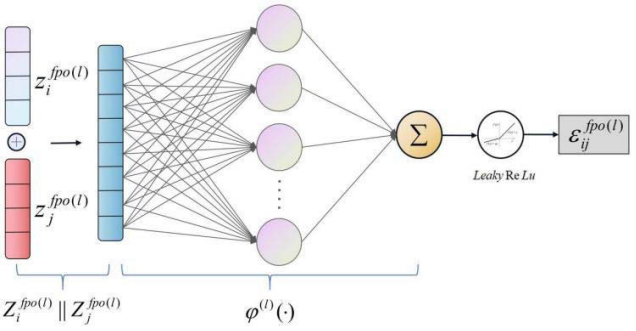


FIGURE 4. Attention coefficient.

To facilitate the comparison of attention coefficients between different entities, the attention coefficients are normalized to obtain the normalized attention coefficients in layer l, denoted as $\alpha_{ij}^{(l)}$, ensuring that the sum of the edge weights between node v_i and all neighboring nodes is equal to 1. Normalization Equation (10) is as follows: This Equation is referred to [18].

$$\begin{aligned} a_{ij}^{(l)} &= soft\,\max(\mathrm{Re}\,Lu(\varepsilon_{ij}^{fpo(l)})) \\ &= \frac{\exp(\varepsilon_{ij}^{fpo(l)})}{\sum_{k\in N(i)}\exp(\varepsilon_{ik}^{fpo(l)})} \end{aligned} \quad (10)$$

where $k\in N(i)$ represents the set of first-order neighboring nodes of node v_i^{fpo}, including the node itself in v_i^{fpo}.

The features of neighboring nodes are aggregated and scaled based on attention coefficients $a_{ij}^{(l)}$, as obtained from Equation (11), resulting in node v_i^{fpo} obtaining the new feature $f_i^{fpo(l+1)}$ by aggregating the features of neighboring nodes. Equation (11) is based on a few studies by [17].

$$f_i^{fpo(l+1)} = \delta(\sum_{j\in N(i)} a_{ij}^{(l)} z_j^{fpo(l)}) \quad (11)$$

where $\delta(\cdot)$ represents a nonlinear activation function, $a_{ij}^{(l)}$ represents the attention coefficients normalized using Equation (10), and $z_j^{fpo(l)}$ represents the embedding vector of node v_i^{fpo}.

To ensure the stability of the learning process for single-head attention and mitigate the risk of overfitting, we employed a multi-head attention mechanism. Figure 5 shows the results of the multi-head graph attention output. Let head=3 be the number of heads and the different layers

represent independent attention outputs. These independent attention outputs are summed and averaged, followed by an activation function to yield the ultimate output, as illustrated in Equation (12), which is based on [17].

$$\hat{f}_i^{fpo(l+1)} = \delta\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N(i)} a_{ij}^k W^k f_j^{fpo(l)}\right) \tag{12}$$

where K represents the number of heads in multi-head attention, a_{ij}^k represents the normalized attention coefficients obtained from Equation (10) of the $k-$th graph attention network, W^k represents the weight matrix for the linear transformation of the $k-$th graph attention network, and $\hat{f}_i^{fpo(l+1)}$ is the new feature obtained by aggregating the outputs of the multi-head graph attention.

Then, the features of node v_i^{fpo} are updated using $\hat{f}_i^{fpo(l+1)}$. Finally, we obtained the human-object interaction score based on facial local features through the FCN layer, denoted as S^{fp}.

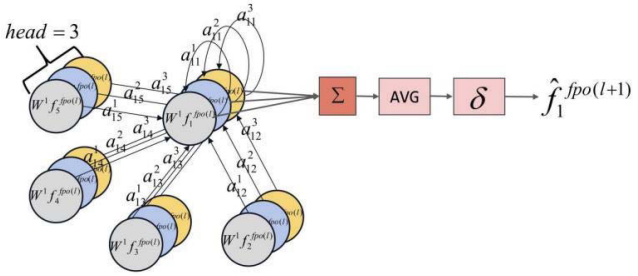


FIGURE 5. Multi-head graph attention output model.

2) HUMAN BODY PART-OBJECT INTERACTION PERCEPTION BRANCH (HBPOI)

Body parts that come into contact with objects often have a strong connection with interaction classification [57]. This branch mainly extracts human body keypoints, where the bounding boxes of the keypoints represent different body parts. The image features of the human body parts are extracted as nodes using a feature extraction model. Spatial distances were established between these features and the object features to create edges. There are two steps in constructing the body part-object interaction multi-head GAT, as described below.

(1) Spatial distance between human body parts and interacting objects

Using human body part keypoints to establish spatial distances with interacting objects, we first calculate the distances between keypoints on human body parts and then determine the distances between these keypoints and the central point of the object. Define $L_{hpo}=L_{obj}\cup L_{hp}$ as the set of coordinates representing the object's center point and human body part positions, where the $i-$th position coordinate is denoted as $l_i^{hpo}\in L_{hpo}$. The distance between the two points, represented as $D_{i,j}=d(l_i^{hpo},l_j^{hpo})$ and computed using the Euclidean formula [58], is given by Equation (13). Because images have different resolutions and the dimensional scale of distances between two points can vary significantly owing to resolution differences, z-score normalization cannot be applied to distance data because maintaining the graph's structure is crucial. Therefore, the spatial distances must be normalized. Following the method in [59], we normalized the spatial distances between each keypoint using Equations (14).

$$D_{i,j} = d(l_i^{hpo}, l_j^{hpo}) = \sqrt{\sum_{i=1}^{n}(l_i^{hpo} - l_j^{hpo})^2} \tag{13}$$

$$\hat{D}_{i,j} = Norm(D_{i,j}) = \frac{D_{i,j}-\mu}{\sigma} \tag{14}$$

where $\hat{D}_{i,j}$ represents the normalized spatial distances and l_i and l_j denote the positions of the two different keypoints. μ is the mean of the spatial distances between human body parts keypoints and objects, σ is the standard deviation of the spatial distances between human body parts keypoints and objects.

The Equations for μ and σ are as follows and are described by Equations (15) and (16):

$$\mu = \frac{\sum D_{i,j}}{Num_{hp}} \tag{15}$$

$$\sigma = \sqrt{\frac{1}{Num_{hp}}\sum (D_{i,j}-\mu)^2} \tag{16}$$

(2) Body Part–object Interaction GAT.

Inspired by [60], the close correlation between human body keypoints and objects can be effectively extracted using GAT. Consider the graph $G_{hpo}=(V_{hpo}, E_{hpo})$, where $V_{hpo}=V_{hp}\cup V_{obj}$ represents a set of nodes for human body part features and interaction object features, and E_{hpo} represents a set of edges between nodes, depicting relationships between image feature nodes. Here, $v_i^{hpo}\in V_{hpo}$ represents the $i-$th node, and $\varepsilon_{ij}^{hpo}=(v_i^{hpo},v_j^{hpo})\in E_{hpo}$ represents an edge connecting nodes v_i^{hpo} and v_j^{hpo}. The graph G_{hpo} is complete. The features of v_i^{hpo} and v_j^{hpo} are denoted as f_i^{hpo} and f_j^{hpo}, where f_i^{hpo} and f_j^{hpo} are flattened feature vectors of F_{inst}^{hp} and F_{inst}^{obj} with dimensions of $m\times 1$. First, the features of nodes v_i^{hpo} and v_j^{hpo} in layer l are denoted by $f_i^{hpo(l)}$ and $f_j^{hpo(l)}$, respectively. They are enhanced through shared linear transformations, yielding embedding vectors $z_i^{hpo(l)}$ and $z_j^{hpo(l)}$, which are initialized with the weight matrix W, as expressed by Equation (9). Subsequently, the attention coefficients $\varepsilon_{ij}^{hpo(l)}$ between nodes v_i^{hpo} and v_j^{hpo} in layer l can be obtained using Equation (7).

