Bi-Causal: Group Activity Recognition via Bidirectional Causality

Youliang Zhang[1,2] Wenxuan Liu[3] Danni Xu[4] Zhuo Zhou[1,2] Zheng Wang[1,2†]
[1]National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, School of Computer Science, Wuhan University [2]Hubei Key Laboratory of Multimedia and Network Communication Engineering
[3]Wuhan University of Technology [4]National University of Singapore

Abstract

Current approaches in Group Activity Recognition (GAR) predominantly emphasize Human Relations (HRs) while often neglecting the impact of Human-Object Interactions (HOIs). This study prioritizes the consideration of both HRs and HOIs, emphasizing their interdependence. Notably, employing Granger Causality Tests reveals the presence of bidirectional causality between HRs and HOIs. Leveraging this insight, we propose a Bidirectional-Causal GAR network. This network establishes a causality communication channel while modeling relations and interactions, enabling reciprocal enhancement between human-object interactions and human relations, ensuring their mutual consistency. Additionally, an Interaction Module is devised to effectively capture the dynamic nature of human-object interactions. Comprehensive experiments conducted on two publicly available datasets showcase the superiority of our proposed method over state-of-the-art approaches. 0ur project page: https://angzong.github.io/bi-causal.github.io/

1. Introduction

Group Activity Recognition (GAR) refers to determining the activities in scenes containing multiple people [9, 20, 32, 36, 40, 53]. The applications of this field include intelligent monitoring, security, and the analysis of team collaborations. This study specifically centers on scenes depicting team collaboration integrated with interactive objects.

Recently, many methods [5, 12, 17, 42] employed attention mechanisms or graph neural networks to model human relations (HRs), which are considered as key information in group activity. However, while these methods have shown progress, the exploration of Human-Object Interactions (HOIs) in group-object scenarios remains largely unexplored. In contrast, studies in other domains of human behavior understanding and computer vision [27, 54, 55] indicate that HOIs can reveal the intentions of individuals and enrich the semantic understanding of human action. Furthermore, Figure 1 (a) shows HRs and HOIs have different salient regions. Understanding the physical movement of objects helps in analyzing key individuals in GAR. These studies and observations motivate us to combine HOIs and HRs in GAR, and it naturally raises a question:

What is the relationship between HRs and HOIs in understanding team behavior?

Based on our observations, we posit bidirectional causality between HRs and HOIs, categorized under predictive causality [10]. Bidirectional causality suggests that HRs and HOIs mutually forecast each other in both directions, implying that a change in one may anticipate the other. Figure 1 (b) illustrates instances of bidirectional causality. Human Relations during preparation indicate the potential and manner in which the ball can subsequently be interacted

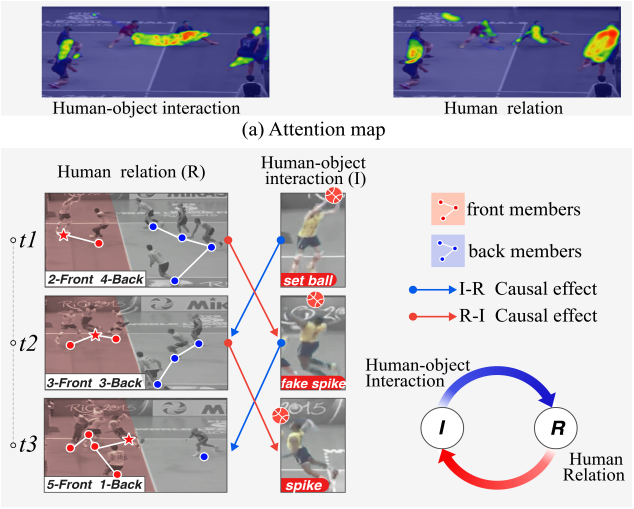


Figure 1. **The illustration for the relationship between HOIs and HRs**. (a) Attention Maps: HOI-based GAR *vs.* HR-based GAR. (b) An example of the bidirectional causality. At $t1$, a front actor's ball set implies the right team will adjust to a 3-front-3 formation at $t2$. The prior 2-front-4 setup also supports a fake spike at $t2$. The fake spike indicates a swift transition to a 5-front-1 formation. The 3-front-3 setup aids a successful spike at $t3$.

†Corresponding Author

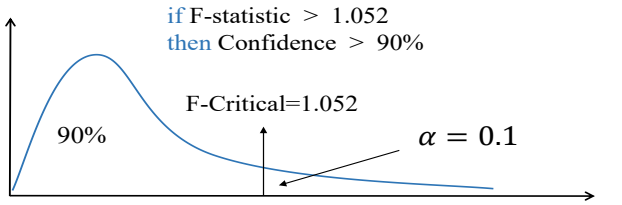
with by humans. Simultaneously, Human-Object Interactions also imply adaptations in human relations. *HRs and HOIs mutually forecast each other and form bidirectional causality*. Methods [43, 47] also adopt a causal perspective to elucidate the causal relationships among humans. These approaches exhibit similarity in their utilization of temporal features contextualized with human activities. However, they neglect to explore the dynamic human-object interactions, leading to a biased focus on irrelevant individuals.

To verify our hypothesis, we perform Granger Causality Tests [14] on the VOLLEYBALL dataset in Section 1.1. The results strongly support the existence of bidirectional causality. Building upon bidirectional causal graphs, we propose a novel framework for Group Activity Recognition that jointly models HRs, HOIs, and their correlation to obtain comprehensive representations of group activities. Our framework, named Bi-Causal, primarily comprises two key components. First, alongside the conventional human relation module (RM), we introduce an Interaction Module (IM) designed to capture HOIs by taking human and object features as inputs and leveraging Graph Convolutional Networks (GCN) to facilitate this process. Second, to harness the bidirectional causality, we establish a Causality Communication Channel to exchange information while modeling HRs and HOIs. We also incorporate the Kullback-Leibler (KL) divergence function between HRs and HOIs to enforce information consistency in the final outcome. Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to introduce the human-object interaction as a causal factor to the GAR task. By modeling HOIs, our IM accurately assists in capturing the region in which behavior occurs.
- We highlight the importance of exploring bidirectional causality in GAR, offering insights for group activity related research. Specifically, we present a novel bidirectional framework, Bi-Causal, simultaneously modeling HRs and HOIs and mutually enhancing each other.
- To showcase the strength of our model, we conduct experiments on widely adopted VOLLEYBALL and COLLECTIVE ACTIVITY datasets. The results demonstrate that our method achieves state-of-the-art performance.

