Learning Group Activity Features Through Person Attribute Prediction

Chihiro Nakatani[1] Hiroaki Kawashima[2] Norimichi Ukita[1]
[1] Toyota Technological Institute, Japan [2] University of Hyogo, Japan

Abstract

This paper proposes Group Activity Feature (GAF) learning in which features of multi-person activity are learned as a compact latent vector. Unlike prior work in which the manual annotation of group activities is required for supervised learning, our method learns the GAF through person attribute prediction without group activity annotations. By learning the whole network in an end-to-end manner so that the GAF is required for predicting the person attributes of people in a group, the GAF is trained as the features of multi-person activity. As a person attribute, we propose to use a person's action class and appearance features because the former is easy to annotate due to its simpleness, and the latter requires no manual annotation. In addition, we introduce a location-guided attribute prediction to disentangle the complex GAF for extracting the features of each target person properly. Various experimental results validate that our method outperforms SOTA methods quantitatively and qualitatively on two public datasets. Visualization of our GAF also demonstrates that our method learns the GAF representing fined-grained group activity classes. Code: https://github.com/chihina/GAFL-CVPR2024.

1. Introduction

A group activity is defined as what multiple people jointly engage in. Group activities are important targets in image and video understanding, such as team plays in sports [36], conversations in social scenes [3], and people flows in surveillance cameras [28].

Group Activity Recognition (**GAR**), in which a frame or video is classified into either of the predefined group activity classes, has been widely investigated in recent years [2, 3, 6, 9, 11, 14, 16, 19, 21, 25, 29, 37–40, 42, 45]. All of these GAR methods are based on supervised learning that requires the ground-truth group activities, as shown in Fig. 1 (a). In addition, it is known that person action recognition, which is also achieved based on a supervised learning manner, supports GAR [2, 3, 6, 9, 11, 14, 19, 21, 25, 29, 37, 38, 40, 45].

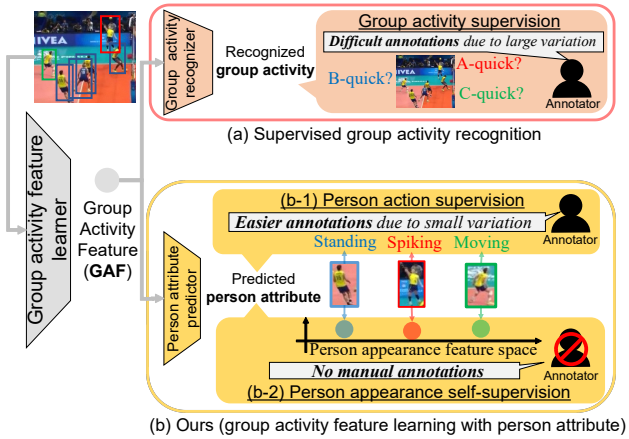

Figure 1. Difference between annotations for GAR and our group activity feature learning. (a) Supervised GAR employs group activity annotations that are difficult due to various similar group activities. (b-1) Our GAF learning employs person action annotations that are easy due to their simplicity. (b-2) We further propose annotation-free GAF learning with person appearance features.

Such supervised learning requires manually-annotated training data. For GAR, group activity annotations are required. In addition to labor-intensive and erroneous annotations, a difficulty peculiar to GAR is the complexity of the group activities. For example, while only four elemental group activity classes are annotated in a widely-used team-sport dataset [15], they may be insufficient for practical purposes such as tactical analysis (e.g., 200 or more plays are defined in American Football [8]). That is, more fine-grained activity classes are required for several applications of group activity analysis. It is, however, difficult to correctly define and annotate such complex, fine-grained activity classes, even manually, because of visually minor but highly contextual differences among those classes.

Such difficulty in manual annotations of group activities motivates us to learn a Group Activity Feature (**GAF**) in which features of multi-person activity are learned as a compact latent vector without group activity annotations. We note here that GAFs, which represent the complex features of multi-person activity, may have enough information to predict the attributes of each person in a group.

With this regard, this paper proposes GAF learning using person attribute (e.g., action and appearance), which is easier to provide compared with the complex group activity annotation, as shown in Fig. 1 (b). With the person attribute related to the group activity, the GAF can narrow down the possible attribute of each person. For example, Fig. 2, in which a spike is observed, shows that the possible person action enclosed by the purple rectangle (i.e., digging) can be narrowed down by the GAF representing the scene context (i.e., spike group activity) with the person's location.

As a person attribute for our GAF learning, person action can be used as shown in Fig. 1 (b-1). This is because person action is strongly related to the group activities and is essential to support GAR as mentioned above. However, manual annotations are still required to use person action. While the annotations of person action are easier than the one of group activities defined with complex people interaction, such action annotations are still labor-intensive.

To alleviate such difficulty in manual annotations, we also propose to learn GAF through the task of appearance feature prediction of each person in a group without manual annotations, as shown in Fig. 1 (b-2). While the reduction of annotation cost have been studied widely, including active learning [13, 26], few-shot learning [35, 44], and self-supervised learning [4, 5, 7, 10, 18, 23, 24, 32, 33, 41, 43], we follows the framework of self-supervised learning, where supervision signals are derived from the input data and any manual annotations are unnecessary.

Our novel contributions are summarized as follows:

- **GAF learning through person attribute prediction:** Unlike supervised GAR, we propose GAF learning through person attribute prediction without group activity annotations. This paper proposes its two variants:
  - **GAF Learning with Person Action Classes (GAFL-PAC):** As a person attribute, we utilize person action classes, which are more easily annotated than group activities defined with complex inter-people interactions (Fig. 1 (b-1)). In addition, for practical use of GAR and other group-related applications, the annotations of person actions are already given in general [3, 14].
  - **GAF Learning with Person Appearance Features (GAFL-PAF):** We also utilize person appearance features that can be obtained by a pre-trained model (e.g., VGG) without manual annotation (Fig. 1 (b-2)). The appearance features are suitable as a person attribute due to the high relationship with multi-person activity.
- **Fine granularity of our GAF:** With the GAF learning through person attribute prediction, our method can learn fine-grained GAF that represents visual-subtle but important differences that are not represented in a manually-defined activity class, as shown in Fig. 6.
- **Location-Guided Person Attribute Prediction Using GAF:** While predicting the person attribute using the GAF is required in our method, extracting the features of each target person from the GAF is difficult. This is because the GAF represents complex features of multi-person activity. This feature extraction is achieved by location guidance in which each person's location feature is embedded into the GAF with positional encoding.

