Interaction-Centric Spatio-Temporal Context Reasoning for Multi-person Video HOI Recognition

Yisong Wang[1], Nan Xi[2(✉)], Jingjing Meng[3], and Junsong Yuan[2]

[1] School of EECS, Peking University, Beijing, China
[2] Department of Computer Science and Engineering, University at Buffalo, Getzville, USA
{nanxi,jsyuan}@buffalo.edu
[3] Amazon Inc., Bellevue, USA

Abstract. Understanding human-object interaction (HOI) in videos represents a fundamental yet intricate challenge in computer vision, requiring perception and reasoning across both spatial and temporal domains. Despite previous success of object detection and tracking, multi-person video HOI recognition still faces two major challenges: (1) the three facets of HOI (*human*, *objects*, and the *interactions* that bind them) exhibit interconnectedness and exert mutual influence upon one another. (2) the complexity of multi-person multi-object combinations in spatio-temporal interaction. To address them, we design a spatio-temporal context fuser to better model the interactions among persons and objects in videos. Furthermore, to equip the model with temporal reasoning capacity, we propose an interaction state reasoner module on top of context fuser. Considering the interaction is a key element to bind human and object, we propose an interaction-centric hypersphere in the feature embedding space to model each category of interaction. It helps to learn the distribution of HOI samples belonging to the same interactions on the hypersphere. After training, each interaction prototype sphere will fit the testing HOI sample to determine the HOI classification result. Empirical results on multi-person video HOI dataset MPHOI-72 indicate that our method remarkably surpasses state-of-the-art (SOTA) method by more than 22% F_1 score. At the same time, on single-person datasets Bimanual Actions (single-human two-hand HOI) and CAD-120 (single-human HOI), our method achieves on par or even better results compared with SOTA methods. Source code is released at the following link: https://github.com/southnx/IcH-Vid-HOI.

Keywords: interaction-centric · spatio-temporal context reasoning · multi-person human-object interaction

Y. Wang and N. Xi—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-73411-3_24.

A. Leonardis et al. (Eds.): ECCV 2024, LNCS 15090, pp. 419–435, 2025.
https://doi.org/10.1007/978-3-031-73411-3_24

1 Introduction

Video-based Human-Object Interaction (HOI) recognition aims to identify the interactions occurring between human and object entities within video frames. Precisely recognizing HOIs in real world scenarios is essential for a bunch of applications, such as assisting patients by recognizing daily activities and predicting pedestrian movements to avoid accidents for autonomous vehicles.

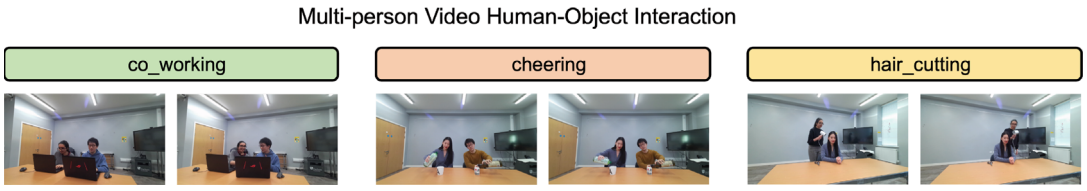

Fig. 1. Examples of multi-person video HOI. Three categories of interactions are presented, including *co_working*, *cheering*, *hair_cutting*.

Most existing HOI recognition research focus on static images [19,36], with much less attention on video-based HOI recognition. Video-based HOI recognition is more demanding than image-based HOI recognition due to the necessity of comprehending complex spatio-temporal dynamics and reasoning about multi-human and multi-object interaction combinations. The complexity is further exacerbated when dealing with multi-person circumstances, as examples shown in Fig. 1. In such cases, various human and object entities mutually influence each other, resulting in intricate interdependencies within the scene. Additionally, the three components (*human*, *object* and *interactions* that the entities are involved in) of HOI exhibit interwind structures, *e.g.*, the possible interactions that can occur within a scene given a human and an object. However, current video HOI recognition methods [23,27] do not fully explore such inherent structural nature of HOI components. Instead, they often opt for disentangled representations for each component, which may bring suboptimal representation capabilities.

To overcome the aforementioned limitations, we introduce an interaction-centric spatio-temporal context reasoning approach for representing video HOIs. For the purpose of enhancing the awareness of complex HOI structures in our representations, we introduce the Context Fuser (CF), which encodes both entity representations and interaction representations. Moreover, to empower our model with the ability to reason about interaction state transitions across video frames, we propose the Interaction State Reasoner (ISR) module for generating interaction representations. In addition, we employ a bidirectional Gated Recurrent Unit (BiGRU) to model temporal dynamics across video frames. This multi-level representation learning framework not only facilitates effective exploration of the interdependencies among structured HOI components but also empowers the model with interaction reasoning capabilities in both spatial and temporal domains. To compute the probability of an HOI in the scene, we introduce an

interaction-centric hypersphere. This approach leverages the concept of a hypersphere, where the *interaction* is located at its center, while the *human* and *object* entities involved in that interaction are situated on the hypersphere's surface. We assume that human-object entities belonging to each interaction class are located on their respective hyperspheres, specific to that interaction class.

Concretely, we generate context-rich and reasoning-aware video HOI representations through three key components: the Context Fuser (CF), Interaction State Reasoner (ISR) and Bidirectional GRU (BiGRU). The CF module integrates context information from human-object entities and interactions. It comprises three fuser blocks for humans, objects, and interactions. The Object Fuser Block processes local video frame data, enhancing object features. The Interaction Fuser Block combines human and object representations with interaction-specific characteristics. Additionally, in multi-person scenes, the Human Fuser Block captures human representations influenced by others. This approach fosters comprehensive HOI representations via effective context fusion. To facilitate interaction reasoning, we place the ISR module on top of the context fuser module, yielding entity representations capable of capturing interaction transition dynamics. These entity representations are then input into the BiGRU module to model temporal dynamics across video frames, thereby ensuring a comprehensive understanding of the evolving context and interactions within the video data. Finally, we determine interaction classes in each frame with the interaction-centric hypersphere, computing the probability of human-object entities belonging to specific interaction classes.