1.1. Motivation—Causality test

Causality testing aims to describe the causal relationship between two entities [51]. The Granger theory involves autoregressive modeling and correlation regression modeling of time series data to analyze regression errors [14, 31]. If introducing environment features results in a reduction of the regression error, it implies a causal relationship between the features and the subject. To enhance the reliability of our hypothesis, we employ Granger Causality Tests to investigate the presence and strength of the bidirectional causality. We build two models: model $\mathcal{R}$ for HRs and model $\mathcal{I} \rightarrow \mathcal{R}$ combining both HRs and HOIs. According to the Granger causality test, if the prediction of $\mathcal{I} \rightarrow \mathcal{R}$ outperforms that of $\mathcal{R}$, we can infer that HOIs causally influence HRs. We utilize the F-statistic $F_{\mathcal{I}\rightarrow\mathcal{R}}$ to quantify the comparison between models, where a higher F-statistic indicates that the $\mathcal{I} \rightarrow \mathcal{R}$ model outperforms the $\mathcal{R}$ model. In the same way, this approach can also be employed to ascertain the causal impact of HRs on HOIs ($\mathcal{R} \rightarrow \mathcal{I}$). The detailed method can be found in the supplementary materials.



F-statistics ↑	T_s=5	T_s=6	T_s=7	T_s=8	T_s=9	T_s=10
$F_{\mathcal{R}\rightarrow\mathcal{I}}$	1.055	1.070	**1.092**	1.069	1.045	1.060
$F_{\mathcal{I}\rightarrow\mathcal{R}}$	1.067	**1.087**	1.069	1.043	1.052	1.052
α	0.150	0.100	0.050	0.020	0.010	0.005
F-Critical	1.041	1.052	1.066	1.084	1.095	1.106
Confidence	85%	90%	95%	98%	99%	99.5%

Figure 2. **The confidence in the bidirectional causality.** F-Critical shows the F-Critical Value at Significance Level α. Both $F_{\mathcal{R}\rightarrow\mathcal{I}}$ and $F_{\mathcal{I}\rightarrow\mathcal{R}}$ are F-statistics that reflect the confidence in causality. When the F-statistic is greater than the F-Critical Value, we consider it as having a corresponding confidence level that causality exists. T_s denotes the sequence window for inference.

Figure 2 shows the values of the two F-statistics for different values of the sequence window T_s. F-statistics indicate the confidence level of the bidirectional causality, *i.e.*, $F_{\mathcal{R}\rightarrow\mathcal{I}}$ and $F_{\mathcal{I}\rightarrow\mathcal{R}}$. Regarding causality from HRs to HOIs, we have a 98% confidence level that causality exists. When T_s equals 5, it performs less optimally but still has a confidence level of over 85%. For causality from HOIs to HRs, the best and worst results are respectively 98% and 85%. The vary caused by T_s may be attributed to the fact that an appropriate sequence length makes causality more evident, while overly long or too short sequences can introduce interference or information deficiency. The experimental results demonstrate the existence of bidirectional causality between HRs and HOIs. We can leverage this causality for modeling to enhance the construction of HRs and HOIs, leading to improved representations of group activities.

2. Related Work

2.1. Group Activity Recognition

GAR has gained considerable prominence owing to its versatile applications. The techniques have transitioned from early manually engineered features and probabilistic graphical models [1–4, 29] to deep learning-based graph mod-

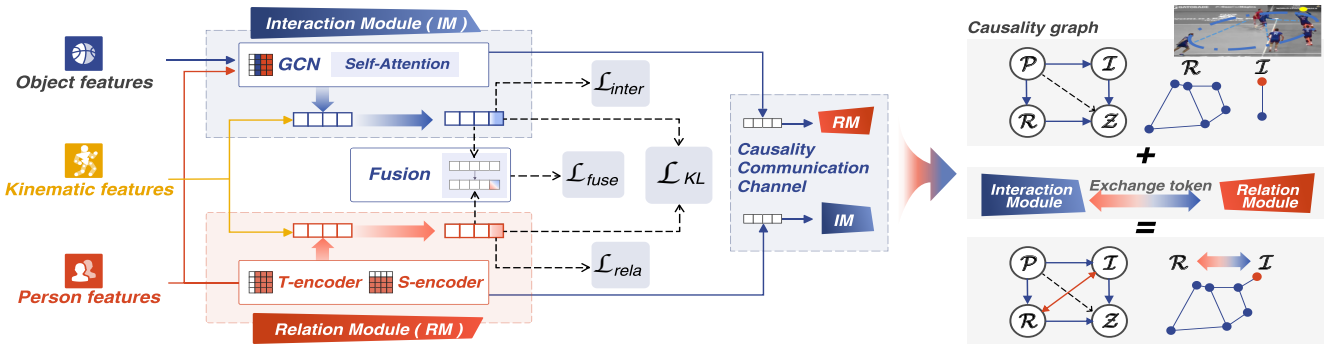


Figure 3. **Illustration of our proposed Bi-Causal.** a) Model structure: Based on preprocessed object features (**O**), person features (**P**), and kinematics features (**K**), Bi-Causal employs two modules RM and IM to extract human relations and human-object interactions from **O** and **P**, integrating them with **K**. In the process, a causality communication channel enables token exchanges between RM and IM. The final representations from RM and IM are merged for the ultimate GAR output. b) Explanation: This structure simulates the bi-causality graph (right) with $\mathcal{P}$ (person features), $\mathcal{Z}$ (final GAR representation), $\mathcal{R}$ (human relations), and $\mathcal{I}$ (human-object interactions). In causality graphs, solid lines indicate direct, dashed lines indicate indirect causal relationships.

els [5, 6, 18, 42, 49]. Recent introductions of attention mechanisms [12, 15, 17, 20, 21, 23, 42] have further improved the adaptability of visual representation.

