TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 01/01 pp1–16
DOI: 10.26599/TST.2025.9010025
Volume x, Number x, xxxx xxxx

Research Article

Exploring Hierarchical Tuple-Based Contextual Correlations for Human-Object Interaction Detection

Xin Hu, Ke Qin*, Tao He, and Guangchun Luo

Abstract: Human-Object Interaction (HOI) detection is a challenging task in computer vision, particularly in complex scenes involving multiple humans and interactions. In this paper, we propose the Hierarchical Tuple-based Contextual Correlations Learning (HTCCL) model, which aims to enhance HOI detection by systematically capturing multi-level contextual relationships. Our approach decomposes an interaction into three hierarchical levels: entity, action, and event. We introduce a heterogeneous graph network with a multi-branch Transformer architecture, where human and object entities are treated as distinct nodes, facilitating fine-grained relational reasoning. Furthermore, we leverage Contrastive Language-Image Pre-training model to embed interaction cues into queries, which are subsequently refined through local and global contextual aggregation modules. The proposed model effectively integrates contextual information across various levels, improving its ability to detect complex interactions within diverse scenes. Our extensive evaluations on standard benchmarks demonstrate the superiority of HTCCL in achieving state-of-the-art performance in HOI detection, particularly in scenarios with high relational complexity.

Key words: Human-Object Interaction (HOI) detection; multimodal models; Transformer

1 Introduction

The Human-Object Interaction (HOI) detection task involves identifying the positions of humans and objects in images and inferring their interactions[1, 2]. Understanding visual relationships in human activities is crucial for various computer vision applications, driving significant research interest in recent years[3–15].

With significant advances in object detection[16, 17] and Transformer networks[18, 19], many HOI detection methods have adopted Transformer as the dominant base architecture[20–24]. However, most existing Transformer-based HOI detection architectures are not well-suited for multitask learning[22, 25–27]. Generally, Transformer-based models employed a single decoder responsible for all sub-tasks, including human detection, object detection, and interaction classification[21, 23, 28]. This design limited the model's ability to adapt to multitask learning[25]. To address this limitation, subsequent methods[24–27, 29] introduced a dual-branch structure, employing two or more independent Transformer decoder branches. Despite partially decomposing the subtasks, the existing structures still face several issues.

Firstly, existing methods often suffer from insufficient contextual information exchange between branches[25, 26, 29–31]. Multi-branch Transformer architectures typically limit the flow of contextual information across branches, thereby hindering the model's ability to comprehend the global scene. Although some studies have introduced additional

• Xin Hu, Ke Qin, Tao He, and Guangchun Luo are with Ubiquitous Intelligence and Trusted Services Key Laboratory of Sichuan Province, University of Electronic Science and Technology of China, Chengdu 611731, China. E-mail: xh1m22@std.uestc.edu.cn; qinke@uestc.edu.cn; tao.he01@hotmail.com; gcluo@uestc.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2024-09-12; revised: 2025-01-06; accepted: 2025-02-28

context propagation mechanisms, these are typically confined to a limited scope, primarily focusing on human-object context and failing to capture deep inter-branch dependencies in complex scenes[20, 32].

Secondly, existing methods lack effective modeling of contextual interactions at different granular levels[22, 27, 33]. HOIs encompass multi-level information, ranging from individual entities to entire events. However, current Transformer architectures often overlook these hierarchical contextual interactions, resulting in inadequate discrimination and exploring fine-grained relationships.

To address these challenges, we propose the Hierarchical Tuple-based Contextual Correlations Learning (HTCCL) model, which incorporates an external graph structure to enhance interaction modeling and effectively capture hierarchical interaction within multi-decoder branch architectures. As shown in Fig. 1, we decompose the relationships in HOI tasks into three levels of interaction: entity level (e.g., "bat" and "human"), behavior level (e.g., "human-hold"), and entity level (e.g., "human-hold-bat"). These levels are interconnected, with their interactions providing crucial context for accurately interpreting the roles and relationships within the event. To explore relationships between entities and actions more deeply, our approach begins by utilizing Contrastive Language-Image Pre-training (CLIP)[34] to

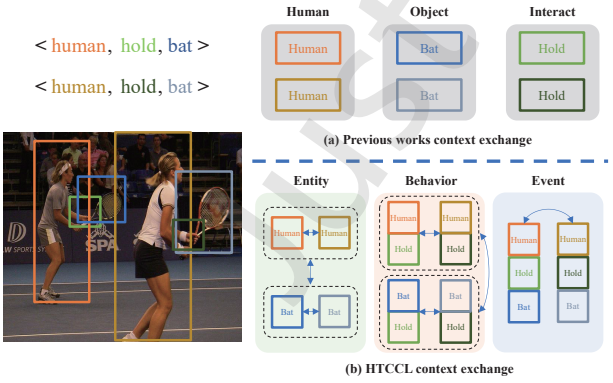

Fig. 1 Illustration of different types of contextual information interactions. (a) Previous works primarily focus on context exchange within the human-object-interaction framework without leveraging these hierarchical distinctions. (b) However, we define three hierarchical levels of contextual information: entity level (human, object), behavior (human-action, object-action), and event (human-object-action). Our method facilitates fine-grained hierarchical context exploitation across these levels, enabling more comprehensive relational reasoning.

embed interaction cues into the interaction queries, providing a rich semantic foundation. We then introduce a heterogeneous graph network into the Transformer structure. We distinctly represent human and object entities as heterogeneous nodes, allowing messages passing between homogeneous and heterogeneous nodes to convey fine-grained relationships at different levels to the three branches. Additionally, the event-level multiplexed context information, composed of the contexts from the three branches, is selectively integrated into each branch.

To sum up, our contributions are as follows:

- We propose the HTCCL model that decomposes HOI tasks into multiple contextual levels, enabling more precise interaction modeling.
- We design an architecture that combines a graph attention network with heterogeneous nodes and a three-branch Transformer to learn intra-class entity relations, and inter-class behavior relations, and propagate multiple-level relations to each branch.
- We evaluate our method on two public benchmarks, V-COCO[35] and HICO-DET[36], and provide comprehensive ablation studies to verify the effectiveness of our model.

2 Related Work

2.1 Two-stage methods

Two-stage HOI detection methods[36–42] have been extensively studied. Typically, in the first stage, these methods use off-the-shelf object detectors (e.g., faster Region-based Convolutional Neural Network (R-CNN)[43]) to detect humans and objects, and then classify the interactions between the detected human-object pairs in the second stage. In the second stage, various methods have been proposed, incorporating different types of information, including the appearance and visual features of humans and objects[44, 45], spatial features such as the bounding boxes of human-object pairs[36], and semantic embeddings of human and object labels[41, 42]. To enrich HOI features, some methods use message-passing mechanisms within instance-centric graph structures[46–48] to perform relation reasoning, leveraging the global context information between human and object instances. Additionally, some methods enhance HOI feature representation by integrating pose information[49], body part details[50], or even detailed three-dimensional body shapes[51].

Furthermore, some works utilize external knowledge from the language domain to further assist in HOI feature learning[52, 53]. In our work, we also embed semantic cues into interaction queries using CLIP[34], allowing the model to leverage these cues more effectively throughout the HOI detection process.

2.2 One-stage methods

One-stage HOI detection methods aim to simplify the detection process by directly predicting HOI triplets in an end-to-end manner[20, 22–24, 28, 54, 55]. These methods are typically faster and more straightforward compared to their two-stage counterparts. Inspired by the success of Transformer-based object detectors like DETR[16], recent one-stage methods have adopted Transformer architectures to push the limits of HOI detection. For example, UnionDet[54] was one of the first attempts to detect human-object pairs' joint regions directly in a one-stage manner. Other methods reformulate the HOI detection task as a keypoint detection problem to enable a one-stage solution[55]. Leveraging Transformer architectures, some methods use a single interaction Transformer decoder to predict a set of HOI triplets[23, 28], optimizing the entire framework end-to-end with a Hungarian loss. Meanwhile, other approaches design parallel Transformer decoders to separately detect interactions and instances, and then associate the outputs to produce the final HOI predictions[22, 24]. These methods demonstrate the potential of Transformer architectures in simplifying the HOI detection pipeline and achieving competitive performance. While these methods simplify the HOI detection pipeline, they often overlook the intricate contextual interactions between entities. In contrast, our approach integrates a three-branch architecture, enabling comprehensive contextual reasoning across human, object, and interaction branches, which further enhances the detection performance in complex scenes.