We assess our model's performance on three video-based HOI datasets: MPHOI-72 [27] (multi-person), Bimanual Actions [4] and CAD-12 [12] (single-person). Our results highlight our model's superiority in multi-person scenarios, achieving an impressive over 22% F_1 score improvement over the current state-of-the-art (SOTA). In single-person scenarios, our method delivers on par or even better performance compared to the current SOTA method. Our major contributions are summarized as follows:

- To represent inherent HOI manifold structures, we propose an interaction-centric hypersphere representation scheme. This scheme explicitly guide the process of learning intrinsic structural nature of HOI and elucidates the inter-dependencies among its components.
- To learn context-rich and reasoning-aware entity representations, we introduce context fuser and interaction state reasoning modules. This enhancement results in entity representations that are highly suitable for video-based HOI tasks.
- Extensive experiment results showcase that our method achieves SOTA performance with a huge improvement of more than 22% F_1 score over existing methods in multi-person scenario. Additionally, our model achieves competitive results in single-person cases compared to SOTA method.

2 Related Works

HOI Detection in Images: HOI detection in images aims at understanding interactions in images between humans and objects. Different methods have been proposed in previous studies. Some works propose Convolutional Neural Networks (CNN)-based methods which can be further divided into one-stage methods [9,17,37] and two-stage methods [5,7,16,33]. However, these methods usually lack of ability to capture global context information. Recently, Transformer-based models [8,10,31,36] became the main approach for the HOI task. Following the architecture of DETR, these models achieved superior performance on HOI detection. Moreover, some works also utilize graph [25] and interactiveness field [19] to achieve better performance. These various approaches to image HOI provide the fundamentals for video HOI recognition.

HOI Recognition in Videos: Video HOI recognition is the foundation of many real-world applications, including understanding surgical activities in the operation room [34] and guiding 3D human reconstruction [20,21]. Video-based HOI recognitions have to deal with both spatial and temporal reasoning. Before the use of neural networks, some early studies formulated this task using the Markov model [12] to utilize temporal cues. In [24], HOI hotspots in videos are learned in a novel approach, with two networks trained jointly to capture spatial regions where actions happen. While [6] focuses on visual relation prediction in open-vocabulary with pretrained visual-language models. Recent works have used Recurrent Neural Networks (RNN) combined with Graph Neural Networks (GNN) [23,26,27,30] to predict human-object relations in videos. Inspired by ViT [3], some works also propose Transformer-based methods to reason spatial relations better [32]. However, RNN-based models usually require complex training strategies or long training time in order to achieve the best performance. Moreover, when multiple persons are involved in an activity jointly, these methods lack the ability to model their collaboration, resulting in poor performance when the interactions are performed by multiple persons.

Hyperspheres for Class Representation: Hyperspheres have been demonstrated to be an effective approach to model class representation [2,22]. Geometrical modeling strategy has been proposed in [2], where the effectiveness has been confirmed. This approach proves advantageous for capturing and representing enriched class-level information, particularly well-suited for creating measurements in Euclidean space. Consequently, it is naturally adaptable to structured prediction tasks.

3 Motivation

Video HOI recognition task involves identifying both the human and object entities engaged in an interaction across a sequence of video frames. This task encompasses spatial and temporal aspects, as it requires understanding the relationships between humans, objects, and their interactions over time. However,

current methods for video HOI recognition often neglect this crucial dependency structure, resulting in the separation of learned representations associated with humans, objects, and their interactions, ultimately compromising their representational accuracy. To address this issue, we propose a novel approach featuring a context fuser and an interaction state reasoner to capture spatio-temporal contextual representations. Additionally, we introduce an interaction-centric hypersphere to represent HOI manifold structure and model the inter-dependency among predicted humans, objects, and interactions.

4 Method

4.1 Problem Formulation

For a video dataset $\mathcal{V}$, given a video clip $V \in \mathcal{V}$ containing T video frames $\{v_1, ..., v_T\}$, video HOI recognition aims to predict the temporal segmentation of interactions between human and object entities across all the video frames. Formally, we aim to learn an HOI recognition model $\mathcal{M}$ that outputs the segmentation of human's sub-activity $\{s_n\}_{n=1}^{N}$ in each frame, where N is the number of human sub-activity segments. Each segment s_n is represented as an interval from its start time t_n to end time t_{n+1}: $s_n = [t_n, t_{n+1})$. The start and end time of each segment are determined from interaction probability prediction $\{u^t\}_{t=1}^{T}$ of each frame, where $u^t \in \mathbb{R}^K$, K is the number of possible interaction classes.

4.2 Model Design

In the following section, we introduce our interaction-centric hypersphere reasoning model for video HOI recognition in detail. As shown in Fig. 2, our major idea is to construct a hypersphere to represent each HOI in the scene. For each hypersphere, the *interaction* locates at the center of the hypersphere, while the corresponding *human-object* entity belongs to that *interaction* locates at the surface of the hypersphere. We construct Context Fuser (CF) module to learn context-rich human-object entity representations. For the aim of enabling model with reasoning ability over interaction state transitions, we propose Interaction State Reasoner (ISR) to reason on whether the current interaction will be continued or stopped. To model the temporal dynamics of HOI in videos, we update human-object entity representations $\{\hat{z}_E^t\}_{t=1}^{T}$ along the temporal domain with bidirectional GRU (BiGRU). Finally, predicted interaction class probability u^t of each frame is computed from the interaction-centric hypersphere.

Context Fuser. For a sequence of video frames $\{v_t\}_{t=1}^{T}$, we follow 2G-GCN [27] to extract feature of humans and objects from backbone network. The extracted human features $z_H \in \mathbb{R}^d$ contain both bounding box information and skeleton keypoint information, where d indicates the feature dimension. Object features $z_O \in \mathbb{R}^d$ contain only bounding box information.