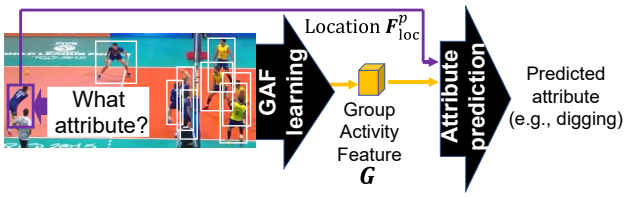

Figure 2. Example of our group activity feature learning. In this example, a group activity feature is learned to extract the scene context (i.e., spike group activity) through prediction of person attribute (e.g., digging). See Fig. 3 for the detailed architecture.

2. Related Work

2.1. Group Activity Recognition

Supervision by group activity labels. While we propose the GAF learning without group activity annotations (Fig. 1 (b)), various GAR methods supervised by group activity annotations (i.e., Fig. 1 (a)) have been proposed. In [16, 40, 42], only the labels of group activity are required for training. In [16], a group activity is recognized from a whole image without any person features. Kim *et al* [42] and Yan *et al* [40] employ the set of person features as input for a Graph Neural Network (GNN) as with [6, 37].

Supervision by the labels of person action and group activity. Different from the aforementioned methods only with group activity supervision, the GAR network is jointly trained with the person action recognition network in [2, 3, 6, 9, 11, 14, 19, 21, 25, 29, 37, 45] for augmenting GAR. In [6, 37], GNN models the interactions between person features. In [9, 11, 19, 29, 45], Transformer [31] improves modeling the interactions between person features for GAR.

While all of the methods introduced in Sec. 2.1 employ the annotation of group activities, our method learns GAF without group activity labels. Following the success of interaction modeling in [9, 11, 19, 29, 45], we also learn the GAF with Transformer.

2.2. Self-supervised Representation Learning

Image-based pretext tasks. Most pretext tasks for self-supervised image representation utilize the transformation of original images. In Gidaris *et al*. [10], Zhang *et al*. [43], and Larsson *et al*. [18], each original image is rotated, affinely- and projectively-transformed, and grayscaled, respectively. The image representation model is trained so that the model undoes each image. While these meth-

ods [10, 18, 43] employ a whole image, patches extracted from the image are used in [4, 23, 24]. Doersch *et al*. [4] predict the spatial configuration (i.e., relative positions) of randomly-sampled patches. In Noroozi *et al*. [23], jigsaw puzzles in which all patches of an image are shuffled are solved. Pathak *et al*. [24] inpaint partially-erased images.

Video-based pretext tasks. While the image-based methods can be applied to a video, video-specific pretext tasks are also proposed in [5, 7, 32, 33, 41]. In Yao *et al*. [41] and Wang *et al*. [33], the paces of video clips are arbitrarily changed (e.g., normal, half, and double speeds), and video representation is learned by solving a pace prediction problem. In Fernando *et al*. [7], sequential video clips are randomly reordered, and video representation is learned so that these reordered and original clips can be classified. Wang *et al*. [32] employ appearance cues as well as motion statistics for spatial-temporal representation learning.

While all of the representation learning methods introduced above extract low-level image features for various downstream tasks such as general image/video classification and prediction, our method focuses on features useful for group-related tasks (e.g., group scene retrieval and clustering). As with our method, Ibrahim *et al*. [14] aims to extract multi-person scene features through GNN in an unsupervised manner, while the relation of each person to a group activity is not fully captured. Our method incorporates such intimate relations between individuals and group activity (e.g., digging in a spike scene) using the task of location-guided person attribute prediction from a GAF.

3. Proposed Method

Our GAF learning network consists of three stages, (a), (b), and (c) (Fig. 3). In stage (a), the features of each person are extracted (Sec. 3.1). In stage (b), the features of several people are masked (i.e., removed) during training for GAF enhancement (Sec. 3.2). Then, the masked person features are fed into the transformer-based GAF learning network (Sec. 3.3). In stage (c), the attribute of each person is predicted from the GAF with location guidance (Sec. 3.4).

3.1. Person Feature Extractor

Overview. As shown in Fig. 3 (a), the set of person features $\boldsymbol{F}_{set} \in \mathbb{R}^{T\times N\times C}$ are extracted from images. $\boldsymbol{F}_{set}$ is composed of the features of N people obtained from T frames in a video. C is the dimension of person p's feature vector, $\boldsymbol{F}^p_{ind} \in \mathbb{R}^C$, indicated by a blue cuboid in Fig. 3 (a). $\boldsymbol{F}^p_{ind}$ consists of appearance features (denoted by $\boldsymbol{F}^p_{app} \in \mathbb{R}^C$) and location features (denoted by $\boldsymbol{F}^p_{loc} \in \mathbb{R}^C$) in our implementation because their effectiveness is validated for GAR [9, 11, 19, 25, 45], while other features (e.g., body keypoints) can also be used.