3 Method

In this section, we propose the HTCCL network to explore structural relationships among humans, objects, and interactions. The network consists of a three-branch transformer, a multimodal knowledge transfer module, and contextual aggregation modules. The model architecture is shown in Fig. 2.

3.1 Model architecture

Consistent with prior research[16, 23, 28], given the raw images $X \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote the image height, width, and number of color channels, we first process it through a ResNet[56] backbone CNN to extract low-resolution visual feature maps $\boldsymbol{V}_c \in \mathbb{R}^{H\times W\times C}$. The feature maps are reduced from $C = 2048$ to $D = 256$ (where D denotes the transformer hidden dimension) using convolution. Positional encodings $p \in \mathbb{R}^{H\times W\times D}$ are then applied to capture spatial information[16]. The resulting features, which are flattened and combined with positional encodings, are input into a transformer encoder to extract serialized visual features $\boldsymbol{V}_e \in \mathbb{R}^{H\times W\times D}$ as follows:

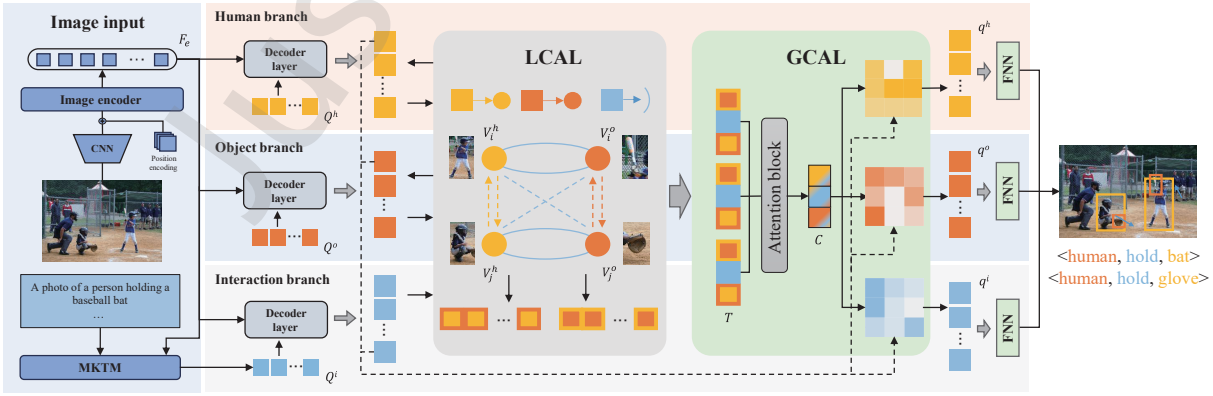

Fig. 2 Overview of HTCCL model. The model comprises three main branches: human branch, object branch, and interaction branch. The input image is first processed through a CNN-based image encoder, with interaction queries initialized by the Multimodal Knowledge Integration Module (MKIM) using CLIP. These queries are then fed into corresponding decoder layers of each branch. The Local Contextual Aggregation Learning (LCAL) module leverages a heterogeneous graph network to model fine-grained relationships between human and object entities at multiple contextual levels, while the Global Contextual Aggregation Learning (GCAL) module propagates event-level context across branches. The outputs from the last layer of each branch are used to predict HOI triplets. FNN: Feedforward Neural Network.

$$\boldsymbol{V}_e = \text{Encoder}^{\text{image}}(\boldsymbol{V}_c,\ p) \tag{1}$$

where $\text{Encoder}^{\text{image}}(\cdot)$ denotes the transformer encoder.

Before sending $\boldsymbol{V}_e$ to the decoder, we first send it to MKIM. Here, we combine text features extracted by CLIP[34] with visual features from the image encoder to generate cue features F_c. These cue features refine the interaction queries Q^i, as shown in the following formula from the randomly initialized queries Q_r^i:

$$Q^i = \text{MKIM}(F_e,\ Q_r^i) \tag{2}$$

Next, we utilize a three-branch decoder framework, comprising the human, object, and interaction decoders. Each branch θ contains l layers, where $\theta \in h$, o, and i represents the human, object, and interaction branches, respectively. At each layer, the queries Q^θ are updated through a transformer decoder layer, followed by the LCAL and GCAL to address the limitations in modeling multi-level contextual interactions.

In the LCAL module, after each decoder layer, the queries from each branch are treated as nodes in a heterogeneous graph, where human and object queries are considered different node types. The queries are updated through message passing within the graph, refining their representations based on intra-class entity and inter-class behavior interactions. These refined queries are then passed to the GCAL module, which aggregates global contextual information across different events. The output queries q^h, q^o, and q^i from the final decoder layer are then used to make the final predictions for HOI triplets.

3.2 MKIM

The MKIM aims to exploit multimodal knowledge in CLIP to embed semantic cues from events and actions into interaction queries. Once these cues are combined with image features, the consolidated knowledge is integrated into the interaction queries, enhancing their contextual relevance and depth. The architecture of MKIM is shown in Fig. 3.

Following Ref. [57], we pass the encoded features through a three-layer Feed-Forward Network (FFN) to predict the HOI candidate labels, selecting the top K HOI labels with the highest confidence scores. Subsequently, we convert the selected HOI labels into textual descriptions using CLIP, employing a standard template such as "a photo of a person [verb-ing] a/an [object]" for each HOI category. Non-interactive verbs are described as "a photo of a person with no interaction with a/an [object]". These sentences are then processed by CLIP's text encoder to produce global semantic representations $\boldsymbol{S}_t \in \mathbb{R}^D$ for each event.

We then use these semantic embeddings to present interaction information. For positive human-object pairs with a single interaction, the corresponding semantic feature $\boldsymbol{S}_i$ is directly integrated into the interaction queries at the initialization process. For positive pairs with multiple interactions, the average of the semantic features $\overline{\boldsymbol{S}}$ is calculated to mitigate semantic ambiguity. For a pair with N interactions, the representation is given by

$$\overline{\boldsymbol{S}}_{\text{pos}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{S}_i \tag{3}$$

For negative pairs, we use the average semantic feature $\overline{\boldsymbol{S}}_{\text{neg}}$ of all M embeddings to differentiate interaction-related features from non-interaction clusters,

$$\overline{\boldsymbol{S}}_{\text{neg}} = \frac{1}{M}\sum_{j=1}^{M}\boldsymbol{S}_j \tag{4}$$

Once the semantic features are obtained, we aggregate them with the image features extracted from the visual backbone. The aggregation combines both the semantic and visual cues into a unified feature representation as follows:

$$F_e = \text{softmax}\left(\frac{\boldsymbol{V}_e \cdot \overline{\boldsymbol{S}}}{\sqrt{D}}\right)\cdot\left[\boldsymbol{V}_e,\ \overline{\boldsymbol{S}}\right] \tag{5}$$

where $[\boldsymbol{V}_e,\ \overline{\boldsymbol{S}}]$ represents the concatenation of the visual and semantic features, and "$\cdot$" denotes matrix multiplication. This dot-product attention mechanism computes the relevance between visual features $\boldsymbol{V}_e$ and semantic embeddings $\overline{\boldsymbol{S}}$, capturing the most contextually relevant interactions for enhancing the representation of interaction queries.

Before incorporating the aggregated features F_e into the interaction queries, we first apply self-attention to the randomly initialized interaction queries. Those queries are then integrated with the aggregated semantic-visual features through a cross-attention mechanism. In this cross-attention process, F_e serves as both the queries and keys, ensuring that the interaction knowledge is enriched with meaningful contextual information from both the image and text domains.

3.3 LCAL

Human activity scenes are typically complex, involving multiple entities. Context learning is essential for inferring relationships between humans and objects. Previous studies[58] focused mainly on behavior and event-level features, overlooking the inherent heterogeneity of unary entities within HOI triplets. To address this gap, we introduce a heterogeneous graph to facilitate both intra-class entity and inter-class behavior level interaction. This design effectively captures the similarities among homogeneous entities in HOI-related contexts while simultaneously modeling the interactivity between heterogeneous entities representing subjects and objects.