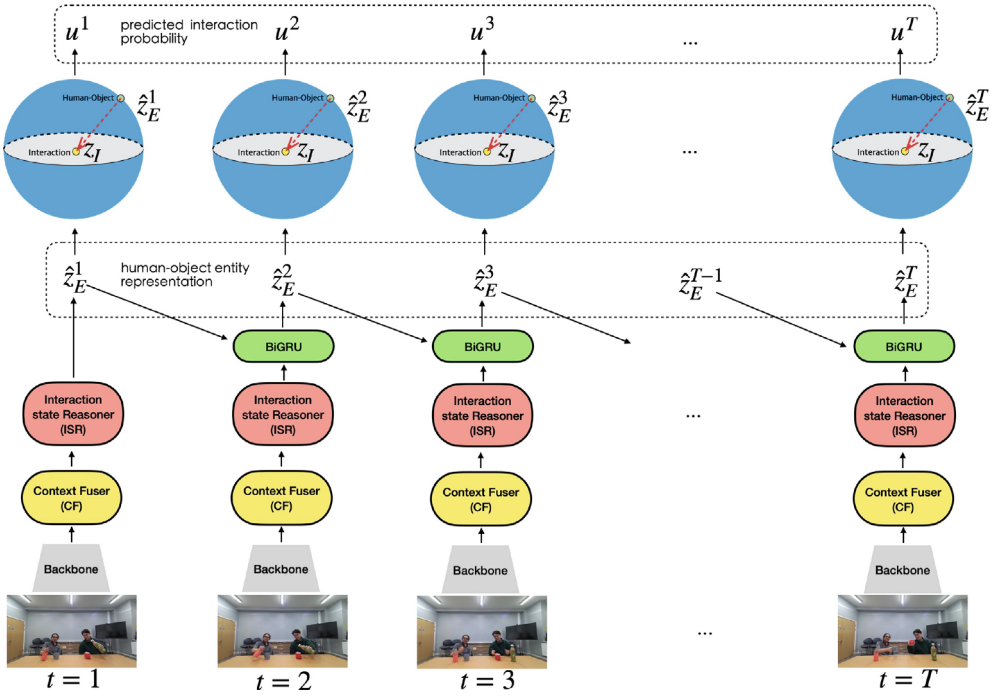

Fig. 2. Model Overview. Each video frame is applied with a backbone for feature extraction. Subsequently, a context fuser and interaction state reasoner is employed for learning interaction representations z_I and human-object entity representations $\{\hat{z}_E^t\}_{t=1}^T$. Bidirectional GRU is further utilized to model temporal dynamics across video frames. The final predicted interaction probability $\{u^t\}_{t=1}^T$ is computed from the interaction-centric hypersphere. The dimension of each hypersphere is d, which is the same with the feature dimension in our model.

We design a context fuser (CF) module shown in Fig. 3 to generate human-object entity representations $\{z_E^t\}_{t=1}^T$ based on human, object and contextual information. In multi-person circumstances, CF contains object fuser block, interaction fuser block and human fuser block sequentially. First, we design an object fuser block to incorporate object representations in local temporal regions into current object representations, generating learned object representation $\hat{z}_{O_1}$ and $\hat{z}_{O_2}$:

$$\begin{aligned} \hat{z}_{O_1} &= FFN(SA(Q = z_{O_1}^t; K, V = z_{O_1}^{\bar{t}}) + z_{O_1}^t), \\ \hat{z}_{O_2} &= FFN(SA(Q = z_{O_2}^t; K, V = z_{O_2}^{\bar{t}}) + z_{O_2}^t), \end{aligned} \tag{1}$$

where SA indicates self-attention, FFN is feed forward network, $z_{O_1}^t \in \mathbb{R}^d$ and $z_{O_2}^t \in \mathbb{R}^d$ are the initial object features, while $z_{O_1}^{\bar{t}} \in \mathbb{R}^{20d}$ and $z_{O_2}^{\bar{t}} \in \mathbb{R}^{20d}$ are stacked object features from a local time window, $\bar{t} \in [t-10, t+10)$. Subsequently, for all the K possible interactions $\{I_i\}_{i=1}^K$ as shown in Fig. 3, each interaction class I_i is prompted as a sentence $s =$"The human is [`interact`]ing in the scene.", where [`interact`] indicates the specific interaction class. Then the sentence is applied with the text encoder ($\mathcal{F}_T$) of large-scale vision-language model CLIP [28] to initialize the interaction feature $z_I = \mathcal{F}_T(s) \in \mathbb{R}^d$. We also generate

context feature $z_C \in \mathbb{R}^d$ to represent the semantic information of each frame. Specifically, we extract frame caption c_i from video frame $i(i = 1, ..., T)$ with BLIP [15] model and apply the caption with CLIP model to extract text embedding. In order to adapt the human (z_{H_1}, z_{H_2}) features, object (z_{O_1}, z_{O_2}) features and context feature (z_C) to the specific interaction feature z_I, we construct an interaction fuser block that contains a cross-attention (CA) module followed by a Feed Forward Network (FFN), generating interaction-aware human features ($z'_{H_1} \in \mathbb{R}^d$, $z'_{H_2} \in \mathbb{R}^d$), object features ($z'_{O_1} \in \mathbb{R}^d$, $z'_{O_2} \in \mathbb{R}^d$) and context feature ($z'_C \in \mathbb{R}^d$). Furthermore, to model the influence between the two humans in multi-person scenarios, we construct a human fuser block, featuring the same architecture with interaction fuser block. For the first human (Human1), the updated human feature $\hat{z}_{H_1} \in \mathbb{R}^d$ is generated as:

$$\hat{z}_{H_1} = FFN(CA(Q = SA(z'_{H_1}); K, V = z'_{H_2}) + SA(z'_{H_1})). \tag{2}$$

Human2 feature $\hat{z}_{H_2} \in \mathbb{R}^d$ is generated in the same way as Eq. 2. Finally, the human-object entity representation $z_{E_1} \in \mathbb{R}^d$ of Human1 is computed by max-pooling operation over all the d dimensions of the four representations shown in Eq. 3:

$$z_{E_1} = MaxPool(\hat{z}_{H_1}, z'_{O_1}, z'_{O_2}, z'_C). \tag{3}$$

The human-object entity representation $z_{E_2} \in \mathbb{R}^d$ of Human2 is computed with similar approach as Eq. 3. The CF module for single-person cases are similar with multi-person, except that the human fuser block is removed and there is only one human feature as query to be fed into the interaction fuser block.