In the interactions between humans and objects, there is a greater likelihood of interaction between a person and nearby objects. Similarly, there is an approximate negative correlation between the distance between body parts and objects and whether they engage in interaction. Furthermore, the connection between distinct body parts and different categories of interaction behaviors varies, and their contributions in determining the interaction between humans

and objects also diverge. Therefore, assigning weights to body part features according to the spatial proximity between the body parts and objects represents a rational strategy. Its definition is given in Equations (17) and (18) below.

$$w_{dist_{ij}} = \frac{F_{dist}(i,j)}{\sum_{j=1}^{M} F_{dist}(i,j)} \tag{17}$$

$$F_{dist}(i,j) = \frac{1}{\hat{D}_{i,j}} \tag{18}$$

where w_{distij} represents the distance coefficient, which is involved in the calculation of attention coefficients for edges in the body part-object interaction multi-head GAT network. $F_{dist}(\cdot)$ represents the negative correlation between the distance between body parts and objects and whether they engage in interaction.

In the node update of the graph attention network, we incorporate the distance coefficient w_{distij} into the node update formula, and after modifying Equation (10), we obtain Equation (19). This allowed us to normalize ε_{ij}^{hpo} to obtain the attention coefficients $a_{ij}^{(l)}$.

$$\begin{aligned} a_{ij}^{(l)} &= soft\max(\mathrm{Re}\,Lu(\varepsilon_{ij}^{hpo(l)})) \\ &= \frac{\exp(\varepsilon_{ij}^{hpo(l)}) + w_{dist_{ij}}}{\sum_{k\in N(i)} \exp(\varepsilon_{ik}^{hpo(l)}) + w_{dist_{ij}}} \end{aligned} \tag{19}$$

To better control the weights of the node updates, we introduced the distance coefficient w_{distij} into the calculation formula of the attention coefficients. This distance coefficient is added to both the numerator and denominator of the attention coefficient. By considering the distance between nodes, we can adjust the attention coefficients, resulting in a greater influence on neighboring nodes and nodes with smaller distances in the node update process.

The multi-head graph attention output $\hat{f}_i^{hpo(l+1)}$ is then obtained using Equation (12). $\hat{f}_i^{hpo(l+1)}$ is used to update the features of the existing node v_i^{hpo}. Finally, the interaction score between a person and an object based on the local features of the body part is obtained through an FCN layer and is denoted as S_{hp}.

D. MODULE 4: GLOBAL FEATURE MODULE (GFM)

To comprehensively consider both the local details of human-object interactions and global context information, this study drew inspiration from Gao et al. work [19]. To enhance the performance of the model, an aggregation of features between local and global context features was executed. In this study, both the absolute spatial pose features and relative spatial pose features of the human body were utilized as global context features. By considering the spatial information between a person and an object, we could capture the relative position and pose relationships between them. Additionally, we considered the relationships between joints within the pose of the human body, as these relationships are closely associated with interaction behaviors [22].

Inspired by [22], the human body poses itself, and the relative spatial information between individual body joints and target objects provides valuable cues for detecting interactions between humans and objects. Especially in densely populated scenes, this information can significantly enhance the HOI detection performance. A relative spatial pose graph (as shown in Figure 6(a)) was used to capture the relative positions and postures between individual body joints and target objects. Through this figure, it becomes clear how the human body's posture aligns with the objects, facilitating a better comprehension of the spatial alignment between the human body and objects during interaction actions. The absolute spatial pose graph (Figure 6(b)) was used to represent the relationships between the joints within the human body. It illustrates the overall structure of the human body posture and the relative positional relationships between joints, thereby enhancing the representation of the shape and posture information of the human body.

(a)Relative spatial pose (b)Absolute spatial pose

FIGURE 6. Human spatial pose features.

1) RELATIVE SPATIAL POSE FEATURES

Spatial features play a crucial role in the inference of human interaction actions as they provide important information. For instance, when a person's bounding box is positioned above a soccer ball, it indicates a higher likelihood of interaction actions like "kick" or "inspect." To construct more refined spatial features, this study delves into the use of human key points. Specifically, based on the relative distance features between human keypoints (comprising 18 keypoints) and the center of the object, this study builds relative spatial pose features. Given that the coordinates of the $i-$th human keypoints are represented as (x_{hp}^i, y_{hp}^i) (as per Equation 3), and the center coordinates of the interacting object are denoted as (x_{obj}, y_{obj}) (as per Equation 2), the relative spatial pose feature f_{rp}^i is defined as in Equation (20) below.

$$f_{rp}^i = \left(\frac{x_{hp}^i - x_{obj}}{W}, \frac{y_{hp}^i - y_{obj}}{H} \right) \tag{20}$$

where (W, H) is the size of the image. All relative spatial pose features for the keypoints are collectively defined as $f^{rp} \in R^{18\times 2}$.

2) ABSOLUTE SPATIAL POSE FEATURES

Usually, different actions lead to variations in the human pose. For example, in the cases of <human, ride, bicycle> and <human, push, bicycle>, the pose of the human will clearly differ. However, sometimes even when interacting with different objects, a person may exhibit similar poses while performing the same action, as seen in <human, ride, horse> and <human, ride, bicycle>. These observations suggest that the internal attributes of a human pose play a role in facilitating the detection of interaction actions. Therefore, this study introduces absolute spatial pose features to leverage the relationships between keypoints within the human pose, thereby improving the effectiveness of HOI detection.

Absolute spatial pose features were obtained by normalizing the coordinates of each keypoint using the center of the human bounding box. The Equation is described in Equation (21) below.

$$f_{ap}^{i} = \left(\frac{x_{hp}^{i}}{x_h}, \frac{y_{hp}^{i}}{y_h}\right) \tag{21}$$

where (x_h, y_h) is the center coordinate of the human obtained from Equation (1). The absolute spatial pose features of all keypoints are denoted as $f_{ap} \in R^{18\times 2}$.

3) POSE FEATURE NEURAL NETWORK MODULE

Figure 7 illustrates the structure of the Pose Feature Neural Network Module proposed in this study. This module consists of four layers and includes two branches: the Relative Spatial Pose Feature Network and the Absolute Spatial Pose Feature Network. These two branches are used to map the relative and absolute spatial poses to high-dimensional features, encode features through multiple fully connected layers, and ultimately generate part-level spatial information features.

In Layer 1, the Relative Spatial Pose and Absolute Spatial Pose branches handle the input features f_{rp} and f_{ap}, respectively, through the FC-36 fully connected layer. In this context, the fully connected layer comprises 36 input nodes, generating 128 sets of feature data that are subsequently transmitted to the following fully connected layer, denoted as FC-128. Subsequently, a ReLU activation function was applied for non-linear transformation, and the features were normalized using Batch Normalization (BN). To minimize the computational parameters and prevent overfitting, Dropout was utilized to randomly drop the neurons.

After feature extraction in layer 1, a fully connected layer comprising 64 nodes was designed to perform additional feature extraction in layer 2. Subsequently, the relative spatial pose and absolute spatial pose were concatenated, and the features were integrated through two additional fully connected layers in layers 3 and 4. Finally, the interaction score, denoted as S_p, which represents the interaction between the person and object, is derived using the fully connected neural network layer (FCN layer).