In our design, the queries from the human branch are treated as human nodes V^h, while the object branch queries serve as object nodes V^o. Interaction queries Q^i are utilized as edge values $\mathcal{E}$ to represent the relationships between these entities. Each object query is aligned with the corresponding human query in Q^h and interaction queries Q^i.

Intra-class entity learning: In our heterogeneous graph, intra-entity context is defined as the interaction similarity among homogeneous nodes. For example, if a group of people are interacting with identical or similar objects, such as rowing boats, they are likely executing the same or related actions. To node V_i^h, the entity message aggregation method can be represented as follows:

$$M_{V_i^h}^{\text{entity}} = \frac{1}{\left|N_{V_i^h}^{\text{entity}}\right|} \sum_{V_j^h \in N_{V_i^h}^{\text{entity}}} \beta_{ij} \cdot g(V_j^h) \tag{6}$$

where β_{ij} is the attention weight computed based on the similarity between node V_i^h and node V_j^h, and $g(\cdot)$ is a function realized by a neural network layer with an activation function, representing the transformation of the feature V_j. $N_{V_i}^{\text{entity}}$ denotes the set of homogeneous neighbors of node V_i, and $|N_{V_i}^{\text{entity}}|$ is the number of such neighbors.

To compute β_{ij}, we first define a pointer $\vec{p}$ for each node that encapsulates its interaction information. This vector is computed by considering the features of the node and its interactions with neighboring nodes,

$$p_i = \max\left\{ g\left(V_i^h + V_l^o + \mathcal{E}_{il}\right),\ \forall V_l^o \in N_{V_i^h}^{\text{behavior}} \right\} \tag{7}$$

where $\mathcal{E}_{il}$ denotes the edge feature associated with the human-object pair (V_i^h, V_l^o).

For nodes V_i^h and V_j^h, we first compute the cosine similarity of $\vec{p}_i$ and $\vec{p}_j$ to represent the intra-entity context between them,

$$\epsilon_{ij} = \frac{\vec{p}_i \cdot \vec{p}_j}{\|\vec{p}_i\| \cdot \|\vec{p}_j\|} \tag{8}$$

The attention weight in the process of gathering homogeneous nodes' messages is their normalization after the softmax function,

$$\beta_{ij} = \frac{\exp(\epsilon_{ij})}{\sum_{k=1}^{K} \exp(\epsilon_{ik})} \tag{9}$$

where k indexes the neighboring entity nodes of V_i.

The computation of object nodes' message passing follows a similar approach, and the graph attention mechanism is applied in the same way. By learning with the intra-class context, the nodes will be able to identify which homogeneous nodes are relevant and gather more messages from them, while the irrelevant nodes will be filtered out during the reasoning process.

Inter-class behavior learning: To calculate inter-class behavior context, we also introduce an interaction weight γ. The interaction weight (γ_{il}) between person node V_i^h and object V_j^o is computed using a fully connected layer with Rectified Linear Unit (ReLU),

$$\gamma_{il} = g(V_i^h + V_l^o + \mathcal{E}_{il}),\ \gamma_{il} \in \mathbb{R} \tag{10}$$

The inter-class behavior message ($M_{V_i^h}^{\text{behavior}}$) received by the human node V_i^h is aggregated from its neighboring object nodes V_l^o, conditioned on the corresponding relation configuration $\mathcal{E}_{V_i^h V_l^o}$. Using a graph attention mechanism, the message aggregation for node V_i^h can be rewritten as

$$M_{V_i^h}^{\text{behavior}} = \sum_{V_l^o \in N_{V_i}^{\text{behavior}}} \delta_{ik} \cdot g(\mathcal{E}_{il} \oplus V_l^o) \tag{11}$$

where $\oplus$ means concatenation, and δ_{ik} is the normalized attention weight computed based on γ_{ik}. The attention weight δ_{ik} is computed using the softmax function,

$$\delta_{ik} = \frac{\exp(\gamma_{il})}{\sum_{k=1}^{K} \exp(\gamma_{ik})} \tag{12}$$

Since inter-class context is expected to imply whether the entities have interaction or not, the attention mechanism can help the inter-class message focus on those interactive heterogeneous nodes and neglect many non-interactive nodes. The inter-class

learning strengthens the relations between heterogeneous nodes which are interactive.

After the message propagates, the updated nodes can be obtained by

$$V' = \sigma\left(V + M_v^{\text{intra}} + M_v^{\text{inter}}\right) \tag{13}$$

These updated nodes are then used to update the human queries Q^h and object queries Q^o. Specifically, the values of the partial human queries and object queries are replaced by the corresponding updated nodes from the heterogeneous graph. This ensures that the information gathered during the message-passing phase is integrated into the original query sets, enhancing their representational capabilities.

3.4 GCAL

While the LCAL captures the unary intra-entity class and binary inter-behavior class contexts, the GCAL aims to mine clues of ternary events and capture implicit dependencies across different events. In the GCAL module, we utilize a combination of self-attention mechanisms and channel attention techniques to aggregate information across different events. This allows the model to effectively learn the inter-event connections and dependencies, ensuring that the contextual information from one event can inform and enhance the understanding of another event. The aggregated global context is then integrated into the human, object, and interaction branches, further refining the predictions of HOI triplets.

The GCAL module concatenates the updated human queries Q^h, object queries Q^o, and interaction queries Q^i to construct the event triplet context, which is then processed through a MultiLayer Perceptron (MLP) to form the event-level context T,

$$T = \text{MLP}(Q^h \oplus Q^o \oplus Q^i) \tag{14}$$

Next, we use cross-attention mechanisms to integrate this event-level context with the visual features F, creating a refined event context. The cross-attention mechanism allows the model to attend to relevant parts of the image that contribute to the understanding of the event,

$$C = \text{CrossAttn}(T, F) \tag{15}$$

The event relationship context C is then propagated to each branch (human, object, and interaction) using channel attention mechanisms. The channel attention mechanism selects the most relevant contextual information for each specific task, ensuring that the refined context is appropriately tailored to the needs of each branch. Formally, the channel attention and the refined tokens Q^θ for the l-th layer of branch are formulated as follows:

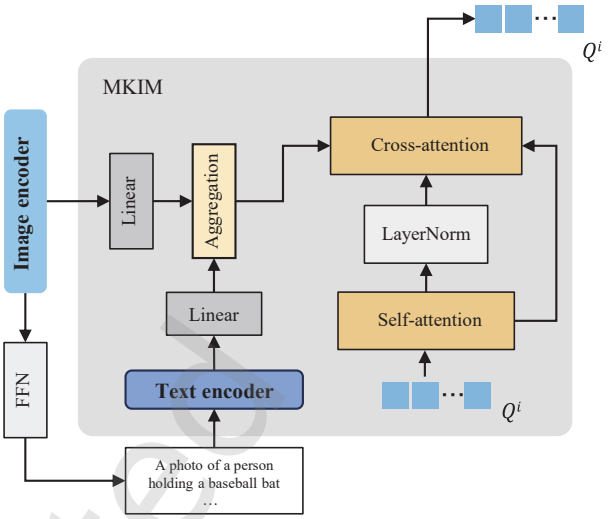

Fig. 3 Illustration of the MKIM. The module utilizes a text encoder, initialized with CLIP, to process textual descriptions of HOIs. These text features are aggregated with visual features extracted by the image encoder. The aggregated features undergo a self-attention mechanism followed by Layer Normalization (LayerNorm). Finally, a cross-attention mechanism refines the interaction queries Q^i, embedding rich semantic cues from both the text and visual domains, which are then passed to subsequent stages in the model.

$$\alpha = \text{Sigmoid}(\text{MLP}([C \oplus Q^\theta])) \tag{16}$$

$$q^{\theta'} = Q^\theta + \alpha \odot \text{MLP}([C \oplus Q^\theta]) \tag{17}$$

where Sigmoid$(\cdot)$ represents logistic function, and $\odot$ is the Hadamard product. By incorporating the global context into each branch, the GCAL module enhances the model's ability to recognize and reason about complex interactions within the scene, leading to more accurate HOI predictions. After the final l-th layer in each branch, the refined outputs $q^{\theta'}$ are fed into distinct FFNs specific to each task. The human branch predicts human bounding boxes, the object branch predicts object bounding boxes and object categories, and the interaction branch predicts the interaction type. These predictions are combined to form the final HOI triplet.