Interaction State Reasoner. To augment the model's ability in interaction state transition reasoning, we introduce an Interaction State Reasoner (ISR) module following CF module. ISR module explicitly empowers the model to determine whether the current interaction should persist or transit to another interaction. Specifically, as shown in Fig. 4, at each time t, the two possible states $state_1$ and $state_2$ represent "continue" or "stop" of an interaction, respectively. Each state is prompted as one sentence, where s_1 ="`This interaction is going to continue.`" and s_2="`This interaction is going to stop and change to another interaction.`". Then the embeddings of the two states z_{state_1} and z_{state_2} are generated from CLIP [28] text encoder $\mathcal{F}_T$: $z_{state_1} = \mathcal{F}_T(s_1) \in \mathbb{R}^d$, $z_{state_2} = \mathcal{F}_T(s_2) \in \mathbb{R}^d$. Interaction state embeddings z_{state_1} and z_{state_2} are further fed to a reasoner block (shown in Fig. 4) together with the interaction embedding $z^t_{\hat{I}} \in \mathbb{R}^d$ at time t, generating state-informed interaction embeddings $\hat{z}_{state_1}$ and $\hat{z}_{state_2}$. The reasoner block contains a FFN and a State Interpolation (SI) module. The SI module generates the weights ω_1, ω_2 for the two interaction states in the following approach:

$$\omega_1, \omega_2 = Softmax(FFN(z^t_E) \cdot [z^\top_{state_1}, z^\top_{state_2}]), \tag{4}$$

where $\top$ indicates transpose operation. Subsequently, the final human-object entity representation $\hat{z}^t_E \in \mathbb{R}^d$ at time t is generated by interpolate over the

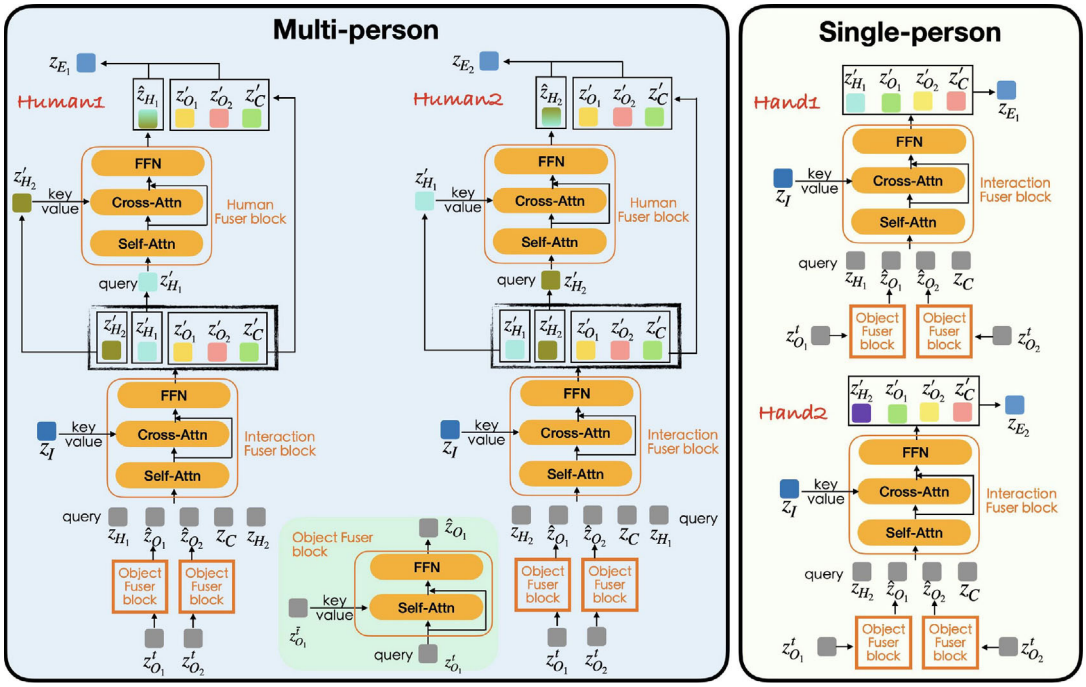

Fig. 3. Context fuser. In multi-person scenario (left), the context fuser consists of object fuser block, interaction fuser block and human fuser block. In single-person case (right), only object fuser block and interaction fuser block are employed. Detailed descriptions of the architecture and annotations can be found in Sect. 4.2.

current entity representation z_E^t at time t and the entity representation z_E^{t-1} at time $t-1$:

$$\hat{z}_E^t = \omega_1 \cdot z_E^{t-1} + \omega_2 \cdot z_E^t. \tag{5}$$

Consequently, the generated human-object entity representation $\hat{z}_E^t$ is able to reason on the possible future interaction state transitions.

Interaction-Centric Hypersphere. With the above generated human-object entity representation $\hat{z}_E^t$ and interaction representation z_I, we need to calculate the probability of a human-object entity E categorizing into the interaction class I_i. To that end, we design an interaction-centric hypersphere with interaction at the center of hypersphere and human-object entity at its surface. We call it "hypersphere" because this is a sphere in the $\mathbb{R}^d$ space, where d is the feature dimension in our model, which is obviously larger than 3. This hypersphere design models the manifold structure of HOI, which is in charge of the

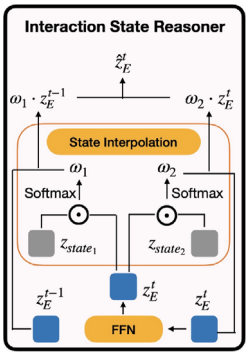

Fig. 4. Interaction State Reasoner. $\odot$ indicates inner product.

specific interaction class. Concretely, we employ a hyperspherical measurement:

$$\mathcal{U}(E, I_i) = \frac{\exp\Big(- \Big[||z_{I_i} - \hat{z}_E||_2 - \lambda \Big]_+ \Big)}{\sum_{j=1}^{K} \exp\Big(- \Big[||z_{I_j} - \hat{z}_E||_2 - \lambda \Big]_+ \Big)}, \tag{6}$$

where z_{I_i} and z_{I_j} indicate the interaction representations of interaction class i and j, respectively. $\hat{z}_E$ denotes the human-object entity representation. $[s]_+ \triangleq max(0, s)$. λ indicates the radius of the hypersphere, which is set to be a constant in our model. The higher value of $\mathcal{U}(E, I_i)$ suggests that the human and object entity E is more likely to be categorized into I_i.