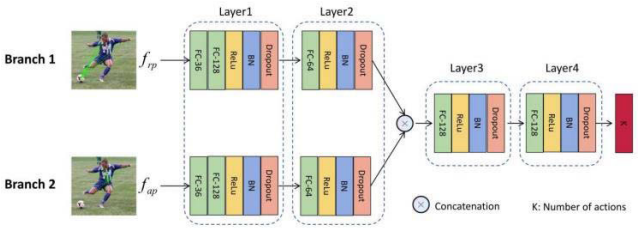


FIGURE 7. The structure of the pose feature neural network module.

E. MODULE 5: SEMANTIC FEATURE MODULE (SFM)

This module primarily focuses on capturing semantic information from textual data for the subsequent stage of human-object reasoning. In this study, a word embedding-based approach was employed to map words into continuous vector representations. These representations are then transformed into semantic representations through multiple fully connected layers.

In this study, word2vec [61] encodes features as semantic features for objects. The Word2vec model is a widely used feature representation method in the field of Natural Language Processing (NLP). It converts input text into vector representations that capture latent semantic or syntactic similarities. The key feature of this model is that a similar vocabulary is assigned similar feature representations.

Specifically, the word2vec model, which is trained on the Google News dataset consisting of approximately 100 billion words, is our foundational model. This model generates 300-dimensional vector representations covering three million words. In the case of the HICO-DET and V-COCO datasets, the object categories they encompass are aligned with the 80 categories established within the COCO dataset. To conduct the experiments, the semantic features of these categories were first saved offline using a pre-trained word2vec model. During the training, we retrieved and obtained the semantic features of the corresponding object categories based on the object detection results for each image. Finally, the human-object interaction scores $S_{semantic}$ were obtained using a fully connected neural network (FCN) layer.

By utilizing word2vec encoding features, semantic information about objects can be effectively extracted, which can then be utilized for various tasks in this study, including the construction of semantic scene graphs. This approach offers an advantage in that similar object categories will have similar feature representations, which aids in better handling the semantic relationships between them.

F. INTEGRATION OF INTERACTION SCORES AND LOSS FUNCTIONS

Through the processes mentioned above, we obtained interaction scores based on local facial features S_{fp}, interaction scores based on local body part features S_{hp}, interaction scores based on global body pose features S_p, and interaction scores based on semantic features $S_{semantic}$.

These scores were then fused to obtain the final interaction relationship score between a person and an object S_{inter}. This is defined in Equation (22). The Equation is based on a few studies by [34] and [41].

$$S_{\text{int}er} = S_{fp} + S_{hp} + S_p + S_{semantic} \tag{22}$$

During the model training phase, because HOI detection is a multi-label and multi-class classification problem, we employed binary cross-entropy loss functions $BCE(\cdot)$ for each action category. As shown in Equation (23), our training objective function uses the common loss function used in [18].

$$Loss = \frac{1}{k}\sum_{j=1}^{k} BCE(\text{s}_j, y_j^{label}) \tag{23}$$

where k represents the total number of possible actions, $s_j \in S_{inter}$ the probability of a specific action, and y_j^{label} the corresponding true label.

IV. EXPERIMENTS

A. EXPERIMENTAL DATASET

In this study, evaluations were conducted on two significant benchmarks for human-object interaction detection: V-COCO and HICO-DET. The V-COCO dataset comprised 10,633 images, encompassing 16,100 person instances and spanning 26 interaction categories. It serves as an evaluation benchmark for models in the context of a relatively small-scale human-object interaction detection task. However, the HICO-DET dataset offers a more extensive dataset, with 47,776 images that involve 80 object categories and 117 action categories, resulting in a total of 600 human-object interaction categories. The complexity and large-scale complexity of the HICO-DET dataset makes it a suitable benchmark for evaluating models for more comprehensive and intricate human-object interaction detection tasks.

In the HICO-DET dataset, there are three different sets of human-object interaction categories: Full Set: This includes all 600 human-object interaction categories and is used to evaluate the overall performance of models over the entire range of human-object interactions. Rare Set: This set comprises 138 human-object interaction categories with fewer than ten training instances. It was used to assess the performance of the models on rare interaction categories. Non-Rare Set: This set includes 462 human-object interaction categories with more than ten training instances. It is used to evaluate the performance of the models in non-rare interaction categories.

In the experiments, as the original HICO-DET and V-COCO datasets did not provide a predefined training and validation split, this study employed a common method for such a split. A random approach is used to ensure data randomness and repeatability. Specifically, 80% of the samples from the original dataset were randomly selected as the training set, whereas the remaining 20% were used as the validation set.

B. EVALUATION METRICS

According to previous studies [3], [8], [30], our evaluation metric of choice is mean Average Precision (mAP). A correct triplet prediction is defined as having an IOU greater than 0.5 between the human box and object box, along with accurate predictions of both the object category and verb category. In the case of HICO-DET, the mAP results are presented for full, rare, and non-rare settings. For V-COCO, we computed $AP^{\#1}_{role}$ and $AP^{\#2}_{role}$ to represent mAP for scenario #1 (including objects) and scenario #2 (ignoring objects) separately. This differentiation allows us to account for objects that might have been missed because of occlusion. In scenario #1, an empty object box is expected to be predicted when the occlusion aligns with the corresponding ground truth, whereas in scenario #2, object boxes are always assumed to match under such circumstances.

C. IMPLEMENTATION DETAILS

This study utilized a deep learning framework based on PyTorch. For experimental details, a parallel training approach was employed, where the four modules were trained separately, and parallelization was achieved using the DistributedDataParallel technique [62] provided by the PyTorch framework. The learning rate was set as 0.0025 during the training process for each module. As the training progressed, when the total number of iterations reached 600 K, a learning rate decay strategy was applied to adjust the learning rate to 0.0001. Learning rate decay is an effective optimization strategy that gradually reduces the learning rate during training, helping the model converge more stably towards the optimal solution in later stages. To prevent overfitting, Dropout techniques [63] were incorporated into the model at a dropout rate of 0.5. The training process for all branches was conducted using 6 NVIDIA Tesla V100 graphics cards. The setting of coefficients α=0.1 for human body part keypoints and coefficients β=0.06 for facial keypoints were also utilized.

D. DATA PROCESSING

The HICO-DET and V-COCO datasets undergo essential data preprocessing steps. Considering the diverse sizes of the original image sets and the prerequisite of model for uniformly sized images, the resolution of each image is initially standardized to 512 × 512 pixels. This standardization guarantees consistent geometric features across all dataset images. The primary aim of this adjustment is to minimize potential geometric distortions stemming from varied resolutions, thereby mitigating the risk of the wedge effect.

V. RESULT AND DISCUSSION

In this section, a series of experiments aimed at assessing the outstanding performance of our approach is presented. These experiments covered the following aspects of investigation: HOI detection accuracy, training efficiency, the impact of

bounding box sizes on body parts and facial keypoints on detection performance, the influence of the number of heads in the multi-head attention mechanism on detection performance, ablation experiments, and qualitative analysis. These results contribute to a better understanding of the potential applicability of our method.