3.5 Training objective

Similar to prior transformer-based approaches[25, 28, 32], to train our proposed method, we utilize the Hungarian matching algorithm[59] to match the predicted HOI

instances with the ground-truth labels. Our model, HTCCL, is optimized using a multi-task loss function that combines four different types of losses: bounding box L1 loss $\mathcal{L}_{\text{box}}$[60], Generalized Intersection over Union (GIoU) loss $\mathcal{L}_{\text{GIoU}}$[61] for bounding box regression, entity classification cross-entropy loss $\mathcal{L}_{\text{ent}}$, and focal loss $\mathcal{L}_{\text{int}}$[62] for interaction classification. The overall training loss $\mathcal{L}_{\text{total}}$ is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{box}} + \lambda_{\text{GIoU}}\mathcal{L}_{\text{GIoU}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}} + \lambda_{\text{int}}\mathcal{L}_{\text{int}} \tag{18}$$

where λ_{GIoU}, λ_{ent}, and λ_{int} are hyper-parameters that balance the contributions of each individual loss component. In addition to this, we improve the representation learning by leveraging intermediate loss which is computed the same as L to adjust at the time of inference. Specifically, different FFNs are added to the decoders with their parameters shared to calculate auxiliary loss.

4 Experiment

4.1 Datasets

For evaluating the performance of our proposed method, we conduct experiments on two widely-used HOI datasets: V-COCO[35] and HICO-DET[36].

V-COCO: The V-COCO dataset[35] is a subset of the Microsoft COCO[63] dataset, specifically designed for HOI detection tasks. It contains 5400 images for training, 2533 images for validation, and 4946 images for testing. The dataset includes annotations for 80 object categories and 29 action categories, resulting in a total of 2533 unique HOI instances. Each instance in the dataset is annotated with bounding boxes for both humans and objects, as well as labels indicating the type of interaction between them.

HICO-DET: The HICO-DET dataset[36] is one of the largest datasets available for HOI detection, containing 47 776 images for training and 9658 images for testing. It covers 600 HOI categories, which are combinations of 80 object categories and 117 action categories. The 600 HOI categories are divided into two groups: 138 for rare and 462 for non-rare, according to the number of instances they contain. Each image in the dataset is annotated with bounding boxes for humans and objects, along with detailed interaction labels.

4.2 Evaluation metrics

To quantitatively evaluate the performance of our proposed method on HOI detection, we use the mean Average Precision (mAP) metric. In the context of HOI detection, an interaction is considered a true positive if both the human and object bounding boxes have an Intersection over Union (IoU) greater than 0.5 with the ground truth boxes, and the interaction label matches the ground truth label. For V-COCO, we report the results in two distinct scenarios: $\text{AP}^{\text{S1}}_{\text{role}}$ and $\text{AP}^{\text{S2}}_{\text{role}}$. The $\text{AP}^{\text{S1}}_{\text{role}}$ metric evaluates the model's ability to predict all HOIs present in the image, regardless of whether the object is linked to the annotated person or not, providing a comprehensive assessment of the model's capacity to detect all possible interactions. In contrast, $\text{AP}^{\text{S2}}_{\text{role}}$ focuses specifically on interactions associated with the annotated person in the image, emphasizing the model's precision in localizing and recognizing interactions for specific individuals.

As for HICO-DET, the evaluations cover three diverse sets of categories: the entire spectrum of HOI categories (full), rare HOI categories in training (rare), and the other HOI categories outside the rare ones (non-rare).

4.3 Implementation details

In our proposed model, the encoder leverages ResNet-50[56] as the CNN backbone, followed by a six-layer transformer encoder. We configure the number of branch layers L to 6. For the training process, we set the number of queries N to 64 for the HICO-DET dataset and 100 for the V-COCO dataset, following the setup in Ref. [25]. The loss weight λ_{L1} is assigned values of 2.5, while the weights for λ_{GIoU}, λ_{cls}, and λ_{int} are all set to 1. The initial parameters of our network are derived from DETR[16], which is pretrained on the MS-COCO dataset[63]. Additionally, we utilize the pretrained ViT-16/B CLIP model to generate text embeddings for initializing query vectors while keeping its parameters fixed. For optimization, we adopt the AdamW optimizer[64] with a weight decay of 1×10^{-4}. The initial learning rates are set to 1×10^{-5} for the CNN backbone and 1×10^{-4} for other components. The learning rates are dropped at the sixty epoch. All experiments are conducted with a batch size of 16 on eight NVIDIA A100 GPUs.

4.4 Performance comparisons

Tables 1 and 2 present the performance comparison between HTCCL and other state-of-the-art HOI

Table 1 Results on the HICO-DET dataset under two evaluation settings. Default computes mAP over the full test set, while known object computes mAP on the subset of test images that contain the corresponding object category. Results are reported on full/rare/non-rare splits. The best results are marked in bold and the second best are underscored. FPN represents Feature Pyramid Network. – indicates that it was not reported in the original paper.

(%)

Method		Source	Backbone	Default			Known object		
				Full	Rare	Non-rare	Full	Rare	Non-rare
Two stage method	iCAN[44]	BMVC'18	ResNet-50	14.84	10.45	16.15	16.26	11.33	17.73
	TIN[38]	PAMI'19	ResNet-50	17.03	13.42	18.11	19.17	15.51	20.26
	VSGNet[65]	CVPR'20	ResNet-152	19.80	16.05	20.91	–	–	–
	VCL[66]	ECCV'20	ResNet-50	23.63	17.21	25.55	25.98	19.12	28.03
	IDN[67]	NeurIPs'20	ResNet-50	26.29	22.61	27.39	28.24	24.47	29.37
	DRG[48]	ECCV'20	ResNet-50-FPN	24.53	19.47	26.04	27.98	23.11	29.43
	SCG[46]	ICCV'21	ResNet-50-FPN	31.33	24.72	33.31	34.37	27.18	36.52
	UPT[26]	CVPR'22	ResNet-50	31.66	25.94	33.36	35.05	29.27	36.77
	CQL[29]	CVPR'23	ResNet-50	31.72	27.40	33.01	34.25	28.77	35.89
	STIP[42]	CVPR'22	ResNet-50	32.22	28.15	33.43	35.29	31.43	36.45
	Opencat[68]	CVPR'23	ResNet-101	32.68	28.42	33.75	–	–	–
One stage method	UnionDet[54]	ECCV'20	ResNet-50-FPN	17.58	11.72	19.33	19.76	14.68	21.27
	IP-Net[69]	CVPR'20	ResNet-50-FPN	19.56	12.79	21.58	22.05	15.77	23.92
	HOTR[24]	CVPR'21	ResNet-50	25.10	17.34	27.42	–	–	–
	HOITrans[23]	CVPR'21	ResNet-101	26.61	19.15	28.84	29.13	20.98	31.57
	AS-Net[20]	CVPR'21	ResNet-50	28.87	24.25	30.25	31.74	27.07	33.14
	QPIC[24]	CVPR'21	ResNet-101	29.90	23.92	31.69	32.38	26.06	34.27
	MSTR[21]	CVPR'22	ResNet-50	31.17	25.31	32.92	34.02	28.83	35.57
	DisTR[32]	CVPR'22	ResNet-50	31.75	27.45	33.03	34.50	30.13	35.81
	CDN[25]	NeurIPs'21	ResNet-101	32.07	27.19	33.53	34.79	29.48	36.38
	MUREN[54]	CVPR'23	ResNet-50	32.87	28.67	34.12	35.52	30.88	36.91
	ERNet-M[31]	TIP'23	EfficientNet-V2	32.94	27.86	34.45	–	–	–
	HOICLIP[30]	CVPR'23	ResNet-50	34.69	**31.12**	35.74	33.75	**34.47**	38.54
	LOGICHOI[33]	NeurIPs'24	ResNet-50	34.53	**31.12**	35.38	**37.04**	<u>34.31</u>	37.86
Our method	HTCCL	–	ResNet-50	<u>34.78</u>±0.04	30.06±0.07	<u>35.86</u>±0.12	35.91±0.10	32.32±0.09	<u>39.88</u>±0.11
		–	ResNet-101	**35.47**±0.15	<u>30.61</u>±0.11	**36.74**±0.08	<u>36.81</u>±0.13	33.06±0.05	**40.77**±0.09