4.3 Learning Objective

The learning objective of our model contains two parts: (i) focal loss $\mathcal{L}_{cls}$ for interaction classification; (ii) interaction feature loss $\mathcal{L}_{fea}$ that controls the smoothness of interaction features in local region.

Focal Loss $\mathcal{L}_{cls}$: We employ focal loss [18] for interaction classification, mitigating the interaction class imbalance problem on model performance. For each video frame $v_i (i = 1, ..., T)$, our model predicts the probability $\hat{y}_i \in \mathbb{R}^K$ of all the interaction classes. The corresponding ground-truth of interaction class $y_i \in \mathbb{R}^K$ is a binary vector. For each interaction class k, the focal loss $\mathcal{L}_{cls}^k$ is formulated as: $\mathcal{L}_{cls}^k = -(1-p_k)^\gamma \log(p_k)$, where γ is a hyperparameter to control the focusing extent, p_k is defined as: $\{p_k = \hat{y}_i^k, \text{ if } y_i^k = 1; p_k = 1 - \hat{y}_i^k, \text{ otherwise}\}$. Subsequently, the focal loss $\mathcal{L}_{cls}$ of each video frame is obtained by combining the focal loss of each individual interaction class k: $\mathcal{L}_{cls} = \Sigma_{k=1}^K \mathcal{L}_{cls}^k$.

Interaction Feature Loss $\mathcal{L}_{fea}$: We introduce interaction feature loss $\mathcal{L}_{fea}$ to control the temporal smoothness of interaction features. Our model outputs the feature of each human-object entity E in each frame. Inspired by [1], in order to improve the continuity, we minimize the feature distance in the same segment and maximize the distance between different segments for each subject. Denote u_E^t as whether the interaction will continue or change to another action for entity E at time t. $u_E^t = 1$ indicates the interaction will stop and change to another action at time t for entity E and $u_E^t = 0$ otherwise. We minimize

$$\mathcal{L}_{fea} = \frac{1}{2} \sum_{t=0}^{T-1} \Big[(1 - u_E^t)(||z_E^t - z_E^{t+1}||_2)^2 + u_E^t (max(L - ||z_E^t - z_E^{t+1}||_2, 0))^2 \Big], \tag{7}$$

where L is a threshold that controls the minimal feature distance when interaction changes. In total, the overall loss is written as:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha \mathcal{L}_{fea}, \tag{8}$$

where α is a hyperparameter to control the weight of each loss.

Model Inference. During model inference, we compute the interaction probability $\hat{y}_i \in \mathbb{R}^K$ for each video frame. The interaction class a with the highest probability is chosen as the predicted interaction for that frame: $a = \arg\max_i \hat{y}_i$.

5 Experiments

5.1 Experimental Setup

Datasets: We evaluate our method on MPHOI-72, Bimanual Actions, and CAD-120 datasets, each representing multi-person collaboration, one person with two hands and a single hand respectively.

(I) MPHOI-72 dataset is proposed in [27], which consists of multiple humans and objects in the scene. The dataset comprises 72 videos featuring 3 human subjects and 6 objects. Within each video, 2 individuals are paired to engage in 3 distinct activities, encompassing a total of 13 sub-activities, while utilizing 2 to 4 objects

(II) Bimanual Actions dataset [4] is the first HOI dataset to include two hands for subjects to perform interactions which is common in reality. There are 540 videos with one person performing activities with both hands. There are 6 subjects performing 9 different activities with 10 repetitions. There are a total of 14 action labels assigned to each hand, and entity-level annotations are provided on a per-frame basis within the video.

(III) CAD-120 dataset [13] is popular for HOI recognition. It contains 120 videos with 10 activities performed by 4 participants. There are 10 human sub-activities labeled per frame.

Evaluation Metric: We report $F_1@k$ metric [14] with thresholds $k = 10\%$, 25%, and 50%. Compared to frame-based metrics which evaluate prediction on every single frame, this metric could measure prediction continuity in action segments because it views each predicted action segment as correct only when it has the Interaction over Union (IoU) with the corresponding ground truth over the threshold k.

5.2 Implementation Details

In the experiment, we use three layers of context fuser for Bimanual and two for CAD-120 and MPHOI. The features of humans and objects are extracted from [27] and their dimension is mapped to 768, 256, and 512 for MPHOI, Bimanual, and CAD-120 respectively. More implementation details can be found in the supplementary material.

5.3 Quantitative Results

Multi-person HOI Recognition The quantitative results of joined segmentation and label recognition of sub-activity on MPHOI-72 in Table 1 show the performance of our method in multi-person HOI circumstance. Our method outperforms SOTA method 2G-GCN [27] by a large margin in all the three evaluation metrics. For F_1@10, F_1@25 and F_1@50 scores, our method surpasses 2G-GCN 23.0%, 23.7% and 22.5%, respectively. The significant improvement achieved by our method indicates that the human fuser block in the CF module effectively improves the context-aware human representation learning under multi-person scenarios.

Single-Person HOI Recognition We show the quantitative results for single-person HOI recognition in Table 2 and Table 3, which are performed on CAD-120 and Bimanual Actions datasets, respectively. Results in Table 2 show that our method performs slightly better on CAD-120 dataset compared to SOTA method 2G-GCN, with around 1% improvement over 2G-GCN on all the three metrics. For the Bimanual Actions dataset, our method performs as good as 2G-GCN in F_1@10 while achieves 0.9% and 5.0% higher than 2G-GCN in F_1@25 and F_1@50 score, respectively. These results indicate that our method achieves generally on par or even better performance on single-person video HOI recognition task.