Among the methods mentioned above, [15, 19, 42, 43, 45, 49] take the RGB modality features with RoIAlign [16] as input, emphasizing the organization of human relations. To enhance the local pose representation for a human and associated relationship, [12, 20, 22, 50] utilize the keypoints modality as auxiliary information. However, these methods overlook the utilization of object information to capture HOIs, leading to a limited comprehension of the overall group activity. Therefore, some methods [24, 35, 53] utilize ball tracklets for GAR, but they mainly examine HOIs based on coordinates or treat objects as additional scene data, rather than prioritizing the object as the central element of the interaction and fully capturing the interaction dynamics. Furthermore, the methods mentioned above neglect the potential bidirectional causality between HRs and HOIs, resulting in the separation of relations and interactions. This separation makes it challenging for HRs and HOIs to perceive and mutually promote each other.

2.2. Causality in Computer Vision

Causality holds promise in the field of computer vision [41, 46] as it contributes to the development of interpretable models. Recently, scholars have shown a growing interest in exploring the causal relationships in GAR [43, 47]. Yuan *et al*. [47] explores the casual graph to incorporate global visual context. Xie *et al*. [43] utilizes Granger causality tests [14, 31] to describe the complex directed causality relationships among individual movements, capturing asynchronous temporal information among actors. In this work, we excavate the **bidirectional causality** that exists in group activities. Through the Granger causality tests, we establish the existence of bidirectional causality using the feature sequences related to HRs and HOIs extracted at the clip level.

3. Bi-Causal Group Activity Recognition

Figure 3 presents the overall framework of Bi-Causal.

3.1. Feature Extractors

Given a video sequence, we perform uniform sampling to select a set of T frames and subsequently extract keypoints from this sequence. We define D as the feature dimension.

Object features are represented by $\mathbf{O} \in \mathbb{R}^{N_o \times T \times D}$, N_o is the number of objects. We calculate information such as movement speed based on the coordinates of the current object. Object annotations are obtained from GIRN [24].

Person features are defined as $\mathbf{P} \in \mathbb{R}^{N_p \times T \times D}$, where N_p denotes the total number of players. For each person, we aggregate all joint information captured by HRnet [33].

Kinematic features are also obtained from joint information, but organized with the inherent grouping information. We define it as $\mathbf{K} \in \mathbb{R}^{N_m \times T \times D}$, where N_m denotes the number of subgroups. We aggregate all joint coordinates from a subgroup to get corresponding kinematic features.

3.2. Interaction Module

Our proposed Interaction Module accepts object features and person features as inputs. Through graph evolution, the IM engenders interaction features that capture the dynamic interactions between objects and human entities.

As illustrated in Figure 4, we employ $\mathbf{V}$ to symbolize the entirety of input node features, with each human and object constituting an individual node. Regarding the node features $\mathbf{v}_t$ at frame t, we execute the dot product of $\mathbf{v}_t$ with its transpose, thereby generating the graph edges $\mathbf{e}_t$, which serves as a reflection of the extent of interaction among distinct nodes. Given that HOIs transpire within video sequences, it becomes imperative to take into account the temporal dynamics inherent in these interactions. Hence, we employ $\mathbf{E} = \{\mathbf{e}_t\}$ to signify the graph edges derived from

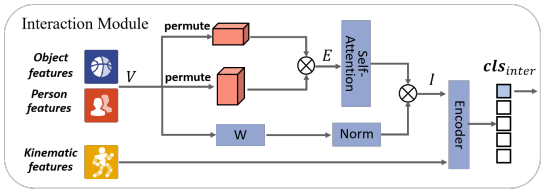


Figure 4. **Illustration of our Interaction Module. V** represents all the node features (object or person). **E** stands for the graph edges. **W** presents learnable weight parameters.

a video clip spanning T frames, thus enabling the capture of temporal interaction dynamics through the application of Self-Attention mechanism to **E**. Consequently, we acquire the interaction graph that encompasses the HOIs, denoted as $\{\mathbf{V}, \mathbf{E}\}$. The interaction graph comprises a set of T graphs, which are then processed by the Self-Attention mechanism to model temporal dynamics across these graphs. The output of IM is derived through the following process:

$$\mathbf{I} = \sigma(\text{Self-Attention}(\mathbf{E}) \cdot \text{Norm}(\mathbf{VW})), \tag{1}$$

where **I** is the human-object interaction features, $\mathbf{W} \in \mathbb{R}^{D\times D}$ is the weight parameters. $\sigma(\cdot)$ represents an activation function, and ReLU [13] is adopted in our method.

Following this, we utilize a Transformer encoder to fuse the kinematic features with HOI features, thereby yielding the classification-oriented token denoted as $\boldsymbol{cls}_{inter}$, akin to the methodology introduced in ViT [11]. The fusion facilitates a more discerning comprehension of inherent group information of human-object interactions. The $\boldsymbol{cls}_{inter}$ is obtained through the minimization of the interaction loss $\mathcal{L}_{inter}$ in the following manner:

$$\mathcal{L}_{inter} = -\sum_{i=1}^{N}\sum_{c=1}^{M} y_{ic}\log(p_{ic}), \tag{2}$$

where p_{ic} designates the predicted probability that sample i pertains to class c, which is determined by classifying $\boldsymbol{cls}_{inter}$. y_{ic} denotes the result of a sign function, assuming the value of 1 when the true class of sample i equals c, and 0 otherwise. N signifies the count of samples within a mini-batch, while M denotes the number of distinct classes.

3.3. Relation Module

We adopt a Transformer-based approach to model human relations, drawing inspiration from [15, 20]. The person features provided as input to the Relation Module encompass both temporal and spatial dimensions.

$$\mathbf{R}_{ST} = \text{T-encoder}(\text{S-encoder}(\mathbf{P})), \tag{3}$$

$$\mathbf{R}_{TS} = \text{S-encoder}(\text{T-encoder}(\mathbf{P})), \tag{4}$$

and

$$\mathbf{R} = \mathbf{R}_{ST} + \mathbf{R}_{TS}, \tag{5}$$

where **R** represents the extracted relation features, and **P** represents person features. The S-encoder and T-encoder correspond to encoders [38] applied to distinct dimensions. $\mathbf{R}_{ST}$ and $\mathbf{R}_{TS}$ are two spatiotemporal modeling patterns for relation evolution by switching the order of space and time.

Similar to the Interaction Module, we likewise fuse kinematic features with relation features to acquire the relation cls-token denoted as $\boldsymbol{cls}_{rela}$ and its associated loss $\mathcal{L}_{rela}$.