A. COMPARISON WITH STATE-OF-THE-ART

To demonstrate the effectiveness of our approach in HOI detection, experiments were conducted on the HICO-DET and V-COCO datasets by comparing our method with current state-of-the-art HOI detection methods. These methods include the iCAN, TIN, DRG, SCG, TMHOI, and ViPLO. To ensure fair comparisons and showcase each model's optimal performance, iCAN and TIN utilize ResNet-50 as their backbone networks, DRG, SCG, and TMHOI employ ResNet-50-FPN as the backbone network, and ViPLO uses ViT-B/ 16 as the backbone network. Our approach utilizes ResNet-50 and ResNet-101 as the backbone networks. The experimental results are listed in Table 1.

Regarding the HICO-DET dataset, our findings indicate that our method performs better than all the existing methods. In both modes (full and non-rare) of data using ResNet50 as the backbone network, the mAP performance of our method was, on average, 0.6% higher than that of ViPLO. Owing to the diversity of rare categories in the rare mode of the HICO-DET dataset, with each category having a small number of instances, the dataset exhibits a long-tail distribution. Therefore, in the Rare mode, our HOI detection results are lower than the state-of-the-art ViPLO performance. ViPLO, which uses ViT as the backbone network and introduces a new feature extraction method, the MOA module, to address spatial quantization issues, achieves an advanced detection performance. In comparison, our method selects fine-grained features, combines them with graph models to thoroughly extract feature relationships between granularities, and provides a deeper understanding of the image, achieving performance similar to ViPLO, and even more significant improvements in the Full and Non-Rare modes.

To further optimize the method, the ViPLO idea was applied, and ViT-B/16 was employed as the backbone network. Next, the self-attention mechanism was used to capture the global information in the image. Therefore, it enhances the interpretability of pixels. The experimental results show that in the rare mode of the Default Situation, the method significantly improved the mAP value from 33.21% to 35.54% (2.33%). Meanwhile, for the rare mode of the Default Situation for ViTB/16, the mAP value of ViPLO compared to the proposed PMGAT improved from 35.45% to 35.54% (0.09%). Similarly, for the rare mode of Known Object Situation, the mAP value of the proposed PMGAT has improved from 37.63% to 38.86% (1.23%), while for ViTB/16, the mAP value of ViPLO compared to the proposed PMGAT, the mAP value has improved from 38.82 to 38.86 (0.04%).

TABLE 1. The mAP(%) value on the HICO-DET dataset.

Method	Backbone Network	mAP(%) value for two situations					
		Default Situation			Known Object Situation		
		Full	Rare	Non-Rare	Full	Rare	Non-Rare
iCAN[10]	ResNet-50	14.84	10.45	16.15	16.26	11.33	17.73
TIN[21]	ResNet-50	17.03	13.42	18.11	19.17	15.51	20.26
DRG[19]	ResNet-50-FPN	24.53	19.47	26.04	27.98	23.11	29.43
SCG[40]	ResNet-50-FPN	31.33	24.72	33.31	34.37	27.18	36.52
TMHOI[41]	ResNet-50-FPN	26.95	21.28	28.56	-	-	-
ViPLO[42]	ViT-B/16	37.22	35.45	37.75	40.61	38.82	41.15
The proposed PMGAT	ResNet-50	37.43	33.21	38.16	41.32	37.63	42.39
	ViT-B/16	**37.82**	**35.54**	**38.32**	**41.51**	**38.86**	**42.62**

The experimental results for the V-COCO dataset are listed in Table 2. Similarly, PMGAT was tested using ResNet-50 and ViT-B/16 as the backbone networks for comparison with ViPLO. In PMGAT, the $AP^{\#1}_{role}$ value was improved from 60.6% to 62.7%, and the $AP^{\#2}_{role}$ value was improved from 65.9% to 68.8% compared to using the ViT-B/16 backbone network and ResNet-50 backbone network, which were 2.1% and 2.9%, respectively. PMGAT achieves optimal performance with the ViT-B/16 backbone, improving $AP^{\#1}_{role}$ values from 62.2% to 62.7% and $AP^{\#2}_{role}$ values from 68.0% to 68.8% compared with the existing state-of-the-art ViPLO by 0.5% and 0.8%, respectively. It should be noted that PMGAT also achieves state-of-the-art performance and outperforms existing state-of-the-art approaches when using the same backbone network support as ViPLO.

TABLE 2. The mAP(%) value on the V-COCO dataset.

Method	Backbone Network	mAP(%) value for two scenarios	
		$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
iCAN[10]	ResNet-50	45.3	52.4
TIN[21]	ResNet-50	47.8	54.2
DRG[19]	ResNet-50-FPN	51.0	-
SCG[40]	ResNet-50-FPN	54.2	60.9
TMHOI[41]	ResNet-50-FPN	-	-
ViPLO[42]	ViT-B/16	62.2	68.0
The proposed PMGAT	ResNet-50	60.6	65.9
	ViT-B/16	**62.7**	**68.8**

For the four different backbone networks, the performance of the proposed PMGAT was tested on the HICO-DET dataset, and the results are presented in Table 3. The test results indicate that as the depth of the ResNet backbone increases from 18 to 34, and then to 50 and 101, there is a significant overall performance improvement for both situations. From ResNets 18 to 34, there is a significant improvement in the mAP value, while from RestNet to 50-101, there are small changes. For example, in the full mode of a default situation, from ResNet 18 to 34, the improvement in mAP value is large, from 19.75% to 27.58% (7.83%), while from ResNet 50 to 101, the improvement is small, from 37.43% to 37.46% (0.03%). Because extracting features is important as input in the interaction classifier, HOI detection performance can vary significantly based on the performance differences of the backbone network. As can be seen in

Tables 1 and 2, PMGAT, which uses the ViT-B/16 backbone network, achieved a significant enhancement compared to the use of ResNet-50.

TABLE 3. Model detection results with different networks backbone.

Method	Backbone Network	mAP(%) value for two situations					
		Default situation			Known Object situation		
		Full	Rare	Non-Rare	Full	Rare	Non-Rare
The proposed PMGAT	ResNet-18	19.75	14.06	21.96	21.38	16.13	23.43
	ResNet-34	27.58	21.36	29.16	28.89	23.77	30.51
	ResNet-50	37.43	33.21	38.16	41.32	37.63	42.39
	ResNet-101	37.46	33.21	38.18	41.37	37.63	42.43

B. COMPARISON OF TRAINING MODEL CONVERGENCE TIME WITH EXISTING METHODS

Owing to the parallel structure design of PMGAT, where each module is trained in parallel using multi-threading and multi-GPU, the model training convergence speed is significantly better than that of existing methods of the same kind. To ensure fairness in testing, the models were trained on both the HICO-DET and V-COCO datasets, utilizing Nvidia Tesla V100 GPUs (6 GPUs in total), a batch size of 50, an AdamW optimizer [64], a weight decay of 0.0001, and 10 epochs. Figure 8 shows the time until model convergence for PMGAT, iCAN, TIN, DGR, SCG, TMHOI, and ViPLO trained on the HICO-DET (red) and V-COCO (cyan) datasets. As can be seen from the figure, PMGAT requires only 30 and 18 min to train the model on both datasets. Obviously, our PMGAT method outperforms the other methods in terms of training time owing to the model's parallel structure and parallel training approach.

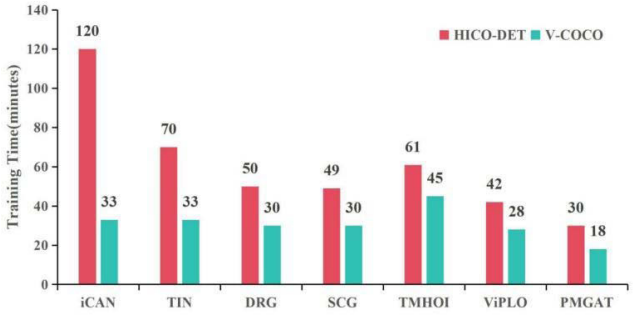


FIGURE 8. Illustrates the training times of different models on the HICO-DET and V-COCO datasets.