Table 1. The results of joined segmentation and label recognition of sub-activity on MPHOI-72. ΔCF: removing context fuser module; ΔISR: removing interaction state reasoner module; $\Delta\mathcal{L}_{fea}$: removing interaction feature loss; $\lambda = 0$: ablating the hypersphere by replacing it with Euclidean distance; ΔCLIP+BLIP: ablating CLIP and BLIP model; Traditional Classifier: ablating the interaction-centric hypersphere by replacing it with a traditional classifier constructed with MLP and Softmax classifier. The improvements of our method over current SOTA method is highlighted with upward arrows.

Model	F_1@10	F_1@25	F_1@50
ASSIGN [23]	59.1 ± 12.1	51.0 ± 16.7	33.2 ± 14.0
2G-GCN [27]	68.6 ± 10.4	60.8 ± 10.3	45.2 ± 6.5
Ours	$\mathbf{91.6 \pm 0.9}$(↑ **23.0**)	$\mathbf{84.5 \pm 2.6}$(↑ **23.7**)	$\mathbf{67.7 \pm 2.2}$(↑ **22.5**)
Ours (ΔCF)	65.8 ± 12.4	57.6 ± 14.0	39.2 ± 12.6
Ours (ΔISR)	80.1 ± 5.5	73.0 ± 8.2	55.6 ± 6.1
Ours ($\Delta\mathcal{L}_{fea}$)	73.5 ± 15.7	69.7 ± 13.3	48.8 ± 13.0
Ours ($\lambda = 0$)	81.2 ± 0.7	74.8 ± 4.2	53.2 ± 0.3
Ours (ΔCLIP+BLIP)	80.0 ± 6.7	73.0 ± 9.9	55.6 ± 7.4
Ours (Traditional Classifier)	84.6 ± 7.2	74.7 ± 10.5	54.6 ± 13.7

Table 2. The results of joined segmentation and label recognition of sub-activity on CAD-120. The notations are the same with Table 1.

Model	F_1@10	F_1@25	F_1@50
rCRF [29]	65.6 ± 3.2	61.5 ± 4.1	47.1 ± 4.3
Independent BiRNN [27]	70.2 ± 5.5	64.1 ± 5.3	48.9 ± 6.8
ATCRF [11]	72.0 ± 2.8	68.9 ± 3.6	53.5 ± 4.3
Relational BiRNN [27]	79.2 ± 2.5	75.2 ± 3.5	62.5 ± 5.5
ACoLP [35]	90.2 ± 2.6	87.4 ± 1.4	76.8 ± 2.3
ASSIGN [23]	88.0 ± 1.8	84.8 ± 3.0	73.8 ± 5.8
2G-GCN [27]	89.5 ± 1.6	87.1 ± 1.8	76.2 ± 2.8
Ours	$\mathbf{90.7 \pm 2.9}(\uparrow \mathbf{0.5})$	$\mathbf{88.1 \pm 2.8}(\uparrow \mathbf{0.7})$	$\mathbf{77.6 \pm 4.7}(\uparrow \mathbf{0.8})$
Ours (ΔCF)	81.1 ± 4.0	77.0 ± 4.8	65.2 ± 5.6
Ours (ΔISR)	88.5 ± 3.7	85.5 ± 3.6	73.9 ± 5.7
Ours ($\Delta\mathcal{L}_{fea}$)	89.3 ± 1.9	85.6 ± 2.1	75.9 ± 4.4
Ours ($\lambda = 0$)	72.0 ± 4.4	65.0 ± 6.9	48.6 ± 6.3
Ours (ΔCLIP+BLIP)	89.4 ± 2.3	85.5 ± 3.9	74.9 ± 5.7
Ours (Traditional Classifier)	79.5 ± 11.0	73.9 ± 11.4	56.6 ± 12.5

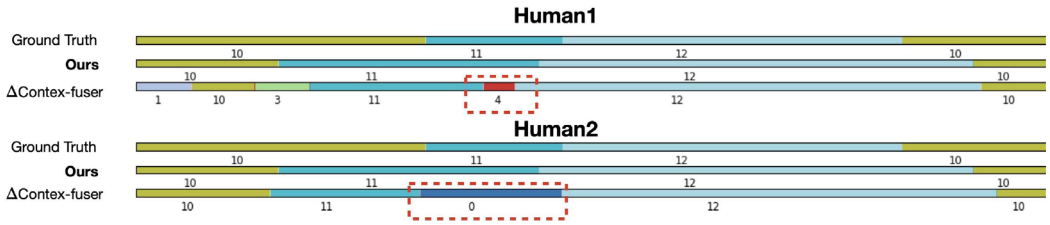

Fig. 5. Qualitative ablation study results on MPHOI-72 dataset. Major prediction errors are highlighted in red dashed boxes.

5.4 Ablation Studies

In this section, we ablate the CF module and the ISR module for validating the effectiveness of these proposed components. As shown in Table 1, removing CF module results in more than 20% F_1 score drop of the three metrics in MPHOI-72 dataset, indicating the essential improvement of CF module in learning context-rich representations. Visualization results of temporal segmentation of interactions in Fig. 5 indicates that removing CF module results in incorrect interaction predictions in both humans (highlighted in red in Fig. 5). Similarly, in CAD-120 and Bimanual Actions datsets, deleting CF module also results in massive F_1 score drop. The visualization results in Fig. 6 suggests an incorrect prediction segment when removing CF module.

Furthermore, we ablate hypersphere by replacing it with Euclidean distance, where $\lambda = 0$ in Eq. 6. Results in Table 1, 2 and 3 indicate that utilizing Euclidean distance results in at least 7%, 10% and 3% F_1 scores drop in MPHOI, CAD-

Table 3. The results of joined segmentation and label recognition of sub-activity on Bimanual Actions. The notations are the same with Table 1.

