COMPOSER: Compositional Reasoning of Group Activity in Videos with Keypoint-Only Modality

Honglu Zhou[1⋆], Asim Kadav[2], Aviv Shamsian[3], Shijie Geng[1], Farley Lai[2], Long Zhao[4], Ting Liu[4], Mubbasir Kapadia[1], and Hans Peter Graf[2]

[1] Department of Computer Science, Rutgers University, Piscataway, NJ, USA
{hz289,sg1309,mk1353}@cs.rutgers.edu
[2] NEC Laboratories America, Inc., San Jose, CA, USA
{asim,farleylai,hpg}@nec-labs.com
[3] Bar-Ilan University, Israel
aviv.shamsian@biu.ac.il
[4] Google Research, Los Angeles, CA, USA
{longzh,liuti}@google.com

Abstract. Group Activity Recognition detects the activity collectively performed by a group of actors, which requires compositional reasoning of actors and objects. We approach the task by modeling the video as tokens that represent the multi-scale semantic concepts in the video. We propose COMPOSER, a Multiscale Transformer based architecture that performs attention-based *reasoning* over tokens at each scale and learns group activity *compositionally*. In addition, prior works suffer from scene biases with privacy and ethical concerns. We only use the keypoint modality which reduces scene biases and prevents acquiring detailed visual data that may contain private or biased information of users. We improve the multiscale representations in COMPOSER by clustering the intermediate scale representations, while maintaining consistent cluster assignments between scales. Finally, we use techniques such as auxiliary prediction and data augmentations tailored to the keypoint signals to aid model training. We demonstrate the model's strength and interpretability on two widely-used datasets (Volleyball and Collective Activity). COMPOSER achieves up to +5.4% improvement with just the keypoint modality [1].

Keywords: Keypoint-only group activity recognition · Compositionality · Multiscale representations · Transformer · Video understanding

1 Introduction

Group Activity Recognition (GAR) detects the activity collectively performed by a group of actors in a short video clip [12,50]. GAR has widespread societal implications in a variety of domains including security, surveillance, kinesiology, sports analysis, robot-human interaction, and rehabilitation [17,40,57,16].

⋆ Work done as a NEC Labs intern.
[1] Code is available at https://github.com/hongluzhou/composer

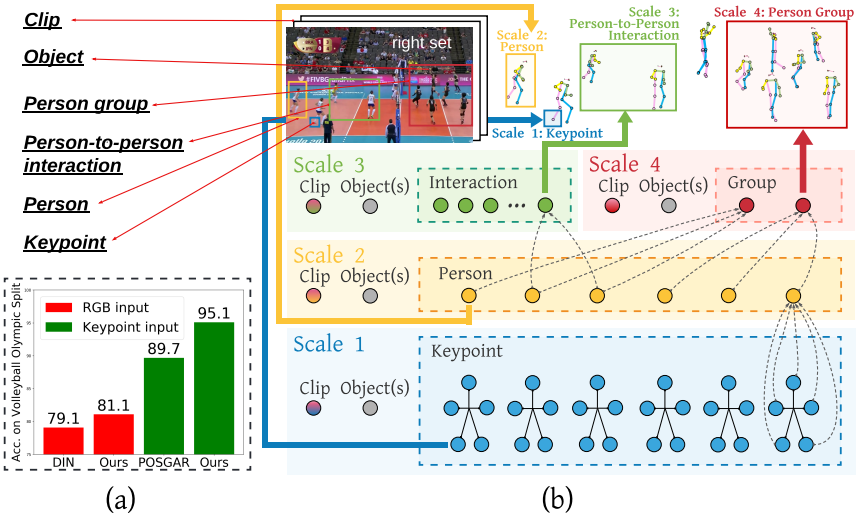

Fig. 1. (a) The keypoint-only setup generalizes better for GAR. The Volleyball Olympic split [46] ensures videos having vastly different scene background between training and testing, which can examine GAR model's scene generalization ability. RGB-based methods severely suffer from scene biases and have poor model generalizability. **(b) Main idea.** We propose `COMPOSER` that uses keypoint only modality for GAR by modeling a video as *tokens* that represent the multiscale semantic concepts in the video, which include *keypoint*, *person*, person-to-person *interaction*, person *group*, *object* if present, and the *clip*. Four scales are formed by grouping actor-related tokens according to their semantic hierarchy. Representations of tokens in coarser scales are learned and aggregated from tokens of the finer scales. `COMPOSER` (Fig. 3) facilitates compositional reasoning of group activity in videos.

The task requires addressing two challenges. First, GAR requires a *compositional understanding* of the scene [2]. Because of the crowded scene, it is challenging to learn meaningful representations for GAR over the entire scene [50]. Since group activity often consists of sub-group(s) of actors and scene objects, the final action label depends on a compositional understanding of these entities [50,56]. Second, GAR benefits from *relational reasoning* over scene elements to understand the relative importance of entities and their interactions [20,54]. For example, in a volleyball game, persons around the ball performing the jumping action are more important than others standing in the scene.

Existing work has proposed to jointly learn the group activity with individual actions [25,42,41,23,6,4] or person sub-groups [31,35,16] for a compositional understanding of the group activity. Meanwhile, graph [57,24,49,20] and transformer [17,31] based models have been proposed for relational reasoning over scene entities. However, these methods do not sufficiently make use of the multiscale scene elements in the GAR task by modeling over entities at either one semantic scale (e.g., person [17,57,49,20]) or two scales (person and person

group [31,35,16], or keypoint and person [39]). More importantly, explicit multiscale modeling is neglected, lacking consistent compositional representations for the group action tasks. Furthermore, majority of the prior GAR methods rely on the RGB modality (see Table. 3), which causes the model more likely to have privacy and ethical issues when deployed in real-world applications [19]. Last but not least, the RGB input hinders the model's robustness to changes in background, lighting conditions or textures, and often results in poor model generalizability due to scene biases (see Fig. 1 (a)) [11,44].