Detail. $\boldsymbol{F}^p_{app}$ is extracted by the following three steps, as with [42]. First, a feature map is extracted from the whole image by VGG [27]. Then, RoIAlign [12] is applied with the bounding box of each person to obtain the feature map for each person. Finally, the feature map of each person is embedded into C-dimensional appearance features with a linear transformation. From the $\boldsymbol{F}^p_{app}$ of N people between T frames, we construct $\boldsymbol{F}_{app} \in \mathbb{R}^{T\times N\times C}$. In addition to $\boldsymbol{F}_{app}$, $\boldsymbol{F}^p_{loc}$ is also essential to understand the spatial structure of a group, as $\boldsymbol{F}^p_{loc}$ is used for GAR [9, 11, 25, 37, 45]. As with [11], the center point of each person bounding box, (x, y), is embedded into $\boldsymbol{F}^p_{loc}$ by a spatial positional encoding so that $\boldsymbol{F}_{loc} \in \mathbb{R}^{T\times N\times C}$ whose dimension is equal to $\boldsymbol{F}_{app}$. Finally, $\boldsymbol{F}_{loc}$ is elementwise added with $\boldsymbol{F}_{app}$ to obtain the set of person features, $\boldsymbol{F}_{set}$. Here, we denote person p's features in T frames (i.e., a slice of $\boldsymbol{F}_{set}$) as $\boldsymbol{F}^p_{ind} \in \mathbb{R}^{T\times C}$ $(p \in \{1, \cdots, N\})$ and, similarly, person p's location features in T frames (i.e., a slice of $\boldsymbol{F}_{loc}$) as $\boldsymbol{F}^p_{loc} \in \mathbb{R}^{T\times C}$ $(p \in \{1, \cdots, N\})$.

3.2. Masked Person Modeling (MPM)

Overview. In the set of person features $\boldsymbol{F}_{set}$, the features of randomly sampled people are masked (e.g., $\boldsymbol{F}^1_{ind}$ indicated by a black cuboid in Fig. 3 (b)) during training. By extracting such features with our Masked Person Modeling (MPM), our network is expected to learn the features of the interaction between unmasked people for predicting the attribute of the masked person.

Detail. The binary mask for p-th person (denoted by $\boldsymbol{M}^p \in \mathbb{R}^{T\times C}$) is initialized by filling 1 in all values. Then, N_{mask} people to be masked are randomly sampled from all N people in a scene. All values in $\boldsymbol{M}^p$ for the randomly sampled people are updated with 0.

From the all $\boldsymbol{M}^p$ of N people, we construct $\boldsymbol{M} = [M^1, \cdots, M^N] \in \mathbb{R}^{T\times N\times C}$. Then, $\boldsymbol{M}$ is elementwise multiplied by $\boldsymbol{F}_{set}$ to obtain the masked person features $\boldsymbol{F}_{mask}$ for our GAF training. Note that all values in $\boldsymbol{M}$ are set to be 0 during inference to preserve features of all people in a GAF.

3.3. Group Activity Feature Learning Network

Overview. The GAF $\boldsymbol{G}$ is extracted from the set of masked person features $\boldsymbol{F}_{mask}$ obtained in Sec. 3.2. For this GAF learning, we employ Transformer to model spatial- and temporal-interactions between people in accordance with [11, 19]. Our transformer-based GAF learning network consists of two branches (i.e., TS and ST branches) similar to [11]. As shown in Fig. 3, the Temporal Transformer encoder (T^T) and the Spatial Transformer encoder (T^S) are used in both branches, but they are placed in reverse order in the two branches. The combination of these two branches can model the spatial-temporal interactions

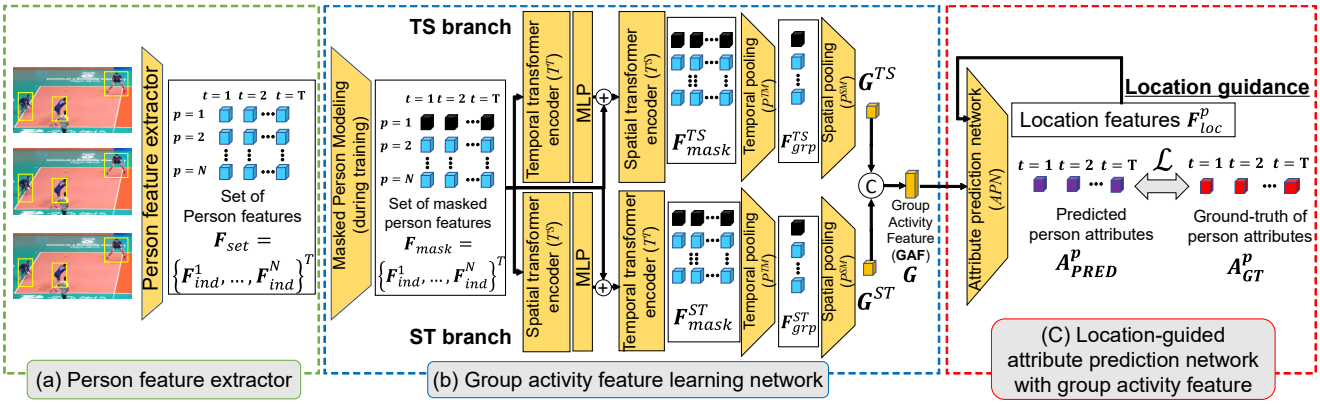

Figure 3. Overview of our GAF learning network. (a) Person feature extractor. The person feature is composed of appearance and location features. (b) GAF learning network. The GAF is learned from extracted people features. (c) Location-guided attribute prediction network with the GAF. The attribute of each person is predicted from the location feature of the person and the GAF extracted in (b).

between people, as proven in GAR [11, 19]. While the transformer-based network is used in our method, the network can be replaced with any SOTA network without difficulties because the stage (i.e., Fig. 3 (b)) is modularized.