C. THE HOI DETECTION IMPACT OF VARIOUS BOUNDING BOX SIZES

To compare the impact of using different bounding box sizes for feature extraction on body part keypoints and facial keypoints, the coefficient sets values as follows, based on Equations (5) and (6):

$$\alpha \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}$$
$$\beta \in \{0.035, 0.04, 0.045, 0.05, 0.055, 0.06, 0.065, 0.07, 0.075, 0.08\}$$

Because determining the optimal values for coefficients α and β in the bounding box formulas for body part keypoints and facial keypoints requires experimentation, this study conducted multiple experiments in the same experimental environment to find the best coefficients.

Figure 9 shows the experimental results of the proposed method using a heat map. The values in the heat map represent the experimental results mAP of PMGAT for HOI detection using different coefficients on the V-COCO dataset, and the x- and y-axes indicate the values of coefficients α and β, respectively. From a Visualization perspective, in the heat map, the colors of each cell fade from red to blue, with redder representing higher mAP and bluer representing lower mAP.

In scenario #2, the mAP obtained the highest 0.68 when using α=0.1 and β=0.06, indicating that α=0.1 and β=0.06 are the most suitable numerical ratios for body part keypoints and facial part keypoints bounding box sizes.

The coefficients α and β starting from 0.1 and 0.06, respectively, show an overall decreasing trend in detection performance, whether increased or decreased. The worst case occurred at α=0.5 and β=0.08 and mAP=0.07. It can be observed that overly large bounding boxes for body parts keypoints and facial keypoints may extract redundant features from the surrounding areas and background, whereas bounding boxes that are too small may not capture enough features, leading to model learning in the wrong direction. This not only does not improve the detection performance, but also adds too many redundant parameters or results in parameter scarcity, thus reducing the model's performance.

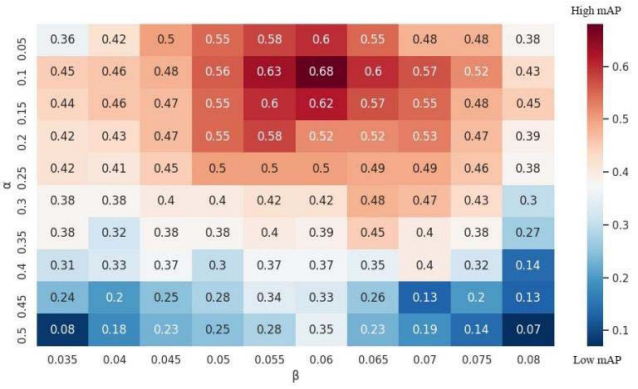


FIGURE 9. The heat map of results for the coefficients α and β on the HOI detection performance mAP of our method on the V-COCO dataset.

D. THE IMPACT OF THE NUMBER OF HEADS IN THE MULTI-HEAD ATTENTION MECHANISM

A series of experiments was conducted to evaluate the impact of the number of heads (NoH) in the graph attention network on attention effectiveness. The experimental results showed that the number of heads had a significant influence on the attention effectiveness. In a single-head attention model, attention is primarily focused on each position, leading to uneven weight allocation. However, increasing the number of

heads can enhance the model's expressive power, resulting in more reasonable attention weight allocation.

Specifically, three sets of experiments are designed using different numbers of attention heads (head=1, head=2, head=3). Figure 10 visualizes the attention weight distribution generated on body parts. In Figure 10, the relationships between the body parts form a graph model. A matrix is used to store the graph structure, where the x- and y-axes represent the body parts, and the matrix cells represent the weights of the relationships between the body parts. Because this matrix is symmetric, we only show the lower triangular matrix.

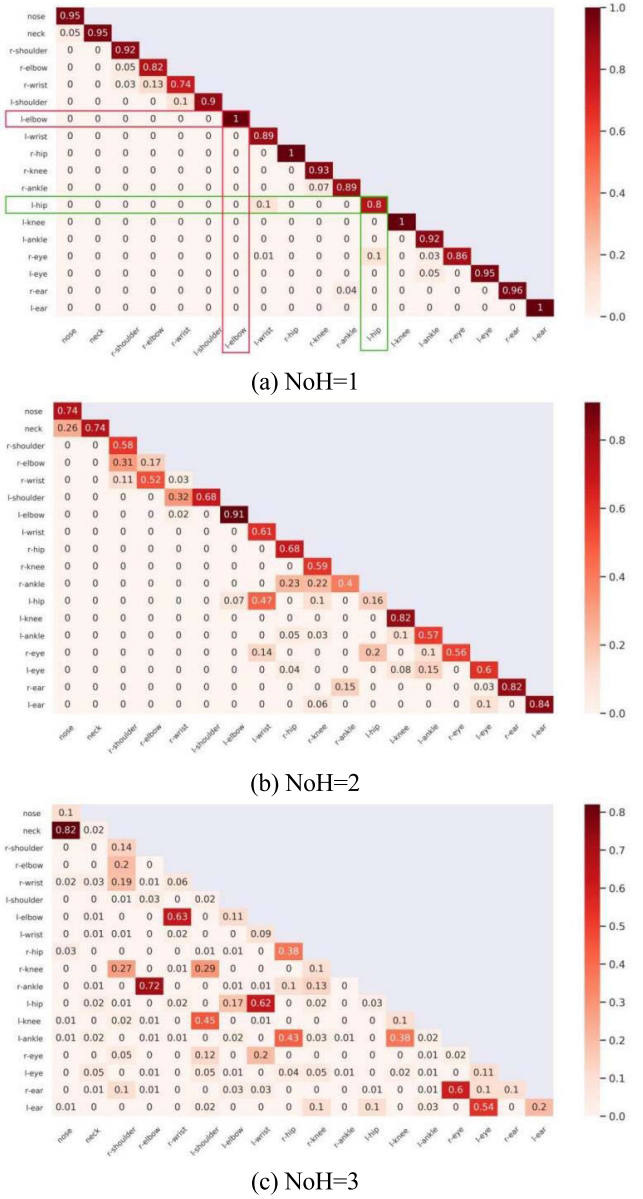
For example, in Figure 10(a), the cell value of the intersection of the r-shoulder on the x-axis and r-wrist on the y-axis is 0.03, which indicates that there is a certain connection between the two body parts when inferring interaction behavior. The larger the cell value, the greater the possibility of proving a connection between body parts and the redder the color of the cell. If the cell value is smaller, it is less likely that there is a connection between body parts, and the color is lighter red.

From the experimental results, when head = 1, the value on the diagonal is larger; for example, for l-elbow on the x-axis and l-elbow on the y-axis, in the cell surrounded by the red box in Figure 10(a), only the diagonal cell has a value, and the color is the reddest; the other cells do not have a value assigned to them, meaning that the l-elbow is not associated with any other body part. However, on the x-axis of the l-hip and y-axis of the l-hip, the cell area is enclosed using a green box where (x=l-hip,y=l-wrist)=0.1, indicating that there is some connection between the l-hip and l-wrist. Because the assignment of the weight values is normalized, the sum of the values of the cells in the region enclosed by the green box is equal to 1. However, as shown in Figure 10(a), most of the values are concentrated on the diagonal, which means that the final attentional weight matrix is mainly focused on the self-localization of each position.