methods on the V-COCO and HICO-DET datasets. As shown in Table 2, the proposed HTCCL method consistently achieves state-of-the-art performance across various settings on the V-COCO dataset, outperforming both existing two-stage and one-stage methods. The three-branch transformer HTCCL model surpasses previous CNN-based models that use graph structures for contextual interaction[46–48, 65], as well as transformer-based methods[21, 26, 28, 32]. This demonstrates the efficacy of allocating different subtasks to separate branches. Notably, our method also exceeds the performance of current parallel multi-branch approaches. These existing methods primarily focus on behavior-level information interaction, propagating contextual information from the instance branch to the interaction branch[20, 24, 32]. In contrast, HTCCL facilitates the interaction of local entity and behavior information, as well as global event information, among its three branches. It selectively propagates the necessary contextual information to each subtask through GCAL. Further evaluation on the HICO-DET dataset revealed similar results to those observed on the V-COCO dataset, as shown in Table 1. HTCCL achieves state-of-the-art performance compared to existing methods on this dataset as well. These results collectively indicate that the three-branch structure and the hierarchical approach using external graphs for fine-grained contextual relationship reasoning provide more discriminative features for addressing each subtask.

Table 2 Results on the V-COCO dataset across two scenarios. "R50" and "R101" refer to ResNet-50 and ResNet-101 backbones, respectively, while "HG104" stands for Hourglass-104, and "ENet2" refers to EfficientNet-V2. The best results are marked in bold, and the second best are underscored. – indicates that it was not reported in the original paper.

Method		Conference	Backbone	AP^{S1}_{role} (%)	AP^{S2}_{role} (%)
Two stage method	iCAN[44]	BMVC'18	R50	45.3	52.4
	TIN[38]	PAMI'19	R50	47.8	54.2
	VSGNet[65]	CVPR'20	R152	51.8	57.0
	VCL[66]	ECCV'20	R101	48.3	–
	IDN[67]	NeurIPs'20	R50	53.3	60.3
	DRG[48]	ECCV'20	R50	51.0	–
	SCG[46]	ICCV'21	R50	54.2	60.9
	UPT[26]	CVPR'22	R50	59.0	64.5
	OpenCat[68]	CVPR'23	R101	61.9	63.2
	STIP[42]	CVPR'22	R50	66.0	70.7
	CQL[29]	CVPR'23	R101	66.8	69.8
One stage method	UnionDet[54]	ECCV'20	R50	47.5	56.2
	IP-Net[69]	CVPR'20	HG104	51.0	–
	GGNet[70]	CVPR'21	HG104	54.7	–
	HOITrans[23]	CVPR'21	R101	52.9	–
	AS-Net[20]	CVPR'21	R50	53.9	–
	HOTR[24]	CVPR'21	R50	55.2	64.4
	QPIC[28]	CVPR'21	R50	58.8	61.0
	MSTR[21]	CVPR'22	R50	62.0	65.2
	CDN[25]	NeurIPs'21	R101	63.9	65.9
	DisTR[32]	CVPR'22	R50	66.2	68.5
	ERNet-M[31]	TIP'23	ENet2	58.2	–
	HOICLIP[30]	CVPR'23	R50	63.5	64.8
	LOGICHOI[33]	NeurIPs'24	R50	64.4	65.6
	MUREN[54]	CVPR'23	R50	68.8	71.0
Our method	HTCCL	–	R50	69.1±0.11	71.5±0.09
		–	R101	**70.1**±0.06	**72.3**±0.13

4.5 Complexity analysis

As shown in Table 3, we present a comparison of the model complexity, measured in terms of Floating Point Operations Per Second (FLOPS) and the number of parameters, between our proposed HTCCL models and other leading Transformer-based HOI detection methods. Following DisTR, FLOPS are computed as the average over the first 100 images in the V-COCO test set using the flop count operators from Detectron2[71]. We can observe from Table 3 that our HTCCL model outperforms the state-of-the-art multi-branch model MUREN, with a comparable increase in computational resources—only 2.7% more FLOPS and slightly higher parameter count. The HTCCL-S variant, derived by reducing the number of layers from 6 to 4 in the architecture, maintains strong performance while using fewer parameters and FLOPS than MUREN. These results demonstrate that even the smaller HTCCL-S model effectively balances performance and computational efficiency.

Table 3 Comparison of model complexity across various Transformer-based HOI detection methods. * indicates multi-branch transformer methods. We report FLOPS and the number of parameters, alongside performance metrics (AP^{S1}_{role} and AP^{S2}_{role}) on the V-COCO dataset. – indicates that it was not reported in the original paper.

Method	Backbone	AP^{S1}_{role} (%)	AP^{S2}_{role} (%)	Number of parameters (mega)	FLOPS (giga)
QPIC[28]	ResNet-50	58.5	61.0	41.68	87.87
AS-NET*[20]	ResNet-50	53.9	–	52.75	88.86
HOITrans[23]	ResNet-101	52.9	–	60.62	156.00
HOTR*[24]	ResNet-50	55.2	66.4	51.41	88.78
CDN*[25]	ResNet-50	63.9	64.4	51.14	93.19
DisTR*[32]	ResNet-50	66.2	68.5	57.31	94.23
LOGICHOI*[33]	ResNet-50	64.4	65.6	70.22	101.45
OpenCat[68]	ResNet-101	61.9	63.2	67.17	96.24
CQL[29]	ResNet-50	64.4	65.6	63.98	98.67
MUREN*[54]	ResNet-50	68.8	71.0	69.32	98.74
HTCCL (ours)	ResNet-50	69.2	71.6	72.8	101.3
HTCCL-S (ours)	ResNet-50	68.9	70.7	68.9	98.1

4.6 Ablation study

We conduct ablation studies in two specific modes on the V-COCO dataset to verify the effectiveness of LCAL, GCAL, and the MKIM.

LCAL. The heterogeneous structure in the LCAL module is crucial for aggregating diverse messages. As shown in Table 4, LCAL enhances the model's performance by 1.5%. To illustrate the impact of our approach, we also assess the effectiveness of aggregating intra-class and inter-class messages separately by computing M^{ent} and denotes the inter-

Table 4 Effect of intra-class entity message passing, inter-class behavior message passing, and heterogeneous structure to model. The best results are marked in bold.

Method	AP^{S1}_{role} (%)	AP^{S1}_{role} (%)
Without M^{ent}	68.7	70.3
Without M^{beh}	67.6	69.5
Homogeneous (M^{ent})	66.1	68.0
Homogeneous (M^{beh})	65.8	67.5
Our method	**69.1**	**71.6**

class behavior message M^{beh} during inference. To further validate the significance of the heterogeneous structure in LCAL, we design a homogeneous graph model, where humans and objects are represented as the same type of nodes. In the homogeneous graph, the message passing process uses the same formula. We set the message passing formulas to be identical to the intra-class entity function and inter-class behavior function used in the heterogeneous graph. As shown in Tables 4, both the aggregation of intra-class messages and inter-class messages lead to significant improvements. These messages enable nodes to effectively gather information from their neighbors, guided by entity context and behavior context. In both configurations, using either method considerably enhances model performance. The results demonstrate that both types of contextual knowledge are crucial for improving HOI feature representation. Additionally, Table 4 indicates that the heterogeneous structure outperforms the homogeneous structure. Under the heterogeneous structure, the model can predict HOIs with greater accuracy.

GCAL. To study the impact of the GCAL module on subtask propagation, we incrementally add the global event context propagation to each branch. When the event relationship context is propagated to any detection branch, we observe consistent performance improvements compared to the baseline, regardless of which branch received the global event context (especially the interaction branch), as shown in Table 5. Our model achieves optimal performance by propagating event information to all branches. These results indicate that propagating global event context is crucial for comprehensive relationship understanding in the model.