Detail. $\boldsymbol{F}_{mask}$ is independently fed into the TS and ST branches to acquire $\boldsymbol{F}^{TS}_{mask}$ and $\boldsymbol{F}^{ST}_{mask}$, respectively:

$$\boldsymbol{F}^{TS}_{mask} = T^S(\boldsymbol{F}_{mask} + MLP(T^T(\boldsymbol{F}_{mask}))) \quad (1)$$

$$\boldsymbol{F}^{ST}_{mask} = T^T(\boldsymbol{F}_{mask} + MLP(T^S(\boldsymbol{F}_{mask}))) \quad (2)$$

where MLP denotes the multi-layer perceptron. $\boldsymbol{G}^{TS} \in \mathbb{R}^C$ and $\boldsymbol{G}^{ST} \in \mathbb{R}^C$ are obtained from $\boldsymbol{F}^{TS}_{mask}$ and $\boldsymbol{F}^{ST}_{mask}$, respectively, by the Temporal Max Pooling (denoted by P^{TM}) and the Spatial Max Pooling (denoted by P^{SM}), as with previous methods [9, 37]:

$$\boldsymbol{G}^{TS} = P^{SM}(P^{TM}(\boldsymbol{F}^{TS}_{mask})) \quad (3)$$

$$\boldsymbol{G}^{ST} = P^{SM}(P^{TM}(\boldsymbol{F}^{ST}_{mask})) \quad (4)$$

Finally, $\boldsymbol{G}^{TS}$ and $\boldsymbol{G}^{ST}$ are concatenated to obtain the final GAF $\boldsymbol{G} \in \mathbb{R}^{2C}$ as follows:

$$\boldsymbol{G} = \boldsymbol{G}^{TS} \oplus \boldsymbol{G}^{ST} \quad (5)$$

During training, $\boldsymbol{G}$ is fed into the attribute prediction network (denoted by APN), as shown in Fig. 3 (c). By back-propagating a loss used in this person attribute prediction not only inside APN but also across the whole network shown in Fig. 3, $\boldsymbol{G}$ can be trained as the GAF. The details of person attribute prediction and the loss used in this prediction are described in Sec. 3.4 and Sec.3.5, respectively.

After predicting $\boldsymbol{G}$ in inference, $\boldsymbol{G}$ can be used in various ways such as a pretrained model for downstream supervised tasks and unsupervised learning tasks (e.g., retrieval and clustering), as mentioned in Sec. 1. The effectiveness of $\boldsymbol{G}$ in these tasks is validated in Sec. 4.

3.4. Attribute Prediction with GAF

Overview. Using $\boldsymbol{G}$ obtained in Sec. 3.3, the attribute of each person is predicted by the APN (Fig. 3 (c)). To predict the attribute of p-th person, we use their location features (i.e., $\boldsymbol{F}^p_{loc}$) as guidance for attribute prediction from $\boldsymbol{G}$.

Detail. $\boldsymbol{G}$ is fed into the attribute prediction network with the location of each person as follows:

$$\boldsymbol{A}^p_{PRED} = APN(\boldsymbol{G}, \boldsymbol{F}^p_{loc}) \quad (6)$$

where $\boldsymbol{A}^p_{PRED} \in \mathbb{R}^{T \times R}$ denotes the predicted attribute of each person obtained from $\boldsymbol{G}$ with their location. The dimension of R changes depending on the type of the predicted person attribute as follows:

(i) Person action: R is the number of action classes and the A^p_{PRED} represents predicted action probabilities.

(ii) Person appearance features: R is the dimension of the appearance features (i.e., C) extracted in Sec. 3.1.

3.5. Loss Function

The whole network is trained with a loss function (denoted by $\mathcal{L}$) for attribute prediction as a pretext task as follows:

(i) Person action: When the person attribute is the action, $\mathcal{L}$ is the cross-entropy loss: $\mathcal{L} = \mathcal{L}_{CE}(\boldsymbol{A}^p_{PRED}, \boldsymbol{A}^p_{GT})$, where $\boldsymbol{A}^p_{GT}$ denotes the ground-truth one-hot vector for action.

(ii) Person appearance features: For appearance features, $\mathcal{L}$ is the mean squared loss function: $\mathcal{L} = \mathcal{L}_{MSE}(\boldsymbol{A}^p_{PRED}, \boldsymbol{A}^p_{GT})$, where $\boldsymbol{A}^p_{GT}$ denotes the extracted appearance features (i.e., $\boldsymbol{F}_{app}$) shown in Sec. 3.1.

4. Experiments

4.1. Datasets

The Volleyball dataset, which contains highly correlated players, is mainly used to validate the effectiveness of our GAF learning. The Collective Activity dataset is also used to validate the generality of our method in Sec. 4.4, while people in this dataset are not highly related to each other compared with the Volleyball dataset.

VolleyBall Dataset (VBD) [15] consists of 4,830 sequences extracted from 55 games. Each sequence is annotated with one of the predefined eight group activity classes, i.e., Left-spike, Right-spike, Left-set, Right-set, Left-pass, Right-pass, Left-winpoint, and Right-winpoint. While each sequence has 41 frames, its center 20 frames have annotations with the full-body bounding boxes of all players and their action classes, i.e., Waiting, Setting, Digging, Falling, Spiking, Jumping, Moving, Blocking, and Standing.

Collective Activity Dataset (CAD) [3] contains 44 videos. In each video, every ten frames are annotated with person action classes, i.e., NA, Crossing, Waiting, Queuing, Walking, and Talking, and their bounding boxes. The group activity class is determined by the largest number of person actions in each frame, while the NA class is not included as a group activity class. We follow the previous methods [34, 42] to merge the Crossing and Walking into Moving.

4.2. Evaluation Protocols

4.2.1 Evaluation Tasks

The quality of the estimated GAF is verified with the following two types of retrieval tasks.

Action set retrieval. We employ the same action set retrieval task as in [14]. Action IoU in [14] evaluates the similarity of action structures between query and retrieved images based on the overlap of action distributions. If the IoU exceeds a predefined threshold (e.g., 0.5), the query and retrieved images are regarded as matched. However, in this action IoU, all action classes are counted equally without weights, although the action class distribution is imbalanced (e.g., the percentage of "standing" people, who are less informative for tactics, is over 68% in VBD).

To resolve this class imbalanced problem, we propose Action Frequency-Inverse Scene Frequency (AF-ISF) inspired by Term Frequency-Inverse Document Frequency (TF-IDF [1]) that evaluates the importance of a word in a document. In AF-ISF, each scene is represented by a feature vector in which each value is computed from frequency statics of action classes as follows:

$$\boldsymbol{FV}^i = [FV_1^i, FV_m^i, \cdots, FV_M^i] \tag{7}$$

$$FV_m^i = AF_m^i \cdot ISF_m \tag{8}$$

where i and M denote the image index and the number of actions, respectively. AF_m^i is the frequency of m-th action class in each image. ISF_M is the inverse frequency of each action class in the dataset. In AF-ISF, the similarity between $\boldsymbol{FV}^j$ and $\boldsymbol{FV}^k$, where j and k denote the IDs of query and retrieved images, evaluates the similarity of action structure of the two images. If the cosine similarity exceeds a predefined threshold (e.g., 0.5), the query and retrieved images are regarded as matched as with the Action IoU.