In the comparison of Figure 10 (a) and (b), it can be seen that the original (x=r-elbow, y=r-elbow) = 0.82 is reduced to 0.17, which assigns the weight values to r-shoulder and r-wrist, with values of 0.31 and 0.52, respectively. The other parts start to be assigned weight values among themselves, but the distribution is not well distributed. It can be concluded that if head = 2, another attention weight matrix with a slightly more reasonable weight distribution is obtained.

As shown in Figure 10(c), when head = 3, the weight values, which were initially concentrated on the diagonal, were reasonably distributed to other positions.

Thus, it can be concluded that as the number of heads increases, the attention weight allocation gradually becomes more reasonable and is no longer overly focused on self-locations. This indicates that a multi-head attention network helps the model better understand the interaction relationships in the image, thereby improving the accuracy and uniformity of attention effectiveness. This is crucial for human-object interaction detection tasks.



(a) NoH=1

(b) NoH=2

(c) NoH=3

FIGURE 10. Visualized attention weight distribution on body parts for different NoH.

E. ABLATION EXPERIMENT

To assess the influence of each component in the module within the PMGAT, ablation experiments were conducted by comparing various variants of the PMGAT structure on both the V-COCO in scenario #1($AP^{\#1}_{role}$) and HICO-DET datasets in the default situation of full mode(Default/Full). The results of the HOI detection performance for various PMGAT variants are presented in Table 4.

The base model is a basic model that does not employ any neural network structure that uses the backbone and semantic features derived from Module 5 without any further neural network for feature extraction. These backbone and semantic features were directly passed to the FCN for the classification output. Only the variant structure of the base model was

named as m0 (Base). Next, different modules were gradually added to the base model, variant structures were made, and changes in HOI performance were observed.

First, HBPOI was added to m0 to compose m1 (Base+HBPOI), which significantly improved the overall performance; the HOI detection performance improved from 47.8% and 17.93% to 57.1% and 27.64% on the V-COCO ($AP^{\#1}_{role}$) and HICO-DET (Default/Full) datasets, respectively. This improvement was attributed to the detailed features of the body parts, which provided important supplementary information for the interaction detection.

Then, the FOI is added to m1 to compose m2 (Base+HBPOI+FOI), which further enhances the performance of the HOI detection model with the assistance of facial features, resulting in an improvement of 3.2% and 5.05% in the m2 HOI detection performance over that of m1.

Next, the GFM is added to m2 to compose m3 (Base+HBPOI+FOI+GFM), which uses both relative and absolute human spatial pose information to strengthen the recognition of human poses and interacting objects based on spatial structures and human poses. This resulted in 1.6% and 3.88% performance improvements in the m3 HOI detection performance over m2.

Finally, by incorporating the HOIPM to obtain m4 (Base + HBPOI + FOI + GFM + HOIPM), the final model PMGAT is formed, and the inclusion of this module leads to performance improvements in all metrics, achieving the highest performance.

TABLE 4. Performance of various modules in PMGAT.

	Method	V-COCO	HICO-DET (Default)
		$AP^{\#1}_{role}$	Full
m0	Base	47.8	17.93
m1	Base+HBPOI	57.1	27.64
m2	Base+HBPOI +FOI	60.3	32.69
m3	Base+HBPOI +FOI+GFM	61.9	36.57
m4	Base+HBPOI +FOI+GFM+HOIPM	**62.7**	**37.82**

F. QUALITATIVE RESULTS

This section presents the results from two different methods, namely SCG and heatmaps. In the SCG method, the results are compared with those of the proposed PMGAT, while the heatmap results show the visualization result of the heatmap with attention.

1) THE HOI DETECTION EFFECT OF SCG AND PMGAT

It is difficult to recognize and compare the HOI recognition performance of the proposed method with that of the SCG method. As shown in Figure 11, these visual examples illustrate the effectiveness of our model in improving the handling of misclassified cases. Our method considers local details in more detail and combines them with context and semantic features for inference. It can consider global features while also paying attention to local details, resulting in more accurate HOI category detection.

For example, in Figure 11(a), the red box represents a human, whereas the green box represents an object. The intersection of the red and green boxes indicates HOI. Previous work used SCG to detect HOI, and the result was a hold-bicycle, but in the proposed method, it detected the image as a ride-bicycle. The correct interaction of the image was a bicycle ride. Therefore, the proposed method detects the correct interactions. Similarly, in Figure 11(b-f), the correct interaction of the image has been detected by the proposed method, whereas the previous SCG pethood incorrectly detected the interaction.

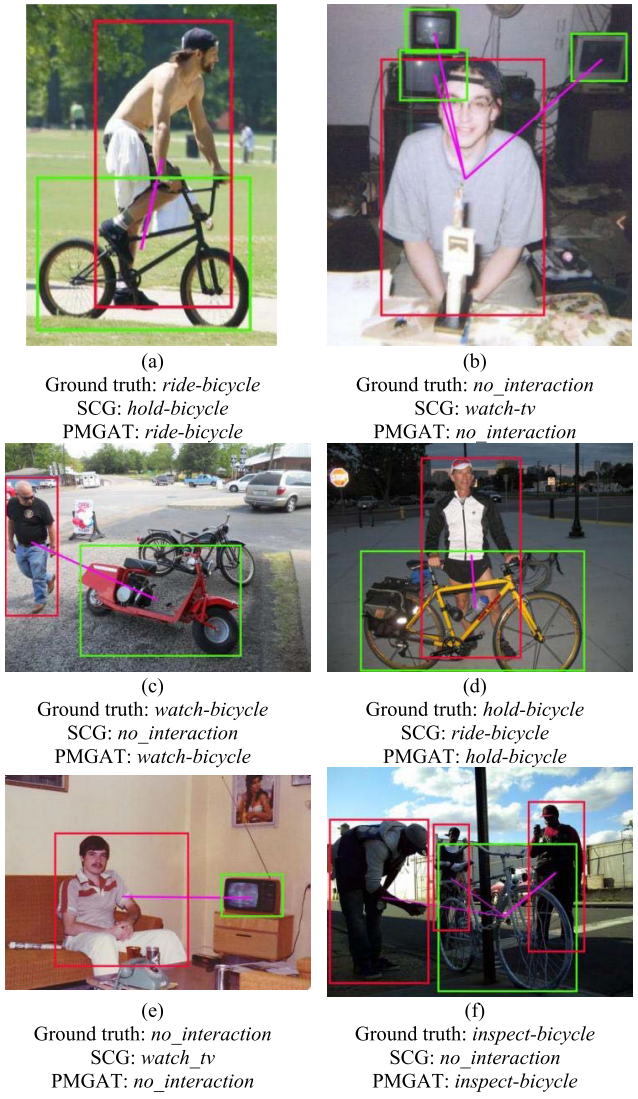


FIGURE 11. Demonstrates the HOI detection performance of PMGAT on misclassified samples.

2) HEATMAP EFFECT OF ATTENTION

Figure 12 shows the attention maps of the proposed method. The attention map for facial parts helps eliminate ambiguity in predicting the actions of the subject. The pose attention map for body keypoints highlights rich information about body parts.