MKIM. We aim to enhance interaction capabilities in HTCCL by transferring textual knowledge via CLIP. We explore different strategies to assess their impact on model performance. The results in Tables 6 and 7 confirm the importance of using a text encoder in MKIM to generate text embeddings, enhancing interaction features. Specifically, MKIM+GCAL improves GCAL by 3.4%, likely because MKIM embeds triplet event knowledge into interaction queries, and GCAL further enhances this prior information by exploiting event relationship dependencies. Additionally, as shown in Table 6, we also utilize the BERT[72] to generate text embeddings. The results show that CLIP outperforms BERT. The BERT model lacks consistency between visual and textual features, leading to incompatibility between interaction features and text embeddings. In contrast, the CLIP model sufficiently encodes triplets, making its text embeddings compatible with visual features and suitable for extracting interaction features.

Furthermore, Table 8 evaluates the effectiveness of simpler structures in MKIM by isolating the contributions of individual components, such as the Image Encoder (Image E) and Text Encoder (Text E). When either the visual (Image E) or textual (Text E) embedding is used alone, the model achieves lower performance compared to the full integration of both embeddings. This result highlights the importance of

Table 5 Effect of global event context to each branch. The "human", "object", and "interaction" refers to the human, object, and interaction branch, respectively. √ denotes these components are used and × represents these components are not used. The best results are marked in bold.

Human	Object	Interaction	AP^{S1}_{role} (%)	AP^{S2}_{role} (%)
×	×	×	63.5	66.2
√	×	×	65.4	67.1
×	√	×	64.2	66.6
√	√	×	64.0	66.3
×	×	√	63.7	65.6
√	√	√	**69.3**	**71.6**

Table 6 Effect of interaction knowledge transfer strategies. √ denotes these components are used and × represents these components are not used. The best results are marked in bold.

CLIP	BERT	LCAL	GCAL	AP^{S1}_{role} (%)	AP^{S2}_{role} (%)
×	√	√	√	63.5	66.2
√	×	×	×	66.4	68.6
×	×	×	√	64.2	66.6
×	×	√	×	66.7	68.9
√	×	√	√	**69.3**	**71.6**

Table 7 Effect of LCAL, global, and MKIM to model. √ donates these components are used and × represents these components are not used. The best results are marked in bold.

MKIM	LCAL	Global	AP^{S1}_{role} (%)	AP^{S1}_{role} (%)
×	×	×	62.6	65.2
√	×	×	63.1	65.8
√	√	×	66.9	68.7
√	×	√	67.6	69.4
√	√	√	**69.1**	**71.6**

Table 8 Effect of interaction knowledge integration strategies. Image E represents the Image Encoder, and Text E represents the Text Encoder. √ denotes these components are used and × represents these components are not used. The best results are marked in bold.

CLIP	BERT	Image E	Text E	AP^{S1}_{role} (%)	AP^{S2}_{role} (%)
×	√	√	√	67.5	68.2
√	×	×	√	64.7	66.3
×	×	√	×	65.1	67.4
√	×	√	√	**69.3**	**71.6**

aggregating both semantic and spatial information for effective interaction feature extraction. Moreover, replacing CLIP with BERT in the presence of both Image E and Text E leads to a performance drop, further supporting the advantage of using CLIP's alignment between visual and textual embeddings. Overall, these experiments validate the superiority of our design and demonstrate that both the quality of embeddings and the integration of complementary modalities are crucial for HOI detection.

Performance analysis in diverse complexity scenarios. The effectiveness of our model is inherently linked to the complexity of the scenes, particularly due to its ability to capture information among entities playing the same role in HOI and across similar events. To assess the impact of scene complexity, we divid the dataset into two subsets: complex scenes involving multiple entities and interactions, and simple scenes with only one human and one object. Based on the dataset annotations, we categorize the HICO-DET test set into around 7k (k = 1000) complex scenes and 2.5k simple scenes, and the V-COCO test set into approximately 4k complex scenes and 600 simple scenes. We evaluate our model on these subsets and compared its performance against other state-of-the-art models. The results, presented in Table 9, demonstrate significant improvements in complex scenarios, where intra-class entity information notably enhanced the outcomes. This suggests that our model effectively discerns relevant homogeneous entities through intra-class contextual reasoning, allowing it to learn more from relevant entities while suppressing irrelevant ones. Additionally, the inter-class behavior reasoning and the propagation of event-level context further strengthen the model's performance. The mAP increases with the number of relationships, which shows that the inter-behavior and inter-event messages allow the model to better understand the relationships between different action types, enhancing its ability to correctly predict interactions in diverse and complex scenes, leading to more accurate predictions of HOI triplets.

Effect of merging human and object branches. In HOI detection, humans play a central and active role, distinct from the relatively passive role of objects. This distinction necessitates a dedicated module to capture relevant attributes and semantics, such as pose and attire. To evaluate the effect of merging the human and object branches, we conduct experiments where the parameters of the two branches are progressively shared. Table 10 presents the results of these evaluations, where we incrementally increase the number of shared layers between the two branches. The results indicate that increasing the number of shared layers leads to a noticeable decline in performance. Compared to the non-shared model HTCCL-(0), the fully shared model HTCCL-(6) results in performance drops of 3.1% and 1.7% for the default and known object scenarios, respectively. This represents a significant degradation relative to HTCCL, which is designed to have separate human and object branches.

4.7 Qualitative results

Figure 4 visualizes the HOI detection results, illustrating the attention distributions across different branches of our model. The heatmaps highlight how

Table 9 Analysis of the model performance (mAP) in different complex scenarios of the V-COCO and HICO-DET datasets. We compare our full model with ablated versions and other state-of-the-art models. Results are reported for different complexity scenarios, where UR indicates Unary-Relation simple scenario, and MR indicates Multi-Relation complex scenario. The best results are marked in bold. W/o: without.

(%)

Dataset	Scenario	Ours	W/o LACL M^{beh}	W/o LACL M^{ent}	W/o GCAL	MUREN	LOGICHOI	Opencat
V-COCO	UR	**63.9**	63.2	61.6	62.7	63.8	61.3	59.6
	MR	**72.8**	71.6	70.5	71.4	71.7	66.4	64.3
HICO-DET	UR	**32.4**	31.7	32.2	31.6	31.2	32.3	30.5
	MR	35.7	35.0	33.9	34.8	34.8	**36.1**	34.2

Table 10 Performance comparison of models with varying levels of parameter sharing between the human and object branches. HTCCL-(0) represents the model with no shared layers, while HTCCL-(6) represents the fully shared model. HTCCL-(2) and HTCCL-(4) represent the design of 2-layer shared parameters and 4-layer shared parameters, respectively. The best results are marked in bold.

Model	Number of parameters (mega)	AP^{S1}_{role} (%)	AP^{S2}_{role} (%)
HTCCL-(0)	72.8	**69.3**	**71.6**
HTCCL-(2)	70.6	68.4	70.7
HTCCL-(4)	67.4	67.9	69.2
HTCCL-(6)	64.2	64.2	66.9

each branch—human, object, and interaction—attends to various parts of the image during the detection process. Notably, our model demonstrates extensive contextual interaction, especially in complex scenarios involving multiple relationships. The HTCCL model effectively learns from surrounding entities, capturing nuanced contextual cues that contribute to more accurate HOI predictions. This behavior is particularly evident in challenging cases where the model leverages nearby objects or actions to inform its understanding of the primary interaction, showcasing the strength of our approach in complex, multi-relational contexts.

5 Conclusion

In this paper, we show the challenges in HOI detection, particularly the limitations in modeling multi-level contextual interactions and insufficient information exchange between entity, action, and event levels. To tackle these challenges, we propose the HTCCL model, which introduces a hierarchical structure to effectively capture fine-grained interactions at the entity, action, and event levels. Our model integrates heterogeneous graph networks to facilitate intra-class and inter-class message passing, while leveraging LCAL and GCAL modules to enhance the understanding of complex interactions to reason about nuanced relationships across different contextual layers, improving robustness in diverse HOI detection tasks. Extensive experiments on two benchmark datasets, V-COCO and HICO-DET, further highlight the effectiveness of our model, outperforming state-of-the-art methods in both accuracy and contextual reasoning. In future work, we will further explore how to leverage the capabilities of the large multimodal models to achieve more accurate detection and enable open-vocabulary HOI detection.