While AF-ISF alleviates the action class imbalanced problem, several action classes are distinctive for representing a group scene even if the action distribution is balanced (e.g., "Spiking" is more important than "Falling" in VBD). Due to this problem, AF-ISF is still improper for the contextual representation of group scenes.

Group activity retrieval. Based on the discussion above, we propose to further evaluate whether or not the group activity class of a query scene matches that of the retrieved scene. Note that the group activity annotations are given to the test data (i.e., all query and retrieved images) only for this evaluation and are not used in our GAF learning.

4.2.2 Evaluation Metrics

With IoU and AF-ISF in the action set retrieval and group activity matching in the group activity retrieval, we compute the Hit@K used in [14]. In addition, the mean Average Precision (mAP) is used in the action set retrieval as with [14]. For mAP, the Euclidean distance of $\boldsymbol{G}$ between query and retrieved images is used as the confidence indicator.

4.3. Training Details

Our network is optimized by Adam [17] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The whole image is resized into 320x640 and 240x360 for VBD and CAD, respectively. We employ the VGG-19 and Inception-v3 models as a person feature extractor (Fig. 3 (a)) for VBD and CAD, respectively. While the person feature extractor trained on person action recognition is fine-tuned through our GAF learning in GAFL-PAC, we freeze the person feature extractor trained on ImageNet in GAFL-PAF following the previous method [14]. Our APN consists of three fully-connected layers. As for the other details, we follow the widely used setting [42]. See the details in the supplementary material.

4.4. Comparative Experiments

We compare our methods with other methods by the following three types of experimental results: (1) the results of retrieval (Sec. 4.4.1), (2) the results of GAR (Sec. 4.4.2), and (3) the visualized distributions of GAFs (Sec. 4.4.3).

Table 1. Quantitative comparison of retrieval on the VolleyBall Dataset (VBD). The results obtained in two experimental settings (i.g., GAFL-PAC and GAFL-PAF) are separated by double lines. The best result in each column is colored in red. Results obtained by the concatenation of output features (i.e., $\boldsymbol{F}_{grp}^{TS}$ and $\boldsymbol{F}_{grp}^{ST}$) and $\boldsymbol{G}$ are denoted as "Ours-ind" and "Ours-grp", respectively.

	Retrieval type	Action set (IoU [14])				Action set (AF-IDF)				Group activity		
	Method	Hit@1	Hit@2	Hit@3	mAP	Hit@1	Hit@2	Hit@3	mAP	Hit@1	Hit@2	Hit@3
GAFL - PAC	HiGCIN [40]	74.3	84.9	89.5	55.7	59.8	73.6	80.3	30.5	50.0	66.3	74.5
	DIN [42]	79.7	90.1	93.4	60.2	74.5	85.2	88.3	39.3	57.0	73.1	81.1
	Dual-AI [11]	67.6	84.7	91.6	56.9	72.6	83.7	88.6	53.0	64.4	76.5	82.0
	Ours-ind	82.7	91.6	95.0	59.1	79.0	86.8	89.8	45.6	82.7	88.8	91.3
	Ours-grp	83.0	92.7	95.5	64.2	80.1	88.4	91.5	59.9	84.8	89.6	91.8
GAFL - PAF	B1-Compact128 [14]	57.9	75.7	84.3	45.8	41.3	60.8	71.4	29.3	30.3	48.0	59.9
	B2-VGG19 [14]	63.8	80.6	86.8	46.8	46.7	65.8	75.7	29.4	35.4	53.6	65.0
	HRN [14]	60.9	78.6	86.0	46.9	40.8	60.9	72.9	28.7	31.2	47.0	57.6
	Ours-ind	64.2	80.8	88.3	45.0	50.4	69.3	77.6	30.1	55.0	72.3	79.2
	Ours-grp	64.8	82.7	90.3	46.4	52.3	71.4	81.0	31.4	61.1	75.1	82.4

Table 2. Quantitative comparison of retrieval on the Collective Activity Dataset (CAD).

	Retrieval type	Action set (IoU [14])				Action set (AF-IDF)				Group activity		
	Method	Hit@1	Hit@2	Hit@3	mAP	Hit@1	Hit@2	Hit@3	mAP	Hit@1	Hit@2	Hit@3
GAFL - PAC	HiGCIN [40]	80.8	85.4	89.7	57.9	81.0	85.2	89.3	61.6	86.1	88.8	91.9
	DIN [42]	71.4	74.1	74.9	51.5	90.1	92.7	94.0	52.8	90.8	92.5	93.2
	Dual-AI [11]	61.0	72.5	76.7	61.5	85.5	86.9	88.1	82.7	82.1	84.1	84.7
	Ours-ind	76.2	82.6	89.4	78.9	94.8	95.6	95.9	82.2	94.9	95.4	95.7
	Ours-grp	81.8	90.7	93.5	69.9	96.1	96.5	96.6	93.9	94.9	95.6	96.3
GAFL - PAF	B1-Compact128 [14]	48.8	60.3	68.2	38.0	81.8	88.2	89.7	52.6	82.4	88.4	90.1
	B2-VGG19 [14]	53.6	61.6	66.1	35.3	71.1	80.3	83.8	46.7	72.2	80.8	84.2
	HRN [14]	37.1	50.1	58.6	22.2	53.2	64.8	72.5	34.2	54.0	64.8	72.4
	Ours-ind	67.6	81.3	85.9	53.3	83.7	88.9	90.2	57.5	88.5	91.2	91.9
	Ours-grp	52.7	70.3	74.1	46.4	74.0	80.5	82.6	60.1	79.2	81.0	82.0

4.4.1 Retrieval

To validate the effectiveness of the GAF for group representation, the set of person features (i.e., the concatenation of $\boldsymbol{F}_{grp}^{TS}$ and $\boldsymbol{F}_{grp}^{ST}$ in Fig. 3 (b)) is also evaluated. Specifically, the output features of P^{TM} in the TS and ST branches ($\boldsymbol{F}_{grp}^{TS}$ and $\boldsymbol{F}_{grp}^{ST}$ in Fig. 3) are concatenated and also used for retrieval in our method. Results obtained by the concatenation of output features (i.e., $\boldsymbol{F}_{grp}^{TS}$ and $\boldsymbol{F}_{grp}^{ST}$) and $\boldsymbol{G}$ are denoted as "Ours-ind" and "Ours-grp", respectively.