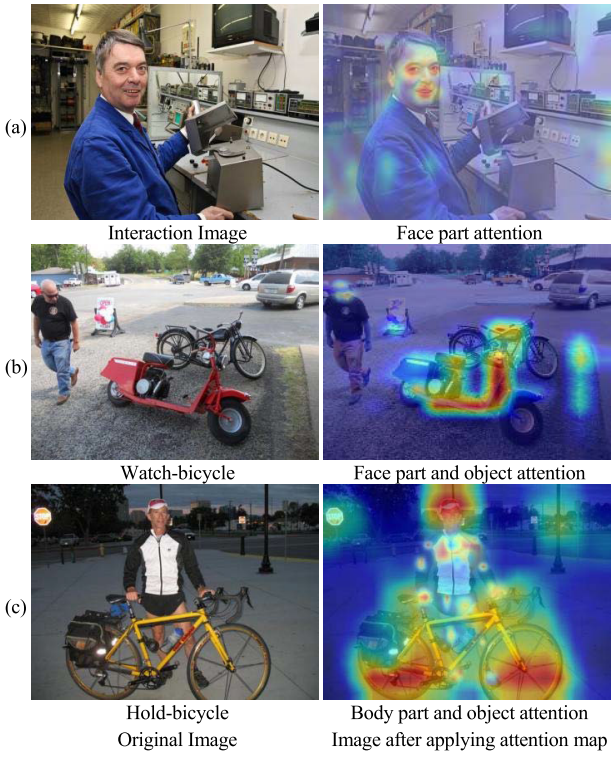


FIGURE 12. *Visualization of the image before and after the attention map.*

For example, in the first row of images, our method focuses on local facial features and captures information regarding key facial parts. In the second row of images, a pedestrian walking on the street looks at a bicycle on the roadside. Our method focuses on the local facial features of a pedestrian and accurately identifies the bicycle being looked at. In the third row of images, it is clear that the proposed method captures the body keypoints of the person and emphasizes the body parts of the person interacting with the object. Because our method incorporates a local feature module specifically designed to perceive local facial and body part features, the visualized attention heatmaps demonstrate that our method pays more attention to the local parts of the interaction between people and objects, effectively utilizing local information to improve the overall performance of the HOI detection in the model.

VI. CONCLUSION AND FUTURE WORK

This study introduces a novel model for human-object interaction (HOI) detection by combining a parallel multi-head graph attention network. The model utilizes facial and body keypoints features along with the relationships between interacting objects to construct a graph-based model, extracting the interaction correlations between keypoints. It then combines spatial pose and semantic features to explain the interactions between humans and objects. A parallel multi-head graph attention network was introduced with detailed design aspects. This research aims to address the challenge of existing methods in overlooking the local details of human-object interactions, while mitigating the high computational cost associated with training existing methods. The model demonstrated impressive results on the V-COCO and HICO-DET public datasets.

In our comparative experiments with state-of-the-art methods, as shown in Tables 1 and 2, our proposed method achieved a performance on par with the state-of-the-art ViPLO on the HICO-DET dataset. In fact, our method outperformed ViPLO in both the Full and Non-Rare scenarios. Different backbone networks affect the HOI detection performance of the PMGAT. As shown in Table 3, the choice of backbone network significantly affects the HOI detection performance, as the extracted features serve as input to the interaction classifier. The parallel structure of the model and parallel training method contributed to the accelerated training efficiency of the model. As shown in Figure 8, our approach significantly reduces the training time compared with the other methods.

The size of the bounding boxes for the body parts and facial keypoints has a significant impact on the detection performance in terms of selecting keypoints feature extraction areas. As shown in Figure 9, it is evident that only when the feature areas with the appropriate bounding box sizes for keypoints are selected can the model achieve optimal detection performance. In a single-head attention model, when encoding information at the current position, attention tends to be overly concentrated at its own position. However, as the number of attention heads in the network increased, the overall expressive power of the model improved, leading to a more reasonable allocation of attention weights across the network. The branches in our approach, including the HBPOI, FOI, GFM, HOIPM, all played a positive role in improving the performance of the model. The results in Table 4 clearly indicate that the introduction of these branches led to significant performance improvements in the ablation experiments conducted on the HICO-DET and V-COCO datasets.

In conclusion, a parallel multi-head graph attention network is proposed for detecting interactions between humans and objects, which is excellent in capturing fine-grained human-object interactions and improves the accuracy of HOI detection. Another significant advantage is the use of a multi-branch parallel structure, which greatly speeds up the training process. Through a series of experiments and comparisons, the proposed PMGAT exhibited superior performance, which is validated on the V-COCO and HICO-DET datasets. However, despite the significant performance advantages achieved by the PMGAT, there are still some limitations. Particularly, the complex structure of the network model leads to a large amount of computation. Future work will focus on addressing these limitations to enhance our approach further.

REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 1–43, 2011.

[2] G. Guo and A. Lai, "A survey on still image based human action recognition," *Pattern Recognit.*, vol. 47, no. 10, pp. 3343–3361, Oct. 2014.

[3] X. Ji and H. Liu, "Advances in view-invariant human motion analysis: A review," *IEEE Trans. Syst., Man, Cybern., C, Appl. Rev.*, vol. 40, no. 1, pp. 13–24, Jan. 2010.

[4] Y. Zhu, X. Li, C. Liu, M. Zolfaghari, Y. Xiong, C. Wu, Z. Zhang, J. Tighe, R. Manmatha, and M. Li, "A comprehensive study of deep video action recognition," 2020, *arXiv:2012.06567*.

[5] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, Nov. 2019.

[6] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and datasets," *Comput. Vis. Image Understand.*, vol. 163, pp. 21–40, Oct. 2017.

[7] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human–object interactions," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 381–389.

[8] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human–object interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8359–8367.

[9] A. Kolesnikov, A. Kuznetsova, C. Lampert, and V. Ferrari, "Detecting visual relationships using box attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1749–1753.

[10] C. Gao, Y. Zou, and J.-B. Huang, "ICAN: Instance-centric attention network for human–object interaction detection," 2018, *arXiv:1808.10437*.

[11] G. Chéron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN features for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3218–3226.

[12] A. Mallya and S. Lazebnik, "Learning models for actions and person-object interactions with transfer to question answering," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 414–428.

[13] H. Wang, W.-S. Zheng, and L. Yingbiao, "Contextual heterogeneous graph network for human–object interaction detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 248–264.

[14] H. Liu, T.-J. Mu, and X. Huang, "Detecting human–object interaction with multi-level pairwise feature network," *Comput. Vis. Media*, vol. 7, no. 2, pp. 229–239, Jun. 2021.

[15] D. Teney, L. Liu, and A. Van Den Hengel, "Graph-structured representations for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3233–3241.

[16] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human–object interactions by graph parsing neural networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 401–417.

[17] L.-M. Xia and W. Wu, "Graph-based method for human-object interactions detection," *J. Central South Univ.*, vol. 28, no. 1, pp. 205–218, Jan. 2021.

[18] Z. Liang, J. Liu, Y. Guan, and J. Rojas, "Visual-semantic graph attention networks for human–object interaction detection," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2021, pp. 1441–1447.

[19] C. Gao, J. Xu, Y. Zou, and J.-B. Huang, "DRG: Dual relation graph for human–object interaction detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 696–712.

[20] H.-S. Fang, J. Cao, Y.-W. Tai, and C. Lu, "Pairwise body-part attention for recognizing human–object interactions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 51–67.

[21] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, and C. Lu, "Transferable interactiveness knowledge for human–object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3580–3589.

[22] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9468–9477.

[23] Z. Liang, J. Liu, Y. Guan, and J. Rojas, "Pose-based modular network for human–object interaction detection," 2020, *arXiv:2008.02042*.