Acknowledgment

This research work was supported by the National Natural Science Foundation of China (Nos. 62176046 and 62306064).

References

[1] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. J. Li, D. A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.

[2] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, Visual

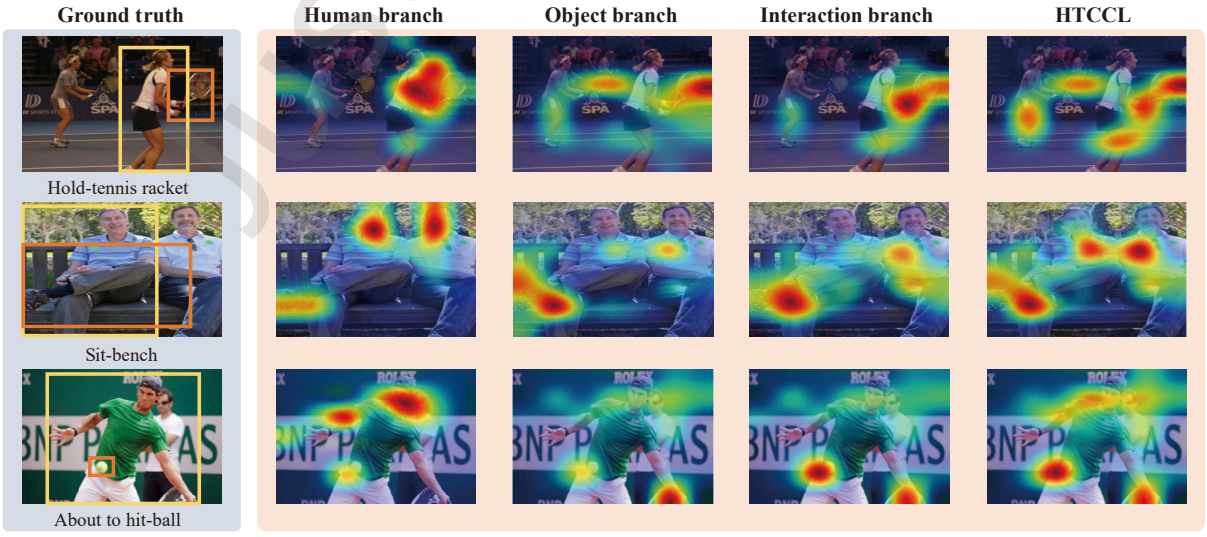

Fig. 4 Visualization of attention maps across different branches of the HTCCL model for various HOI instances. The first column shows the ground truth, followed by attention distributions from the human, object, interaction branches, and the final HTCCL output.

relationship detection with language priors, in *Proc. 14th European Conf. Computer Vision*, Amsterdam, The Netherlands, 2016, pp. 852–869.

[3] A. Gordo and D. Larlus, Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval, in *Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 5272–5281.

[4] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, Image captioning: Transforming objects into words, in *Proc. 33rd Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2019, p. 999.

[5] G. Moon, H. Kwon, K. M. Lee, and M. Cho, IntegrAlaction: Pose-driven feature integration for robust human action recognition in videos, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 3334–3343.

[6] H. Wu, M. Wang, W. Zhou, H. Li, and Q. Tian, Contextual similarity distillation for asymmetric image retrieval, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 9479–9488.

[7] M. Wu, X. Zhang, X. Sun, Y. Zhou, C. Chen, J. Gu, X. Sun, and R. Ji, DIFNet: Boosting visual information flow for image captioning, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 17999–18008.

[8] T. Yao, Y. Pan, Y. Li, and T. Mei, Exploring visual relationship for image captioning, in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 711–727.

[9] S. Yoon, W. Y. Kang, S. Jeon, S. Lee, C. Han, J. Park, and E. S. Kim, Image-to-image retrieval by learning similarity between scene graphs, in *Proc. 35th AAAI Conf. Artificial Intelligence*, Virtual Event, 2021, pp. 10718–10726.

[10] N. Ma, Z. Wu, Y. M. Cheung, Y. Guo, Y. Gao, J. Li, and B. Jiang, A survey of human action recognition and posture prediction, *Tsinghua Science and Technology*, vol. 27, no. 6, pp. 973–1001, 2022.

[11] E. Khezri, H. Hassanzadeh, R. O. Yahya, and M. Mir, Security challenges in Internet of Vehicles (IoV) for ITS: A survey, *Tsinghua Science and Technology*, vol. 30, no. 4, pp. 1700–1723, 2025.

[12] G. V. Reddy, K. Deepika, L. Malliga, D. Hemanand, C. Senthilkumar, S. Gopalakrishnan, and Y. Farhaoui, Human action recognition using difference of Gaussian and difference of wavelet, *Big Data Mining and Analytics*, vol. 6, no. 3, pp. 336–346, 2023.

[13] S. Al-Janabi and A. H. Salman, Sensitive integration of multilevel optimization model in human activity recognition for smartphone and smartwatch applications, *Big Data Mining and Analytics*, vol. 4, no. 2, pp. 124–138, 2021.

[14] B. Yang, Z. Jin, Y. Cheng, X. Ji, and W. Xu, Adversarial robustness analysis of LiDAR-included models in autonomous driving, *High-Confid. Comput.*, vol. 4, no. 1, p. 100203, 2024.

[15] Y. Yang, P. Hu, J. Shen, H. Cheng, Z. An, and X. Liu, Privacy-preserving human activity sensing: A survey, *High-Confid. Comput.*, vol. 4, no. 1, p. 100204, 2024.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, End-to-end object detection with transformers, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 213–229.

[17] J. Zhang, C. Xu, S. Shen, J. Zhu, and P. Zhang, MFF-YOLO: An improved YOLO algorithm based on multi-scale semantic feature fusion, *Tsinghua Science and Technology*, vol. 30, no. 5, pp. 2097−2113, 2025.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[19] L. Gao, L. Cui, S. Chen, L. Deng, X. Wang, X. Yan, and H. Zhu, AMTrans: Auto-correlation multi-head attention transformer for infrared spectral deconvolution, *Tsinghua Science and Technology,* vol. 30, no. 3, pp. 1329−1341, 2025.

[20] M. Chen, Y. Liao, S. Liu, Z. Chen, F. Wang, and C. Qian, Reformulating HOI detection as adaptive set prediction, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 9000–9009.

[21] B. Kim, J. Mun, K. W. On, M. Shin, J. Lee, and E. S. Kim, MSTR: Multi-scale transformer for end-to-end human-object interaction detection, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 19556–19565.

[22] S. Kim, D. Jung, and M. Cho, Relational context learning for human-object interaction detection, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 2925–2934.

[23] C. Zou, B. Wang, Y. Hu, J. Liu, Q. Wu, Y. Zhao, B. Li, C. Zhang, C. Zhang, Y. Wei, et al., End-to-end human object interaction detection with HOI transformer, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 11820–11829.

[24] B. Kim, J. Lee, J. Kang, E. S. Kim, and H. J. Kim, HOTR: End-to-end human-object interaction detection with transformers, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 74–83.

[25] A. Zhang, Y. Liao, S. Liu, M. Lu, Y. Wang, C. Gao, and X. Li, Mining the benefits of two-stage and one-stage HOI detection, in *Proc. 35th Int. Conf. Neural Information Processing Systems*, Virtual Event, 2021, p. 1316.

[26] F. Z. Zhang, D. Campbell, and S. Gould, Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 20072–20080.

[27] W. Jiang, W. Ren, J. Tian, L. Qu, Z. Wang, and H. Liu, Exploring self- and cross-triplet correlations for human-object interaction detection, in *Proc. 38th AAAI Conf. Artificial Intelligence*, Vancouver, Canada, 2024, pp.

2543–2551.

[28] M. Tamura, H. Ohashi, and T. Yoshinaga, QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 10405–10414.

[29] C. Xie, F. Zeng, Y. Hu, S. Liang, and Y. Wei, Category query learning for human-object interaction classification, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 15275–15284.

[30] S. Ning, L. Qiu, Y. Liu, and X. He, HOICLIP: Efficient knowledge transfer for HOI detection with vision-language models, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 23507–23517.