3.4. Causality Communication Channel

To leverage the bidirectional causality between human relations and human-object interactions, we introduce exchange tokens. First, we incorporate an empty token into the construction process of both the relation and interaction graphs, treating it as a node. We then extract information from all graph nodes to formulate the relation and interaction exchange tokens, respectively. Subsequently, we employ exchange tokens to reconstruct both the relation and interaction graphs, guided by the prior information conveyed by exchange tokens (e.g., utilizing the relation exchange token to guide the reconstruction of the interaction graph).

To better establish bidirectional causality between human relations and human-object interactions, we implement the KL Divergence, as in mutual learning [52] to facilitate their mutual guidance and enhancement. The KL distance $\mathcal{D}_{KL}$ from $\boldsymbol{cls}_{inter}$ to $\boldsymbol{cls}_{rela}$ is computed as:

$$\hat{\boldsymbol{p}}_{inter} = \exp(\delta(\boldsymbol{cls}_{inter})), \hat{\boldsymbol{p}}_{rela} = \exp(\delta(\boldsymbol{cls}_{rela})), \tag{6}$$

$$\mathcal{D}_{KL}(\hat{\boldsymbol{p}}_{inter}||\hat{\boldsymbol{p}}_{rela}) = \sum_{i=1}^{N} \hat{\boldsymbol{p}}^{i}_{inter}\log(\frac{\hat{\boldsymbol{p}}^{i}_{inter}}{\hat{\boldsymbol{p}}^{i}_{rela}}), \tag{7}$$

where N denotes the number of samples in a mini-batch and δ means a classifier. $\hat{\boldsymbol{p}}_{inter}$ represents the group score. To be noticed, due to the asymmetry of KL Divergence, we compute $\mathcal{D}_{KL}(\hat{\boldsymbol{p}}_{rela}||\hat{\boldsymbol{p}}_{inter})$ as well. The total KL Divergence can be formulated as:

$$\mathcal{L}_{KL} = \mathcal{D}_{KL}(\hat{\boldsymbol{p}}_{rela}||\hat{\boldsymbol{p}}_{inter}) + \mathcal{D}_{KL}(\hat{\boldsymbol{p}}_{inter}||\hat{\boldsymbol{p}}_{rela}). \tag{8}$$

3.5. Feature Fusion and Objective Function

Once the HR features and HOI features are acquired, we proceed to employ a Transformer encoder to fuse them, resulting in the final representation of group activities. This representation is subsequently used during the inference phase. The fusion encoder can be trained using $\mathcal{L}_{fuse}$, akin to the fusion process within the Interaction Module.

In the interaction module, interaction representation is trained by minimizing $\mathcal{L}_{inter}$. The relation module is trained by minimizing $\mathcal{L}_{rela}$. To discern the bidirectional causality between human relations and human-object interactions, we minimize $\mathcal{L}_{KL}$. The fusion of interaction and

relation is achieved by minimizing $\mathcal{L}_{fuse}$. We infer the individual actions by averaging the predictions from the node features of the interaction and relation graphs. The total objective function is constituted as a weighted summation of all the aforementioned losses:

$$\mathcal{L}_{group} = \lambda_r \mathcal{L}_{rela} + \lambda_i \mathcal{L}_{inter} + \lambda_f \mathcal{L}_{fuse}, \quad (9)$$

$$\mathcal{L}_{total} = \mathcal{L}_{group} + \mathcal{L}_{person} + \lambda_{KL} \mathcal{L}_{KL}, \quad (10)$$

where $\mathcal{L}_{group}$ represents the loss from GAR, $\mathcal{L}_{person}$ is from the individual actions recognition. λ_r, λ_i, λ_f and λ_{KL} are hyper-parameters that govern the relative significance of each loss. $\mathcal{L}_{total}$ is employed to train our framework.

4. Experimental Results and Analysis

4.1. Datasets and metric

VOLLEYBALL dataset [18] consists of 55 volleyball videos, comprising a total of 4,830 labeled clips, with 3,493 clips in the training set and 1,337 clips in the testing set. The individual action labels encompass 9 distinct actions, while the group activity labels encompass 8 activities.

RE-ANNOTATED VOLLEYBALL dataset [8] was created through the process of re-annotating a subset of group labels within the VOLLEYBALL dataset that were deemed inappropriate. A total of 497 reannotations were conducted, representing approximately 10% of the entire dataset. After excluding 9 video clips due to changes in camera angles, the refined dataset now comprises 4,821 clips.

COLLECTIVE ACTIVITY dataset [7] comprises 44 videos with comprehensive annotations. It includes 5 group activity labels. For our training and testing, we follow the same split as in previous studies [26, 49], using 32 videos for training and 12 videos for testing.

In our evaluation, we use the metric of group activity accuracy, in alignment with the method adopted by [20, 53]. We also validated our method within the NBA dataset [45], which is commonly utilized in weakly supervised group activity recognition. Please refer to the supplementary materials for the experimental results of NBA dataset.

4.2. Implementation details

The feature dimension D is set as 256. For consistent evaluation, we follow a standardized approach consistent with prior research [12, 42, 48, 49, 53], employing input size of $T = 10$ frames for both training and testing. The dimension of the Feed-forward Network layer in all Transformer encoders is 1024, with ReLU activation functions. Annotations of objects in the VOLLEYBALL dataset are obtained from [24]. During the training phase, we utilized the Adam optimizer with a learning rate of 0.001 and a batch size of 128. Our network is implemented using PyTorch and trained for 80 epochs on a single NVIDIA Tesla V100 GPU. Further details are available in the supplementary material.

Model	Keypoint	RGB	Flow	Backbone	VD ↑	CAD ↑
CERN [30]		✓		VGG-16	83.3	87.2
stagNet [26]		✓		VGG-16	89.3	89.1
HRN [18]		✓		VGG-19	89.5	–
SSU [6]		✓		Inception-v3	90.6	–
HiGCIN [44]		✓		ResNet-18	91.5	93.4
ARG [42]		✓		Inception-v3	92.5	91.0
CRM [5]		✓		I3D	93.0	–
DIN [49]		✓		VGG-16	93.6	–
DECOMPL [8]		✓		VGG-16	93.8	95.5
GroupFormer [20]		✓		Inception-v3	94.1	93.6
Dual [15]		✓		Inception-v3	94.4	–
ACCG [43]		✓		VGG-16	95.5	95.0
Tamura *et al.* [34]		✓		I3D	**96.0**	**96.5**
Dual [15]		✓	✓	Inception-v3	95.5	–
GIRN [24]	✓	✓	✓	I3D+OpenPose	94.0	95.2
GroupFormer [20]	✓	✓	✓	I3D+AlphaPose	**95.7**	**96.3**
SACRF [25]	✓	✓	✓	I3D+AlphaPose	95.0	95.2
AT [12]	✓	✓		I3D+HRNet	93.5	91
GIRN [24]	✓			OpenPose	92.2	–
AT [12]	✓			HRNet	92.3	–
POGARS [35]	✓			Hourglass	93.9	–
COMPOSER [53]	✓			HRNet	94.6	94.1†
Ours	✓			HRNet	**96.1**	**94.7**