Volleyball dataset (GAFL-PAC). Our method is compared with the SOTA GAR methods [11, 40, 42] which are only trained with person action labels as with our method. Table 1 (top) shows that our method is best in action set and group activity retrieval. The results validate that our method learns features about the action structure and the group activity of a scene into the compact latent vector (i.e., $\boldsymbol{G}$) efficiently. The large gain in the group activity retrieval shows that the features of multi-person activity are learned in our GAF in contrast to the SOTA GAR methods [11, 40, 42] where the set of person features is used for the retrieval.

Volleyball dataset (GAFL-PAF). Our method is compared with the SOTA method [14], the only existing GAF learning method using no person-action and group-activity annotations, and baseline methods as with [14]. From Table 1 (bottom), we see that our method performs best in all metrics except that "HRN" is better than our method in mAP of Action set (IoU). However, the performance difference is small (i.e., 46.9 and 46.4 in "HRN" and "Ours-grp," respectively). Furthermore, our method is better than "HRN" in Action set (AF-IDF). From these results, we can say that our GAF is better even in the action set retrieval. Regarding group activity retrieval, our method significantly outperforms the SOTA method. The results validate that our method learns the contextual features of multi-person activity better than the SOTA method.

Collective activity dataset (GAFL-PAC). The results in GAFL-PAC on CAD are shown in Table 2 (top). The results validate that our method is better than Dual-AI [11] in all metrics. While "Ours-grp" is better than "Ours-ind" in most metrics, "Ours-ind" is better in mAP of Action set (IoU). The action set (IoU) evaluates the similarity of the number of people, so the results may come from that the change in the number of people between scenes on this dataset is addressed well in "Ours-ind." Specifically, we can consider that such information about the number of people may be preserved well in "Ours-ind." where the set of person features is used for retrieval.

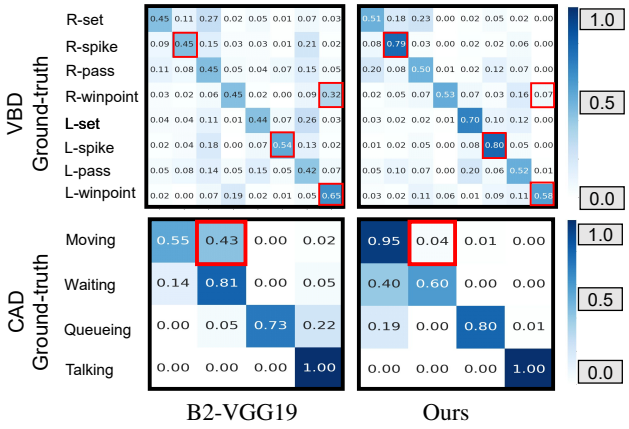

Figure 4. Confusion matrices of GAR by nearest neighbor retrieval on VBD and CAD in GAFL-PAF. Each row and column show the ground-truth and recognized group activity, respectively. Results of the other methods are shown in the supplementary material.

Collective activity dataset (GAFL-PAF). Table 2 (bottom) shows that our method is the best in all metrics among the other methods. The results validate the wide applicability of GAF learned even in general scenes included in CAD. While "Ours-grp" is better than "Ours-ind" on VBD, "Ours-ind" is better on CAD. These opposite results may come from the difference between the number of people in each image. While $N = 12$ people are observed in most images in VBD, around five people on average on CAD. As the number of features in $\boldsymbol{F}_{grp}^{TS}$ and $\boldsymbol{F}_{grp}^{ST}$ increase in proportion to the number of observed people, the difficulty in learning $\boldsymbol{F}_{grp}^{TS}$ and $\boldsymbol{F}_{grp}^{ST}$ becomes higher. Since this learning difficulty might occur on VBD, spatial max pooling used in "Ours-grp" is effective for reducing the feature dimension from NC of $\boldsymbol{F}_{grp}^{TS}$ and $\boldsymbol{F}_{grp}^{ST}$ to C of $\boldsymbol{G}^{TS}$ and $\boldsymbol{G}^{ST}$.

4.4.2 Group Activity Recognition

While no group annotation is used in our GAF learning, the group activity class retrieved by Hit@1, which is equal to 1-nearest neighbor classification, can be regarded as the result of GAR, as done in [20]. While results only in GAFL-PAF are shown in this section, results in GAFL-PAC are available in the supplementary material.

Volleyball dataset (GAFL-PAF). As shown in Fig. 4, our method outperforms "B2-VGG19" in all group activity classes except for L-winpoint. While "B2-VGG19" is better than "Ours" in L-winpoint shown at the bottom right of each confusion matrix, "B2-VGG19" often misrecognizes the R-winpoint scene as L-winpoint, as shown in the red rectangle cells. Left-spike and Right-spike are especially recognized better in our method than the others. This superiority of our method can be interpreted as follows. In Left-spike and Right-spike, spiking and blocking players who

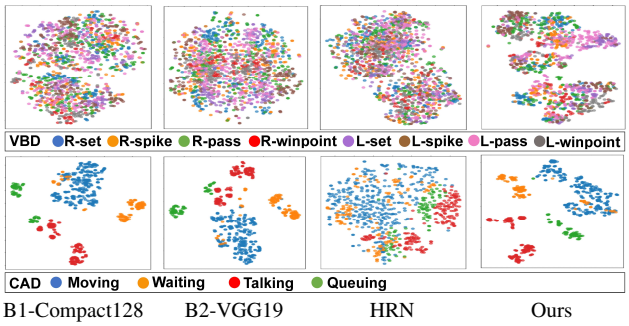

Figure 5. Visualization of the learned GAF on VBD and CAD in GAFL-PAF. The color of each sample shows the ground-truth of the group activity label corresponding to each test sample.