[24] X. Sun, X. Hu, T. Ren, and G. Wu, "Human object interaction detection via multi-level conditioned network," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2020, pp. 26–34.

[25] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10312–10321.

[26] O. Ulutan, A. S. M. Iftekhar, and B. S. Manjunath, "VSGNet: Spatial attention network for detecting human object interactions using graph convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13614–13623.

[27] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proc. ICML*, 2017, pp. 1263–1272.

[28] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-RNN: Deep learning on spatio-temporal graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5308–5317.

[29] R. Li, M. Tapaswi, R. Liao, J. Jia, R. Urtasun, and S. Fidler, "Situation recognition with graph neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4183–4192.

[30] K. Marino, R. Salakhutdinov, and A. Gupta, "The more you know: Using knowledge graphs for image classification," 2016, *arXiv:1612.04844*.

[31] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3097–3106.

[32] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with graph LSTM," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 125–143.

[33] Y. Yuan, X. Liang, X. Wang, D.-Y. Yeung, and A. Gupta, "Temporal dynamic graph LSTM for action-driven video object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1819–1828.

[34] W. Yang, G. Chen, Z. Zhao, F. Su, and H. Meng, "ICGPN: Interaction-centric graph parsing network for human-object interaction detection," *Neurocomputing*, vol. 502, pp. 98–109, Sep. 2022.

[35] S. P. R. Sunkesula, R. Dabral, and G. Ramakrishnan, "LIGHTEN: Learning interactions with graph and hierarchical temporal networks for HOI in videos," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 691–699.

[36] Q. Li, X. Xie, J. Zhang, and G. Shi, "Few-shot human–object interaction video recognition with transformers," *Neural Netw.*, vol. 163, pp. 1–9, Jun. 2023.

[37] A. Agarwal, R. Dabral, A. Jain, and G. Ramakrishnan, "Skew-robust human–object interactions in videos," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 5087–5096.

[38] Z. Ni, E. Valls Mascaró, H. Ahn, and D. Lee, "Human–object interaction prediction in videos through gaze following," *Comput. Vis. Image Understand.*, vol. 233, Aug. 2023, Art. no. 103741.

[39] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "HICO: A benchmark for recognizing human-object interactions in images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1017–1025.

[40] F. Z. Zhang, D. Campbell, and S. Gould, "Spatially conditioned graphs for detecting human–object interactions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13299–13307.

[41] L. Zhu, Q. Lan, A. Velasquez, H. Song, A. Kamal, Q. Tian, and S. Niu, "TMHOI: Translational model for human-object interaction detection," 2023, *arXiv:2303.04253*.

[42] J. Park, J.-W. Park, and J.-S. Lee, "ViPLO: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 17152–17162.

[43] P. Zhou and M. Chi, "Relation parsing neural network for human–object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 843–851.

[44] Z. Su, Y. Wang, Q. Xie, and R. Yu, "Pose graph parsing network for human-object interaction detection," *Neurocomputing*, vol. 476, pp. 53–62, Mar. 2022.

[45] Y. Y. Ghadi, M. Waheed, M. Gochoo, S. A. Alsuhibany, S. A. Chelloug, A. Jalal, and J. Park, "A graph-based approach to recognizing complex human object interactions in sequential data," *Appl. Sci.*, vol. 12, no. 10, p. 5196, May 2022.

[46] M. Tamura, H. Ohashi, and T. Yoshinaga, "QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10405–10414.

[47] S. Chan, W. Wang, Z. Shao, and C. Bai, "SGPT: The secondary path guides the primary path in transformers for HOI detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 7583–7590.

[48] J. Zhang, Z. Mohd Yunos, and H. Haron, "Interactivity recognition graph neural network (IR-GNN) model for improving human–object interaction detection," *Electronics*, vol. 12, no. 2, p. 470, Jan. 2023.

[49] S. Gupta and J. Malik, "Visual semantic role labeling," 2015, *arXiv:1505.04474*.
[50] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.
[51] U. Watchareeruetai, B. Sommana, S. Jain, P. Noinongyao, A. Ganguly, A. Samacoits, S. W. F. Earp, and N. Sritrakool, "LOTR: Face landmark localization using localization transformer," *IEEE Access*, vol. 10, pp. 16530–16543, 2022.
[52] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[53] S. Liu, S. Huang, W. Fu, and J. C.-W. Lin, "A descriptive human visual cognitive strategy using graph neural network for facial expression recognition," *Int. J. Mach. Learn. Cybern.*, vol. 2022, pp. 1–17, Oct. 2022.
[54] Q. Bao, B. Gang, W. Yang, J. Zhou, and Q. Liao, "Attention-driven graph neural network for deep face super-resolution," *IEEE Trans. Image Process.*, vol. 31, pp. 6455–6470, 2022.
[55] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.
[56] J. Xu, Z. Li, B. Du, M. Zhang, and J. Liu, "Reluplex made more practical: Leaky ReLU," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jul. 2020, pp. 1–7.
[57] Q. Li, X. Xie, W. Liu, X. Jin, and C. Zhang, "Two-stage body-part attention network for detecting human–object interactions," in *Proc. IEEE 5th Int. Conf. Signal Image Process. (ICSIP)*, Oct. 2020, pp. 415–419.
[58] P.-E. Danielsson, "Euclidean distance mapping," *Comput. Graph. Image Process.*, vol. 14, no. 3, pp. 227–248, Nov. 1980.
[59] Y. Chen, X. Tang, X. Qi, C.-G. Li, and R. Xiao, "Learning graph normalization for graph neural networks," *Neurocomputing*, vol. 493, pp. 613–625, Jul. 2022.
[60] J. Liu, J. Rojas, Y. Li, Z. Liang, Y. Guan, N. Xi, and H. Zhu, "A graph attention spatio-temporal convolutional network for 3D human pose estimation in video," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 3374–3380.
[61] K. W. Church, "Word2Vec," *Nat. Lang. Eng.*, vol. 23, no. 1, pp. 155–162, 2017.
[62] S. Li, Y. Zhao, R. Varma, O. Salpekar, P. Noordhuis, T. Li, A. Paszke, J. Smith, B. Vaughan, P. Damania, and S. Chintala, "PyTorch distributed: Experiences on accelerating data parallel training," 2020, *arXiv:2006.15704*.
[63] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 56, pp. 1929–1958, 2014.
[64] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

JIALI ZHANG received the B.S. degree in software engineering from the University of Electronic Science and Technology of China, Zhongshan Institute, in 2010, and the M.S. degree in computer science from the Wuhan University of Technology, in 2017. He is currently pursuing the Ph.D. degree with UTM. His current research interests include deep learning, image processing, and human–object interaction detection.

ZURIAHATI MOHD YUNOS received the Diploma degree in computer science from UiTM, in 1999, and the bachelor's, master's, and Ph.D. degrees in computer science from UTM, in 2001, 2006, and 2017, respectively. She is currently a Senior Lecturer with the Faculty of Computing, UTM. Her current research interests include forecasting, optimization, classification, and soft computing.

HABIBOLLAH HARON (Senior Member, IEEE) received the Diploma and bachelor's degrees in computer science from UTM, in 1987 and 1989, respectively, the M.Sc. degree in computer technology in manufacture from the University of Sussex, East Sussex, U.K., in 1995, and the Ph.D. degree in computer-aided geometric design from UTM, in 2004. He is currently a Professor of soft computing techniques with the Faculty of Computing, UTM. His current research interest includes soft computing (SC) techniques for prediction, optimization, and planning.