Table 1. **Comparison with SOTA on the VOLLEYBALL dataset (VD) and the COLLECTIVE ACTIVITY dataset (CAD) for group activity accuracy.** Keypoint, RGB, and flow (optical flow) are widely used information modalities, and ✓ in the table means they are used in the corresponding model. † indicates that this data is from our replication results. ↑ represents a higher value is better.

4.3. Comparison with the State-of-the-Art

We conduct a comparative analysis between our Bi-Causal and SOTA methods, on the VOLLEYBALL dataset and the COLLECTIVE ACTIVITY dataset.

VOLLEYBALL dataset. The comparison results are presented in Table 1. Our method demonstrates superior performance when compared to previous approaches, both using the same keypoint inputs as ours and those relying on different data sources. Notably, we achieve SOTA performance even when many methods incorporate additional information such as optical flow. Among the RGB-only methods, compared to recent ACCG [43], which does not use object features, our method utilizes HOIs to have a more comprehensive understanding of group activities, allowing for accuracy improvements of 0.6%. Without the constraint of bounding boxes, Tamura *et al*. [34] uses a detection-based method to identify and aggregate features to perceive scene contexts. Though this method reduced information loss, the lack of HOIs still makes [34] 0.1% lower than ours.

Compared to the current state-of-the-art only using the keypoint modality, our Bi-Causal achieves considerable performance. In contrast to COMPOSER [53] with multiscale representations, our method models HOIs instead of treating objects as auxiliary features, resulting in a notable improvement of 1.5%. Furthermore, by exploring HOIs at a higher level and fully leveraging the bidirectional causality between HRs and HOIs, our method achieves a significant improvement of 3.9% compared to GIRN [24].

Model	RGB	Keypoint	Accuracy ↑
SACRF [25]	✓		92.8
DIN [49]	✓		94.3
GroupFormer [20]	✓		94.4
DECOMPL [28]	✓		95.2
COMPOSER [53]		✓	96.2
Ours		✓	**96.8**

Table 2. **Comparison with SOTA on the RE-ANNOTATED VOLLEYBALL dataset.**

Method	Temporal	Multi-graph	Accuracy ↑
GAT [39]		✓	95.4
ARG [42]		✓	95.5
GIRN [24]	✓		95.8
IM	✓	✓	96.1

Table 3. **Comparison of different Interaction Module.** Temporal means whether the temporal dynamics of HOIs are considered. Multi-graph means whether the module contains multiple graphs.

Base model		Causality		Accuracy ↑
RM	IM	$\mathcal{R} \rightarrow \mathcal{I}$	$\mathcal{I} \rightarrow \mathcal{R}$	
	✓			93.1
	✓	✓		94.7
✓	✓			94.3
✓				94.0
✓			✓	95.4
✓	✓	✓	✓	**96.1**

Table 4. **Effect of Causality.** RM and IM denote using only the relation module or the interaction module for GAR. $\mathcal{R} \rightarrow \mathcal{I}$ means causality from human relations to human-object interactions is considered and $\mathcal{I} \rightarrow \mathcal{R}$ means the opposite.

For the multi-modal methods, our method outperforms all other methods, despite these approaches utilizing additional information. This achievement can be attributed to the full exploitation of bidirectional causality between HOIs and HRs, enabling them to mutually promote each other and resulting in a comprehensive perception of various activities. In contrast, methods such as GroupFormer [20], AT [12], and Dual [15], which primarily model HRs, face challenges when distinguishing activities with comparable HRs, despite the inclusion of additional information.

COLLECTIVE ACTIVITY dataset. The COLLECTIVE ACTIVITY dataset primarily consists of simpler actions and human relations, compared to the VOLLEYBALL dataset. However, it does not include fixed interactive objects. Therefore, when working with the COLLECTIVE dataset, we employ a strategy to detect the most salient object in the scene that is closest to individuals. We then consider the interactions between this object and these individuals. The comparison results are detailed in Table 1, and our framework consistently delivers a commendable performance. When compared to keypoints-only approaches, our method continues to yield strong results. Due to the lack of crucial scene information [34], keypoint-based methods often exhibit comparatively modest performance compared to RGB-based methods. Despite this, compared to model AT, which utilizes both keypoints and RGB information, our method exhibits a significant improvement of 3.7% due to the incorporation of bidirectional causality between HRs and HOIs. Compared to GIRN [24] with multi-modal information, we achieve a significant improvement on the VOLLEYBALL dataset, while the difference on the COLLECTIVE dataset is relatively modest. This indicates that our causal framework yields better results in scenarios where there is substantial interaction between people and objects. The results on the COLLECTIVE ACTIVITY dataset reaffirm the highlights and generalizability of our proposed method.

RE-ANNOTATED VOLLEYBALL dataset. The results of our comparison with other SOTA methods are presented in Table 2. Our method surpasses all other approaches on this dataset, achieving the highest performance. In contrast to Groupformer, which excels in the RGB modality, our method achieves a notable improvement of 2.4%. Similarly, we observe a performance gain of 0.6% compared to COMPOSER. These results further highlight the discriminative capability of our approach in the field of GAR.

Interaction modeling comparison. To demonstrate the effectiveness of our IM, we adapted some graph-based methods to model HOIs and conducted experiments by replacing the IM while keeping all other experimental conditions unchanged. 1) We employed either human or object features as graph nodes and calculated edges using an attention mechanism to implement the GAT [39] method. 2) By keeping the input at the joint level, we reproduced the Person-Object module of GIRN [24] to extract HOIs features. 3) We implemented ARG [42] to construct multiple graphs within a same frame to model HOIs. These methods have been proven to be highly effective in modeling HRs. As shown in Table 3, our IM performs favorably against all the other methods. The reason for the superior performance of our IM compared to ARG and GAT is that these methods construct interaction graphs but do not consider the temporal dynamics of HOIs. In contrast, we construct T graphs for each frame and employ a transformer encoder to incorporate the edges of T graphs with temporal dynamics. In contrast to GIRN, our method considers humans and objects as interaction subjects rather than focusing solely on joints. This broader perspective allows us to capture interactions beyond mere contact and shared motion, resulting in a more comprehensive representation of HOIs.