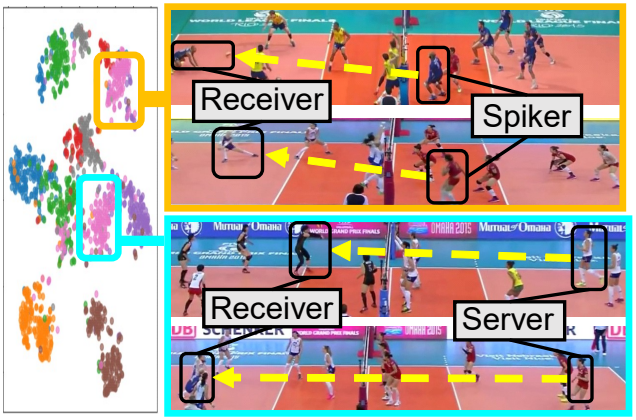

Figure 6. Visualization of the learned GAF on VBD in GAFL-PAC. The magenta data points (i.e., "L-pass") are divided into two sub-categories based on the context (i.e., whether the receiving is caused by the spiking or serving of the opposite player).

are distinctive both in appearance and location cues are always observed. Since these distinctive people are important in predicting not only their attributes but also the attributes of other people, the features of these distinctive people are extracted well even in the compact latent vector (i.e., GAF).

Collective activity dataset (GAFL-PAF). The confusion matrices are shown in Fig. 4. The results show that our method is the best in all activity classes. In particular, while VGG19 gets many false negatives in Moving, the number of false negatives in Moving of our method is almost zero.

4.4.3 Visualization of Learned GAFs

The distribution of learned GAFs in all test images is visualized in a 2D space by t-SNE [30]. The color of each point shows its annotated group activity class. Figure 5 shows that our method can learn the GAFs better than the other methods because (1) while the inner-class variance is small, the inter-class variance is large, and (2) the data points of similar group activities (e.g., Left-pass and Left-set) are closer.

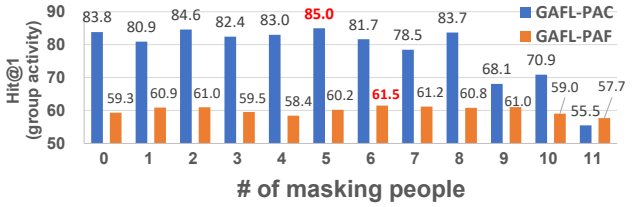

Figure 7. Performance changes depending on the number of masking people on VBD.

Table 3. Effectivenss of our location-guidance on VBD. Results obtained by "Ours-grp" are shown as "Ours."

	Retrieval type	Action set (IoU)	Action set (AF-IDF)	Group activity
	Method	Hit@1	Hit@1	Hit@1
GAFL-PAC	Ours w/o $\boldsymbol{F}_{loc}$	80.0	75.6	69.5
	Ours	83.0	80.1	84.8
GAFL-PAF	Ours w/o $\boldsymbol{F}_{loc}$	64.1	51.7	53.2
	Ours	64.8	52.3	61.1

Our GAFs in GAL-PAC on VBD are also shown in Fig. 6. This figure shows that data points with the same group activity labels are divided into sub-activities. For example, L-pass indicated by magenta data points are divided into two clusters. While the cluster enclosed by the orange rectangle represents L-pass where the person is receiving a ball from the opposite spiker, the other cluster enclosed by the light blue rectangle represents L-pass where the person is receiving a ball from the opposite server. As shown in this example, our GAFs are learned well enough to represent visually subtle but important differences that are not represented in the manually defined activity classes.

4.5. Detailed analysis

Comparison of the number of masked persons. We explore the optimal number of masked people (i.e., N_{mask}) for our MPM. For this comparison, we change N_{mask} from 0 to $N-1$. $N_{mask}=0$ means that $\boldsymbol{F}_{set}$ is directly fed into the transformer encoder without the MPM.

Figure 7 shows that the performance changes depending on N_{mask} on VBD. The best performance is obtained when N_{mask} is the middle number (i.e., 5 and 6 in the GAFL-PAC and GAFL-PAF, respectively), while the performance gain from $N_{mask}=0$ is insignificant. On the other hand, we can also see that the performance with a large masking ratio (i.e., $N_{mask}\in\{10,11\}$) drops. These results reveal that the extreme difficulty in the attribute prediction of the masked person from a few non-masked people leads GAF learning to failure.

Effect of location-guidance in GAF learning. We ablate $\boldsymbol{F}^p_{loc}$ which is used for guidance to extract the features of each person in the action prediction network in Fig. 3 (c) (see also Sec. 3.4) on VBD.

Table 4. Comparison with supervised GAR. Double lines separate the results obtained by VBD and CAD.

Dataset	Method	Accuracy
VBD	Dual-AI [11]	92.1
	Ours w/ group activity labels	92.4
CAD	Dual-AI [11]	94.1
	Ours w/ group activity labels	96.6

In GAFL-PAC, "Ours" is better than "Ours w/o $\boldsymbol{F}^p_{loc}$" in the all metrics. In particular, the performance gain in the group activity retrieval is large. We can interpret the reason as follows. In "Ours w/o $\boldsymbol{F}^p_{loc}$," only $\boldsymbol{G}$ is used to predict the attribute of each person in a scene. It makes the model predict not the attribute of each person but the attribute distribution of people in a scene. Therefore, the GAF can be used for action set retrieval to some extent because the similarity of action distribution is evaluated in this retrieval. However, spatial interaction between people is not learned in the GAF. It causes a significant performance drop in the group activity retrieval. In GAFL-PAF, we can also see the high-performance gain of "Ours" in the group activity retrieval. From these results, we can conclude that our location guidance is important for learning group activity.