Model	F_1@10	F_1@25	F_1@50
Dreher *et al.* [4]	40.6 ± 7.2	34.8 ± 7.1	22.2 ± 5.7
Independent BiRNN [27]	74.7 ± 7.0	72.0 ± 7.0	61.8 ± 7.3
Relational BiRNN [27]	77.7 ± 3.9	75.0 ± 4.2	64.8 ± 5.3
ASSIGN [23]	84.0 ± 2.0	81.2 ± 2.0	68.5 ± 3.3
2G-GCN [27]	$\mathbf{85.0 \pm 2.2}$	82.0 ± 2.6	69.2 ± 3.1
Ours	85.0 ± 2.5	$\mathbf{82.9 \pm 2.9}(\uparrow \mathbf{0.9})$	$\mathbf{74.2 \pm 4.3}(\uparrow \mathbf{5.0})$
Ours (ΔCF)	82.5 ± 5.0	80.5 ± 5.5	71.1 ± 7.0
Ours (ΔISR)	84.1 ± 2.3	81.8 ± 2.8	73.0 ± 3.7
Ours ($\Delta\mathcal{L}_{fea}$)	84.5 ± 4.6	82.0 ± 5.2	71.8 ± 6.9
Ours ($\lambda = 0$)	76.7 ± 5.2	74.3 ± 6.0	65.2 ± 6.3
Ours (ΔCLIP+BLIP)	84.3 ± 1.4	81.8 ± 1.8	73.2 ± 2.7
Ours (Traditional Classifier)	82.0 ± 3.6	79.8 ± 4.1	71.0 ± 5.6

Fig. 6. Qualitative ablation study results on CAD-120 dataset. Major prediction errors are highlighted in red dashed boxes.

120 and Bimanual Actions dataset, respectively. Therefore, we conclude that Euclidean distance do not introduce HOI structure priors, ignoring valuable structure information of HOI for guiding predictions. Subsequently, we ablate CLIP and BLIP models by randomly initialize interaction features and context features. Results in Table 1, 2 and 3 indicate that removing CLIP and BLIP models results in some drop of model performance, but is still on par with or better than SOTA methods. Thus, it is the intricately designed structure of our model that substantiates the substantial enhancement in performance. Finally, we ablate the interaction-centric hypersphere by replacing it with a traditional classifier constructed with multi-layer perceptron (MLP) and Softmax classifier. The outcomes, as presented in Table 1, 2 and 3, reveal a notable decline of over 7% in the F_1 score within the MPHOI dataset when employing the traditional classifier. Likewise, in the CAD-120 dataset, the traditional classifier results in a substantial decrease of more than 11%in the F_1 score. Additionally, within the Bimanual Actions dataset, the traditional classifier induces a decline exceeding 2% in the F_1 score. These findings unanimously underscore the efficacy of modeling HOI manifold structures by the hypersphere module.

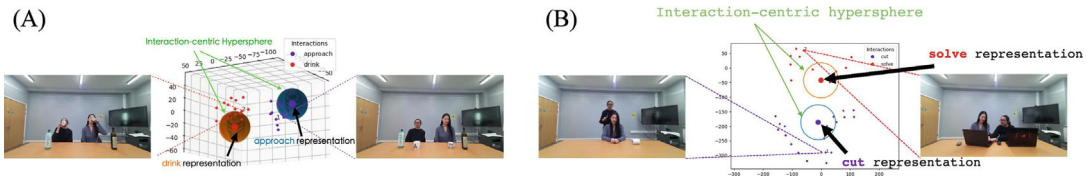

Fig. 7. (A) Visualization of interaction-centric hyperspheres, the learned interaction representations (large points locate at the center of hyperspheres) and human-object entity representations (small points surrounding the hyperspheres) in embedding space. Selected frame samples are shown for each interaction-centric hypersphere. (B) 2D visualization of interaction-centric hyperspheres.

5.5 Qualitative Results

We show some visualization results on MPHOI in Fig. 8 to compare our method with SOTA method 2G-GCN. The red highlighted boxes indicate that 2G-GCN tend to generate unreasonable interaction predictions, while our method generates more reasonable interaction predictions. The visualization results in Fig. 9 show similar prediction pattern where 2G-GCN predicted some unreasonable short segments (highlighted in red boxes) while our method predicts more accurately. We also visualize the interaction-centric hypersphere, the learned interaction representations and the human-object entity representations in embedding space in Fig. 7. Results in Fig. 7(A) show that human-object entity representations (small dots in Fig. 7(A)) belonging to the specific interaction class locates near the surface of the corresponding hypersphere. Similar with Fig. 7(A), we also visualize the intraction-centric hyperspheres in 2D space in Fig. 7(B). Results in Fig. 7(B) indicate that HOIs distribute around the sphere of different interaction classes. The distributions of HOIs on each sphere also differ a lot, where the sphere belongs to the *solve* interaction class captures HOIs on the upper part of its sphere, while the sphere belongs to the *cut* interaction class captures HOIs on the lower part of its sphere. These results suggest that our model successfully model the manifold structure of HOI.

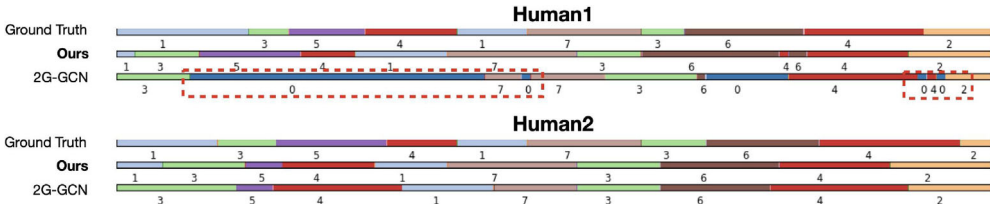

Fig. 8. Visualization results on MPHOI-72 dataset. Major prediction errors are highlighted in red dashed boxes. (Color figure online)

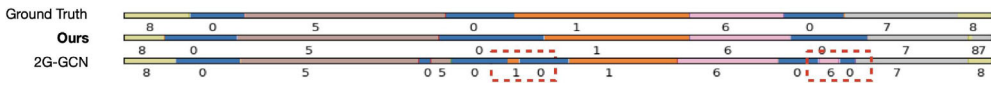

Fig. 9. Visualization results on CAD-120 dataset. Major prediction errors are highlighted in red dashed boxes. (Color figure online)

6 Conclusion

In this work, we propose an interaction-centric spatio-temporal context reasoning network for multi-person video HOI recognition. Specifically, we propose a context fuser and an interaction state reasoner to learn spatio-temporal context-rich and reasoning-aware entity representations. We further represent HOI components with an interaction-centric hypersphere for HOI classification. Experiment results show that our method outperforms SOTA method by more than 22% F_1 score in multi-person scenarios, and achieves competitive results on single-person cases compared to SOTA methods.