4.4. Ablation Studies

In this subsection, we conduct ablation studies in terms of group activity accuracy on the VOLLEYBALL dataset to investigate the contribution of each component in our model.

Effect of Causality. To validate the beneficial effect of bidirectional causality on human relations and human-

(a) Effectiveness of interaction module

Manner	Accuracy ↑
erase	91.9
w/o ball	92.7
w/o self-attention	93.6
IM	**95.3**

(b) Effectiveness of relation module.

S-encoder	T-encoder	Path	Accuracy ↑
✓			93.6
	✓		94.1
✓	✓	S-T	94.5
✓	✓	T-S	94.6
✓	✓	Dual	**95.3**

(c) Effectiveness of feature fusion method

Manner	Accuracy ↑
sum	94.0
concat	94.3
w/o kinematic	94.7
Ours	**95.8**

Table 5. **Experiment results of ablation studies on the VOLLEYBALL dataset.**

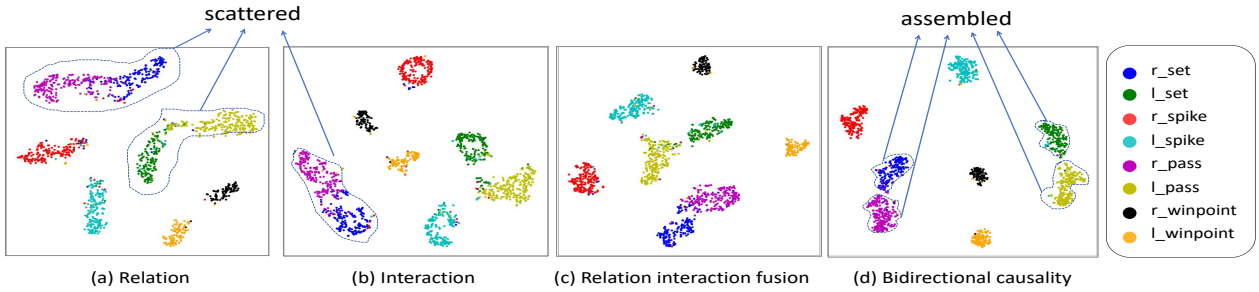


Figure 5. **t-SNE visualization of activity representation on the VOLLEYBALL dataset.** Learned from different models: model based on human relations, model based on human-object interactions, relation and interaction fusion model, and our bidirectional causality method.

object interactions, we perform a series of experiments. 1) **RM** and 2) **IM** denote using only human relations or human-object interactions to describe activities, respectively. In 3) **IM** ***w/*** $\mathcal{R} \rightarrow \mathcal{I}$, causality from HRs to HOIs is employed to optimize the information related to human-object interactions. In 4) **RM** ***w/*** $\mathcal{I} \rightarrow \mathcal{R}$, causality from HOIs to HRs is employed to optimize human relations. 5) '**RM** and **IM**' incorporates both HRs and HOIs for GAR without utilizing causality. The corresponding experimental results are shown in Table 4. Optimizing HRs and HOIs based on unidirectional causality (settings **RM** ***w/*** $\mathcal{I} \rightarrow \mathcal{R}$ and **IM** ***w/*** $\mathcal{R} \rightarrow \mathcal{I}$) results in a 1.4% and 1.6% improvement, respectively, compared to using HRs and HOIs alone, which signifies the impact of causality. Besides, the 0.4% improvement observed when transitioning from 5) '**RM** and **IM**' to 3) **IM** ***w/*** $\mathcal{R} \rightarrow \mathcal{I}$ indicates that, compared to solely introducing HRs information, the addition of causality from HRs to HOIs results in a more effective model. Overall, the inclusion of bidirectional causality results in significant improvements of 2.1% and 3% compared to using HRs and HOIs alone, emphasizing the crucial role it plays in GAR.

Variations of Interaction. To assess the efficacy of IM, we investigate its effectiveness across four distinct settings. 1) The **erase** setting replaces the GCN used in our IM with a feed-forward network. 2) ***w/o* ball** setting modifies our interaction module to exclude the ball information. 3) ***w/o* self-attention** setting eliminates the Transformer encoder previously employed along the edges of our graph for T frames. All the other aspects of these variants remain constant, and the results are presented in Table 5 (a). In contrast to the **erase** method, our approach demonstrates a notable 3.4% increase, providing compelling evidence for the crucial role played by our Interaction Module. Compared to the full method, the ***w/o* ball** variant shows a 2.6% decrease in accuracy, underscoring the pivotal role of objects for accurately capturing HOIs. In the ***w/o* self-attention** setting, there is a 1.7% decrease compared to the full method. This highlights the capability of our encoder to integrate temporal information into HOI features. The above experiments on interaction variations indicate that GAR performance benefits from our dynamically modeled HOIs.

Variations of Relation. Table 5 (b) presents four different settings to examine the effect of RM. The S-encoder and T-encoder are responsible for modeling HRs in the spatial and temporal dimensions, respectively. In the S-T setting, HRs are processed in a spatial-temporal order, while in the T-S setting, a temporal-spatial order is employed. The Dual setting combines the results of the S-T and T-S paths using Equation 5. The result shows that the Dual setting achieves the best result, with a 0.8% increase against S-T and a 0.7% increase against T-S. Additionally, Dual achieves 1.7% and 1.2% point gain compared to solely utilizing S-encoder and T-encoder respectively. These findings highlight the importance of considering multiple dimensions in modeling HRs, and using spatial and temporal information in different orders facilitates a more comprehensive characterization of HRs. The experimental findings reaffirmed the role of the dual paths postulated by Han *et al*. [15].