[31] J. Lim, V. M. Baskaran, J. M. Y. Lim, K. Wong, J. See, and M. Tistarelli, ERNet: An efficient and reliable human-object interaction detection network, *IEEE Trans. Image Process*., vol. 32, pp. 964–979, 2023.

[32] D. Zhou, Z. Liu, J. Wang, L. Wang, T. Hu, E. Ding, and J. Wang, Human-object interaction detection via disentangled transformer, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 19546–19555.

[33] L. Li, J. Wei, W. Wang, and Y. Yang, Neural-logic human-object interaction detection, in *Proc. 37th Int. Conf. Neural Information Processing Systems*, New Orleans, LA, USA, 2024, p. 924.

[34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in *Proc. 38th Int. Conf. Machine Learning*, Virtual Event, 2021, pp. 8748–8763.

[35] S. Gupta and J. Malik, Visual semantic role labeling, arXiv preprint arXiv: 1505.04474, 2015.

[36] Y. W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, Learning to detect human-object interactions, in *Proc. 2018 IEEE Winter Conf. Applications of Computer Vision* (*WACV*), Lake Tahoe, NV, USA, 2018, pp. 381–389.

[37] D. J. Kim, X. Sun, J. Choi, S. Lin, and I. S. Kweon, Detecting human-object interactions with action co-occurrence priors, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 718–736.

[38] Y. L. Li, X. Liu, X. Wu, X. Huang, L. Xu, and C. Lu, Transferable interactiveness knowledge for human-object interaction detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3870–3882, 2022.

[39] 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. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 17152–17162.

[40] A. Bansal, S. S. Rambhatla, A. Shrivastava, and R. Chellappa, Detecting human-object interactions via functional generalization, in *Proc. 34th AAAI Conf. Artificial Intelligence*, New York, NY, USA, 2020, pp. 10460–10469.

[41] X. Liu, Y. L. Li, and C. Lu, Highlighting object category immunity for the generalization of human-object interaction detection, in *Proc. 36th AAAI Conf. Artificial Intelligence*, Virtual Event, 2022, pp. 1819–1827.

[42] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, and C. W. Chen, Exploring structure-aware transformer over interaction proposals for human-object interaction detection, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 19526–19535.

[43] S. Ren, K. He, R. Girshick, and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in *Proc. 29th Int. Conf. Neural Information Processing Systems*, Montreal, Canada, 2015, pp. 91–99.

[44] C. Gao, Y. Zou, and J. B. Huang, iCAN: Instance-centric attention network for human-object interaction detection, in *Proc. British Machine Vision Conf. 2018*, Newcastle, UK, 2018, p. 41.

[45] G. Gkioxari, R. Girshick, P. Dollár, and K. He, Detecting and recognizing human-object interactions, in *Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 8359–8367.

[46] F. Z. Zhang, D. Campbell, and S. Gould, Spatially conditioned graphs for detecting human-object interactions, in *Proc. 2021 IEEE/CVF Int. Conf. Computer Vision*, Montreal, Canada, 2021, pp. 13299–13307.

[47] S. Qi, W. Wang, B. Jia, J. Shen, and S. C. Zhu, Learning human-object interactions by graph parsing neural networks, in *Proc. 15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 407–423.

[48] C. Gao, J. Xu, Y. Zou, and J. B. Huang, DRG: Dual relation graph for human-object interaction detection, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 696–712.

[49] T. Gupta, A. Schwing, and D. Hoiem, No-frills human-object interaction detection: Factorization, layout encodings, and training techniques, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 9676–9684.

[50] P. Zhou and M. Chi, Relation parsing neural network for human-object interaction detection, in *Proc. 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 843–851.

[51] Y. L. Li, X. Liu, H. Lu, S. Wang, J. Liu, J. Li, and C. Lu, Detailed 2D-3D joint representation for human-object interaction, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 10163–10172.

[52] Y. Liu, J. Yuan, and C. W. Chen, ConsNet: Learning consistency graph for zero-shot human-object interaction detection, in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, 2020, pp. 4235–4243.

[53] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli, Learning to detect human-object interactions with knowledge, in *Proc. 2019 IEEE/CVF Conf. Computer*

Vision and Pattern Recognition (*CVPR*), Long Beach, CA, USA, 2019, pp. 2019–2028.
[54] B. Kim, T. Choi, J. Kang, and H. J. Kim, UnionDet: Union-level detector towards real-time human-object interaction detection, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 498–514.
[55] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, PPDM: Parallel point detection and matching for real-time human-object interaction detection, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 479–487.
[56] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770–778.
[57] A. S. M. Iftekhar, H. Chen, K. Kundu, X. Li, J. Tighe, and D. Modolo, What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions, in *Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 5343–5353.
[58] H. Wang, W. S. Zheng, and L. Yingbiao, Contextual heterogeneous graph network for human-object interaction detection, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 248–264.
[59] H. W. Kuhn, The Hungarian method for the assignment problem, *Nav. Res. Logist. Quar.*, vol. 2, nos. 1&2, pp. 83–97, 1955.
[60] S. Ren, K. He, R. Girshick, and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.
[61] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, Generalized intersection over union: A metric and a loss for bounding box regression, in *Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 658–666.
[62] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, Focal loss for dense object detection, in *Proc. 2017 IEEE Int. Conf. Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 2999–3007.
[63] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, Microsoft COCO: Common objects in context, in *Proc. 13th European Conf. Computer Vision*, Zurich, Switzerland, 2014, pp. 740–755.
[64] I. Loshchilov and F. Hutter, Decoupled weight decay regularization, in *Proc. 7th Int. Conf. Learning Representations*, New Orleans, LA, USA, 2019. https://openreview.net/forum?id=Bkg6RiCqY7.
[65] O. Ulutan, A. S. M. Iftekhar, and B. S. Manjunath, VSGNet: Spatial attention network for detecting human object interactions using graph convolutions, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 13614–13623.
[66] Z. Hou, X. Peng, Y. Qiao, and D. Tao, Visual compositional learning for human-object interaction detection, in *Proc. 16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 584–600.
[67] Y. L. Li, X. Liu, X. Wu, Y. Li, and C. Lu, HOI analysis: Integrating and decomposing human-object interaction, in *Proc. 34th Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2020, p. 421.
[68] S. Zheng, B. Xu, and Q. Jin, Open-category human-object interaction pre-training via language modeling framework, in *Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, Canada, 2023, pp. 19392–19402.
[69] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, Learning human-object interaction detection using interaction points, in *Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 4115–4124.
[70] X. Zhong, X. Qu, C. Ding, and D. Tao, Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection, in *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021, pp. 13229–13238.
[71] Y. Wu, A. Kirillov, F. Massa, W. Y. Lo, and R. Girshick, Detectron2, https://github.com/facebookresearch/detectron2, 2019.
[72] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (*Long and Short Papers*), Minneapolis, MN, USA, 2019, pp. 4171–4186.

Ke Qin received the MEng and PhD degrees from University of Electronic Science and Technology of China, Chengdu, China, in 2006 and 2010, respectively. He is currently a full professor at University of Electronic Science and Technology of China, Chengdu, China. His research interests include neural networks, machine learning, and machine reasoning.

Xin Hu received the master degree from University of Southampton, Southampton, UK, and the bachelor degree from Southwest Jiaotong University, Chengdu, China. Now, he is pursuing the PhD degree at University of Electronic Science and Technology of China, Chengdu, China. His research interests primarily focuse on computer vision.

Guangchun Luo received the MEng and PhD degrees from University of Electronic Science and Technology of China, Chengdu, China, in 1999 and 2004, respectively. He is now the director of Ubiquitous Intelligence and Trusted Services Key Laboratory of Sichuan Province, University of Electronic Science and Technology of China, Chengdu, China. His research interests include deep learning, as well as decision theory and its applications.

Tao He received the PhD degree from Monash University, Melbourne, Australia, in 2022. He is currently a deputy researcher at University of Electronic Science and Technology of China, Chengdu, China. He has published over ten top-tier conference and journal papers, such as *IJCV*, *ICCV*, *ECCV*, and *AAAI*. His research interests lie in scene graph generation, HOI detection, and image retrieval.