Acknowledgements. This material is based upon work supported under the AI Research Institutes program by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education through Award #2229873 - AI Institute for Transforming Education for Children with Speech and Language Processing Challenges. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

References

1. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, pp. 539–546 (2005). https://doi.org/10.1109/CVPR.2005.202
2. Deng, S., et al.: Low-resource extraction with knowledge-aware pairwise prototype learning. Knowl.-Based Syst. **235**, 107584 (2022)
3. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Dreher, C.R.G., Wächter, M., Asfour, T.: Learning object-action relations from bimanual human demonstration using graph networks. IEEE Robot. Autom. Lett. (RA-L) **5**(1), 187–194 (2020). https://doi.org/10.1109/LRA.2019.2949221
5. Gao, C., Xu, J., Zou, Y., Huang, J.-B.: DRG: dual relation graph for human-object interaction detection. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12357, pp. 696–712. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_41
6. Gao, K., Chen, L., Zhang, H., Xiao, J., Sun, Q.: Compositional prompt tuning with motion cues for open-vocabulary video relation detection. In: The Eleventh International Conference on Learning Representations (2022)

7. Gupta, T., Schwing, A., Hoiem, D.: No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9677–9685 (2019)
8. Iftekhar, A., Chen, H., Kundu, K., Li, X., Tighe, J., Modolo, D.: What to look at and where: semantic and spatial refined transformer for detecting human-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5353–5363 (2022)
9. Kim, B., Choi, T., Kang, J., Kim, H.J.: UnionDet: union-level detector towards real-time human-object interaction detection. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12360, pp. 498–514. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58555-6_30
10. Kim, B., Lee, J., Kang, J., Kim, E.S., Kim, H.J.: HOTR: end-to-end human-object interaction detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 74–83, June 2021
11. Koppula, H.S., Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. IEEE Trans. Pattern Anal. Mach. Intell. **38**(1), 14–29 (2015)
12. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from RGB-D videos. Int. J. Robot. Res. **32**(8), 951–970 (2013)
13. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from RGB-D videos. Int. J. Rob. Res. **32**(8), 951–970 (2013). https://doi.org/10.1177/0278364913478446
14. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 156–165 (2017)
15. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, pp. 12888–12900. PMLR (2022)
16. Li, Y.L., et al.: Transferable interactiveness knowledge for human-object interaction detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3585–3594 (2019)
17. Liao, Y., Liu, S., Wang, F., Chen, Y., Qian, C., Feng, J.: PPDM: parallel point detection and matching for real-time human-object interaction detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 482–490 (2020)
18. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
19. Liu, X., Li, Y.L., Wu, X., Tai, Y.W., Lu, C., Tang, C.K.: Interactiveness field in human-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20113–20122 (2022)
20. Luan, T., Wang, Y., Zhang, J., Wang, Z., Zhou, Z., Qiao, Y.: PC-HMR: pose calibration for 3D human mesh recovery from 2D images/videos. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2269–2276 (2021)
21. Luan, T., et al.: High fidelity 3D hand shape reconstruction via scalable graph frequency decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16795–16804 (2023)
22. Mettes, P., Van der Pol, E., Snoek, C.: Hyperspherical prototype networks. Adv. Neural Inf. Process. Syst. **32** (2019)

23. Morais, R., Le, V., Venkatesh, S., Tran, T.: Learning asynchronous and sparse human-object interaction in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16041–16050 (2021)
24. Nagarajan, T., Feichtenhofer, C., Grauman, K.: Grounded human-object interaction hotspots from video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8688–8697 (2019)
25. Park, J., Park, J.W., Lee, J.S.: ViPLO: vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17152–17162, June 2023
26. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.-C.: Learning human-object interactions by graph parsing neural networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11213, pp. 407–423. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01240-3_25
27. Qiao, T., Men, Q., Li, F.W., Kubotani, Y., Morishima, S., Shum, H.P.: Geometric features informed multi-person human-object interaction recognition in videos. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13664, pp. 474–491. Springer, Cham (2022)
28. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
29. Sener, O., Saxena, A.: rCRF: recursive belief estimation over CRFs in RGB-D activity videos. In: Robotics: Science and systems (2015)
30. Sunkesula, S.P.R., Dabral, R., Ramakrishnan, G.: Lighten: learning interactions with graph and hierarchical temporal networks for hoi in videos. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 691–699 (2020)
31. Tamura, M., Ohashi, H., Yoshinaga, T.: QPIC: query-based pairwise human-object interaction detection with image-wide contextual information. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10410–10419, June 2021
32. Tu, D., Sun, W., Min, X., Zhai, G., Shen, W.: Video-based human-object interaction detection from tubelet tokens. Adv. Neural. Inf. Process. Syst. **35**, 23345–23357 (2022)
33. Wang, T., et al.: Deep contextual attention for human-object interaction detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5694–5702 (2019)
34. Xi, N., Meng, J., Yuan, J.: Chain-of-look prompting for verb-centric surgical triplet recognition in endoscopic videos. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 5007–5016 (2023)
35. Xi, N., Meng, J., Yuan, J.: Open set video HOI detection from action-centric chain-of-look prompting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3079–3089 (2023)
36. Zhang, Y., Pan, Y., Yao, T., Huang, R., Mei, T., Chen, C.W.: Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19548–19557 (2022)
37. Zhong, X., Qu, X., Ding, C., Tao, D.: Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13234–13243 (2021)