Integration Effect. In our IM and RM, we fuse kinematic features with interaction features and relation features, respectively. We then fuse the interaction features and relation features to derive the final activity representation. Ablation experiments on our fusion method are conducted and presented in Table 5 (c). 1) After projecting features into the same dimension, **sum** simply adds the features to be fused. 2) **concat** directly concatenates the fea-

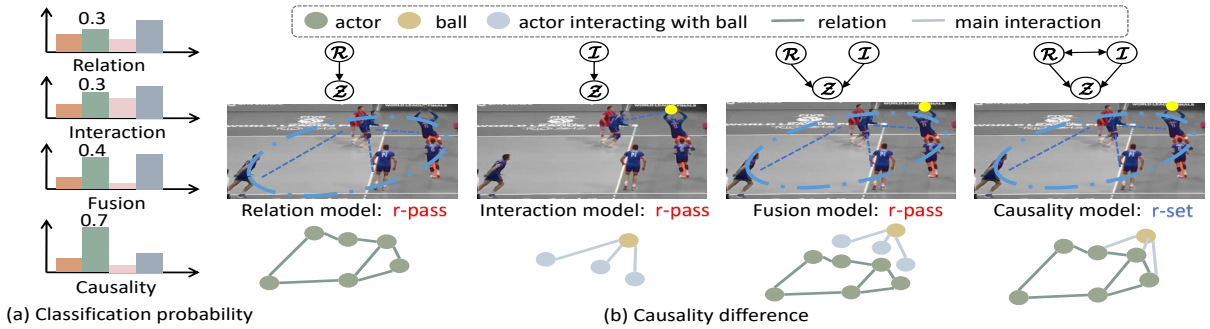


Figure 6. **Visualization of the distinctions between our proposed bidirectional causality model and other models.** In causality graphs (above), $\mathcal{Z}$ represents the final representation for GAR. $\mathcal{R}$ and $\mathcal{I}$ denote HRs and HOIs, respectively. (a) presents the classification results of various models on an r-set data sample and the classification possibilities of the correct class are annotated in the figure (green).

$\mathcal{L}_{KL}$	$\mathcal{L}_{inter}$	$\mathcal{L}_{rela}$	$\mathcal{L}_{person}$	$\mathcal{L}_{fuse}$	Accuracy ↑
				✓	91.6
			✓	✓	93.6
✓			✓	✓	94.7
	✓		✓	✓	94.8
		✓	✓	✓	94.9
✓	✓	✓	✓	✓	**95.8**

Table 6. **Effectiveness of our multiple loss function.**

tures to be fused. 3) ***w/o* kinematic** represents not using the kinematic features to optimize the interaction and relation features. Table 5 (c) demonstrates that compared to the **sum** and **concat** methods, the attention mechanism dynamically fuses HR features and HOI features, resulting in performance improvements of 1.8% and 1.5% respectively. The omission of kinematic features leads to the absence of raw motion information and grouping information, resulting in a slight decline compared to our fusion strategy.

Effect of Multiple Loss Functions We assess the impact of different components of our loss function on the performance of our network. Table 6 shows that utilizing multiple losses consistently outperforms using only the final fuse loss function. This improvement stems from the introduction of constraints of different modules in a multi-module network, which enhances the consistency of our model and leads to better performance. By leveraging all components of our loss, our network achieves the best results.

4.5. Visualization

Group Representation Visualization. Figure 5 displays the t-SNE [37] visualization of the extracted activity representations. These representations are high-dimensional features obtained from the test set of the VOLLEYBALL dataset. We utilize t-SNE to project the group activity representations onto a two-dimensional plane. The results depicted in the figure show that describing activities using both HRs and HOIs concurrently yields superior performance compared to using either HRs or HOIs in isolation. Moreover, exploring bidirectional causality between HRs and HOIs further enhances the representation of group activities. The visualization supports the effectiveness of our framework.

Causality distinction of different models. As illustrated in Figure 6, when viewed from a causal perspective, the Relation Model and Interaction Model describe group activities from only one aspect (HRs or HOIs), providing an incomplete understanding of group activities. The Fusion model combines features from both $\mathcal{R}$ and $\mathcal{I}$ but does not consider the potential mutual influence between HRs and HOIs, leading to a separation between them. In contrast, our approach establishes bidirectional causality between $\mathcal{R}$ and $\mathcal{I}$, considering their mutual impact and providing comprehensive support for GAR. Hence, the first three models incorrectly classify the given data sample as "right pass", while our Bi-Causal correctly identifies it as "right set".

5. Conclusion

This paper introduces human-object interaction as a causal factor in Group Activity Recognition. We establish a bidirectional causality relationship between HRs and HOIs through empirical evidence. We present Bi-Causal, a novel framework that concurrently models HRs and HOIs, fostering mutual enhancement via a Causality Communication Channel. Our Interaction Module dynamically captures spatial and temporal interactions between objects and humans through graph evolution. Our comprehensive experiments not only showcase the performance of our model but also highlight the contributions and impacts of each module. We believe our work provides valuable insights for future research in this field. However, there are limitations to consider, such as improving bi-causality representation and expanding our approach to other GAR scenarios.

Acknowledgements. This work was supported by National Natural Science Foundation of China (62171325) and Hubei Key R&D Project (2022BAA033). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] Mohamed R Amer and Sinisa Todorovic. Sum product networks for activity recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(4):800–813, 2015.

[2] Mohamed R Amer, Dan Xie, Mingtian Zhao, Sinisa Todorovic, and Song-Chun Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *Proc. Eur. Conf. Comput. Vis.*, 2012.

[3] Mohamed R Amer, Sinisa Todorovic, Alan Fern, and Song-Chun Zhu. Monte carlo tree search for scheduling activity recognition. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1353–1360, 2013.

[4] Mohamed Rabie Amer, Peng Lei, and Sinisa Todorovic. Hirf: Hierarchical random field for collective activity recognition in videos. In *Proc. Eur. Conf. Comput. Vis.*, pages 572–585, 2014.

[5] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional relational machine for group activity recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7892–7901, 2019.

[6] Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4315–4324, 2017.

[7] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *Proc. Int. Conf. Comput. Vis. Workshops*, pages 1282–1289, 2009.

[8] Berker Demirel and Huseyin Ozkan. Decompl: Decompositional learning with attention pooling for group activity recognition from a single volleyball image. *arXiv:2303.06439*, 2023.

[9] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4772–4781, 2016.

[10] Francis X Diebold. *Elements of forecasting*. Citeseer, 1998.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.

[12] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 839–848, 2020.

[13] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *International conference on artificial intelligence and statistics*, pages 315–323, 2011.