Fine-tuning for Group Activity Recognition. Table 4 shows that the comparison of "Dual-AI" and "Ours" with group activity labels on VBD and CAD. In "Dual-AI," the network is trained with group activity labels from scratch. In "Ours," after the pretraining by our GAF learning without the group activity supervision, the network is fine-tuned for GAR with group activity labels.

On both datasets, our method is better than "Dual-AI." In particular, on CAD, the GAR accuracy obtained by our method is 2.5% better than the GAR accuracy obtained by "Dual-AI." The results demonstrate that our GAF learning is effective as a pre-training for supervised GAR.

5. Concluding Remarks

Instead of group activity annotations which is difficult due to a variety of similar group activities, our method learns GAF through person attribute prediction without group activity annotations. Quantitative comparisons and visualized results show that our method can learn informative GAF compared with other methods on the two public datasets.

While our method outperforms all the other methods in our experiments not only for GAF learning but also for GAR, our method is understandably inferior to GAR supervised by group activity annotations. Exploring other pretext tasks such as predicting the joint attention of a group (e.g., [22]) is important future work for further GAF enhancement.

References

[1] Akiko N. Aizawa. An information-theoretic perspective of tf-idf measures. *Inf. Process. Manag.*, 39(1):45–65, 2003. 5

[2] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional relational machine for group activity recognition. In *CVPR*, 2019. 1, 2

[3] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *ICCVW*, 2009. 1, 2, 5

[4] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015. 2, 3

[5] Haodong Duan, Nanxuan Zhao, Kai Chen, and Dahua Lin. Transrank: Self-supervised video representation learning via ranking-based transformation recognition. In *CVPR*, 2022. 2, 3

[6] Mahsa Ehsanpour, Alireza Abedin, Fatemeh Sadat Saleh, Javen Shi, Ian D. Reid, and Hamid Rezatofighi. Joint learning of social groups, individuals action and sub-group activities in videos. In *ECCV*, 2020. 1, 2

[7] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *CVPR*, 2017. 2, 3

[8] Richard B. Foster. *American Football Playbook: 210 Field Templates*. Createspace Independent Pub, 2016. 1

[9] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees G. M. Snoek. Actor-transformers for group activity recognition. In *CVPR*, 2020. 1, 2, 3, 4

[10] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 2, 3

[11] 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 *CVPR*, 2022. 1, 2, 3, 4, 6

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, 2017. 3

[13] Fabian Caba Heilbron, Joon-Young Lee, Hailin Jin, and Bernard Ghanem. What do I annotate next? an empirical study of active learning for action localization. In *ECCV*, 2018. 2

[14] Mostafa S. Ibrahim and Greg Mori. Hierarchical relational networks for group activity recognition and retrieval. In *ECCV*, 2018. 1, 2, 3, 5, 6

[15] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, 2016. 1, 5

[16] Dongkeun Kim, Jinsung Lee, Minsu Cho, and Suha Kwak. Detector-free weakly supervised group activity recognition. In *CVPR*, 2022. 1, 2

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[18] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *CVPR*, 2017. 2, 3

[19] 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 *ICCV*, 2021. 1, 2, 3, 4

[20] Kengo Nakata, Youyang Ng, Daisuke Miyashita, Asuka Maki, Yu-Chieh Lin, and Jun Deguchi. Revisiting a knn-based image classification system with high-capacity storage. In *ECCV*, 2022. 7

[21] Chihiro Nakatani, Kohei Sendo, and Norimichi Ukita. Group activity recognition using joint learning of individual action recognition and people grouping. In *MVA*, 2021. 1, 2

[22] Chihiro Nakatani, Hiroaki Kawashima, and Norimichi Ukita. Interaction-aware joint attention estimation using people attributes. In *ICCV*, 2023. 8

[23] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 2, 3

[24] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2, 3

[25] Rizard Renanda Adhi Pramono, Yie-Tarng Chen, and Wen-Hsien Fang. Empowering relational network by self-attention augmented conditional random fields for group activity recognition. In *ECCV*, 2020. 1, 2, 3

[26] Aayush Jung Rana and Yogesh S. Rawat. Hybrid active learning via deep clustering for video action detection. In *CVPR*, 2023. 2

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3

[28] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, 2018. 1

[29] Masato Tamura, Rahul Vishwakarma, and Ravigopal Vennelakanti. Hunting group clues with transformers for social group activity recognition. In *ECCV*, 2022. 1, 2

[30] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 7

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 2

[32] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *CVPR*, 2019. 2, 3

[33] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *ECCV*, 2020. 2, 3

[34] Minsi Wang, Bingbing Ni, and Xiaokang Yang. Recurrent modeling of interaction context for collective activity recognition. In *CVPR*, 2017. 5

[35] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Mingqian Tang, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. Hybrid relation guided set matching for few-shot action recognition. In *CVPR*, 2022. 2

[36] Dekun Wu, He Zhao, Xingce Bao, and Richard P. Wildes. Sports video analysis on large-scale data. In *ECCV*, 2022. 1

[37] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *CVPR*, 2019. 1, 2, 3, 4

[38] Zhao Xie, Tian Gao, Kewei Wu, and Jiao Chang. An actor-centric causality graph for asynchronous temporal inference in group activity. In *CVPR*, 2023. 1

[39] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Social adaptive module for weakly-supervised group activity recognition. In *ECCV*, 2020.

[40] 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.*, 45(6):6955–6968, 2023. 1, 2, 6

[41] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *CVPR*, 2020. 2, 3

[42] Hangjie Yuan, Dong Ni, and Mang Wang. Spatio-temporal dynamic inference network for group activity recognition. In *ICCV*, 2021. 1, 2, 3, 5, 6

[43] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. AET vs. AED: unsupervised representation learning by auto-encoding transformations rather than data. In *CVPR*, 2019. 2, 3

[44] Sipeng Zheng, Shizhe Chen, and Qin Jin. Few-shot action recognition with hierarchical matching and contrastive learning. In *ECCV*, 2022. 2

[45] 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 *ECCV*, 2022. 1, 2, 3