[14] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[15] Mingfei Han, David Junhao Zhang, Yali Wang, Rui Yan, Lina Yao, Xiaojun Chang, and Yu Qiao. Dual-ai: Dual-path actor interaction learning for group activity recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2990–2999, 2022.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 2961–2969, 2017.

[17] Guyue Hu, Bo Cui, Yuan He, and Shan Yu. Progressive relation learning for group activity recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 980–989, 2020.

[18] Mostafa S Ibrahim and Greg Mori. Hierarchical relational networks for group activity recognition and retrieval. In *Proc. Eur. Conf. Comput. Vis.*, pages 721–736, 2018.

[19] Dongkeun Kim, Jinsung Lee, Minsu Cho, and Suha Kwak. Detector-free weakly supervised group activity recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 20083–20093, 2022.

[20] Shuaicheng Li, Qianggang Cao, Lingbo Liu, Kunlin Yang, Shinan Liu, Jun Hou, and Shuai Yi. Groupformer: Group activity recognition with clustered spatial-temporal transformer. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 13668–13677, 2021.

[21] Wenxuan Liu, Xian Zhong, Zhuo Zhou, Kui Jiang, Zheng Wang, and Chia-Wen Lin. Dual-recommendation disentanglement network for view fuzz in action recognition. *IEEE Transactions on Image Processing*, 32:2719–2733, 2023.

[22] Lihua Lu, Huijun Di, Yao Lu, Lin Zhang, and Shunzhou Wang. Spatio-temporal attention mechanisms based model for collective activity recognition. *Signal Process. Image Commun.*, 74:162–174, 2019.

[23] Lihua Lu, Yao Lu, Ruizhe Yu, Huijun Di, Lin Zhang, and Shunzhou Wang. Gaim: Graph attention interaction model for collective activity recognition. *IEEE Trans. Multimedia*, 22(2):524–539, 2019.

[24] Mauricio Perez, Jun Liu, and Alex C Kot. Skeleton-based relational reasoning for group activity analysis. *Pattern Recognit.*, 122:108360, 2022.

[25] Rizard Renanda Adhi Pramono, Wen-Hsien Fang, and Yie-Tarng Chen. Relational reasoning for group activity recognition via self-attention augmented conditional random field. *IEEE Transactions on Image Processing*, 30:8184–8199, 2021.

[26] Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. stagnet: An attentive semantic rnn for group activity recognition. In *Proc. Eur. Conf. Comput. Vis.*, pages 101–117, 2018.

[27] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. Objects in context. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1–8, 2007.

[28] Kohei Sendo and Norimichi Ukita. Heatmapping of people involved in group activities. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pages 1–6, 2019.

[29] Tianmin Shu, Dan Xie, Brandon Rothrock, Sinisa Todorovic, and Song Chun Zhu. Joint inference of groups, events and

human roles in aerial videos. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4576–4584, 2015.

[30] Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. Cern: confidence-energy recurrent network for group activity recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 5523–5531, 2017.

[31] Elsa Siggiridou and Dimitris Kugiumtzis. Granger causality in multivariate time series using a time-ordered restricted vector autoregressive model. *IEEE Transactions on Signal Processing*, 64(7):1759–1773, 2015.

[32] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Proc. Adv. Neural Inf. Process. Syst.*, 27, 2014.

[33] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 5693–5703, 2019.

[34] Masato Tamura, Rahul Vishwakarma, and Ravigopal Vennelakanti. Hunting group clues with transformers for social group activity recognition. In *Proc. Eur. Conf. Comput. Vis.*, pages 19–35, 2022.

[35] Haritha Thilakarathne, Aiden Nibali, Zhen He, and Stuart Morgan. Pose is all you need: The pose only group activity recognition system (pogars). *Mach. Vis. Appl.*, 33(6):95, 2022.

[36] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 4489–4497, 2015.

[37] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Proc. Adv. Neural Inf. Process. Syst.*, 30, 2017.

[39] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

[40] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1430–1439, 2018.

[41] Xiao Wang, Zheng Wang, Wu Liu, Xin Xu, Qijun Zhao, and Shin'ichi Satoh. Towards causality inference for very important person localization. In *ACM International Conference on Multimedia*, pages 6618–6626, 2022.

[42] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 9964–9974, 2019.

[43] Zhao Xie, Tian Gao, Kewei Wu, and Jiao Chang. An actor-centric causality graph for asynchronous temporal inference in group activity. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 6652–6661, 2023.

[44] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Higcin: Hierarchical graph-based cross inference network for group activity recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[45] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Social adaptive module for weakly-supervised group activity recognition. In *Proc. Eur. Conf. Comput. Vis.*, pages 208–224, 2020.

[46] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1472–1481, 2023.

[47] Hangjie Yuan and Dong Ni. Learning visual context for group activity recognition. In *Proc. AAAI Conf. Artif. Intell.*, pages 3261–3269, 2021.

[48] Hangjie Yuan and Dong Ni. Learning visual context for group activity recognition. In *Proc. AAAI Conf. Artif. Intell.*, pages 3261–3269, 2021.

[49] Hangjie Yuan, Dong Ni, and Mang Wang. Spatio-temporal dynamic inference network for group activity recognition. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 7476–7485, 2021.

[50] Xiaolin Zhai, Zhengxi Hu, Dingye Yang, Lei Zhou, and Jingtai Liu. Spatial temporal network for image and skeleton based group activity recognition. In *Proc. Asian Conf. Comput. Vis.*, pages 20–38, 2022.

[51] Chi Zhang, Baoxiong Jia, Mark Edmonds, Song-Chun Zhu, and Yixin Zhu. Acre: Abstract causal reasoning beyond covariation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 10643–10653, 2021.

[52] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4320–4328, 2018.

[53] Honglu Zhou, Asim Kadav, Aviv Shamsian, Shijie Geng, Farley Lai, Long Zhao, Ting Liu, Mubbasir Kapadia, and Hans Peter Graf. Composer: Compositional reasoning of group activity in videos with keypoint-only modality. In *Proc. Eur. Conf. Comput. Vis.*, 2022.

[54] Xingyi Zhou, Anurag Arnab, Chen Sun, and Cordelia Schmid. How can objects help action recognition? In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2353–2362, 2023.

[55] Yingying Zhu, Nandita M Nayak, and Amit K Roy-Chowdhury. Context-aware modeling and recognition of activities in video. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2491–2498, 2013.