In this paper, we present `COMPOSER` that addresses *compositional learning* of entities in the video and *relational reasoning* about these entities. Inspired by how humans are particularly adept at representing objects in different granularities meanwhile reasoning their interactions to turn sensory signals into a high-level knowledge [22,30], we approach GAR by modeling a video as tokens that represent the multi-scale semantic concepts in the video (Fig. 1 (b)). Compared to the aforementioned prior works, we consider more fine-grained scene entities that are grouped into *four* scales. By combining the scales together with Multiscale Transformer (Fig. 4), `COMPOSER` provides attention-based reasoning over tokens at each scale, which makes the higher-level understanding of the group activity possible. Moreover, `COMPOSER` uses only the keypoint modality. Using only the 2D (or 3D) keypoints as input, our method can prevent the sensor camera from acquiring detailed visual data that may contain private or biased information of users [2]. Keypoints also allow the model to focus on the action-specific cues, and help the model be more invariant to the scene biases. `COMPOSER` generalizes much better to testing data with different scene backgrounds (see the Volleyball Olympic split results in Table. 1).

`COMPOSER` learns *consistent* multiscale representations which boost the performance for GAR (Fig. 2). This is achieved by contrastive clustering assignments of clips. Intuitively, a model can recognize the group activity using representations of entities at just one particular scale. Hence, we consider representations of the clip token learned across scales as representations of different *views* of the clip. Such perspective allows us to cluster clip representations learned at all scales while enforcing consistency between cluster assignments produced from different scales of the *same* clip. In order to enforce this consistency, we follow [8] and use a *swapped* prediction mechanism where we predict the cluster assignment of a scale from the representation of another scale. However, distinct from related works [8,3,10], which use information from multiple augmentations or modalities for self-supervised learning from unlabelled images or videos, we use information from multiple scales for the task of group activity recognition. Contrasting clustering assignments enhance our intermediate representations and the overall performance. Finally, we use techniques such as auxiliary prediction at each scale and data augmentation methods such as *Actor Dropout* to aid training.

Our contributions are three-fold:

[2] Even for the keypoint extraction backbone which our method is agnostic to, there are existing works [19] that perform privacy-preserving keypoint estimation.

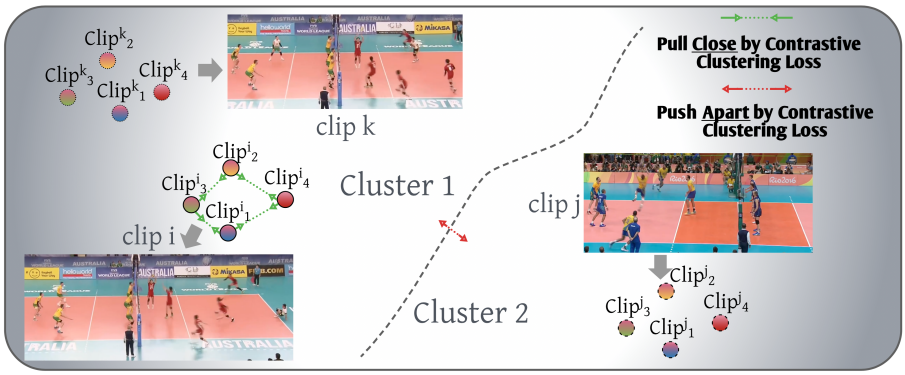

Fig. 2. Embedding space learned by COMPOSER. COMPOSER exploits a contrastive clustering objective (Sec. 3.3) to learn *consistent* multiscale representations for GAR. This is achieved by clustering clip representations learned at all scales. The clustering objective encourages an "agreement" between scales on the high-level knowledge learned ('Pull Close' representations of the same clip). Contrastive learning is performed on the clusters, which also helps the model to discriminate between clips with different semantic characteristics ('Pull Close' representations of the semantically-similar clips and 'Push Apart' those that are semantically-different). In the illustration, we use subscript to denote the scale and use superscript to indicate different clips.

1. We present COMPOSER for compositional reasoning of group activity in videos. COMPOSER can distill and convey high-level semantic knowledge from the elementary elements of the human-centered videos. We learn contrastive clustering assignment to improve the multiscale representations. By maintaining a consistent cluster assignment across the multiple scales of the *same* clip, an agreement between scales on the high-level knowledge learned can be promoted to optimize the representations across scales.
2. We use only the keypoint modality that allows COMPOSER to address the privacy and ethical concerns and to be robust to changes in background, with auxiliary prediction and data augmentation methods tailored to learning group activity from the keypoint modality.
3. We demonstrate the model's strength and interpretability on two commonly-used datasets (Volleyball and Collective Activity) and COMPOSER achieves up to +5.4% improvement using just the keypoint modality.

2 Related Work

Much of the recent research on GAR explores how to capture the actor relations [24,5,49,20,40]. Several works tackle this problem from a graph-based perspective [24,33,54,53]. Some utilize attention modeling [41,51,33,56] including using Transformers [17,31]. Existing works have primarily used RGB- and/or optical-flow-based features with RoIAlign [18] to represent actors [53,41,49,6].

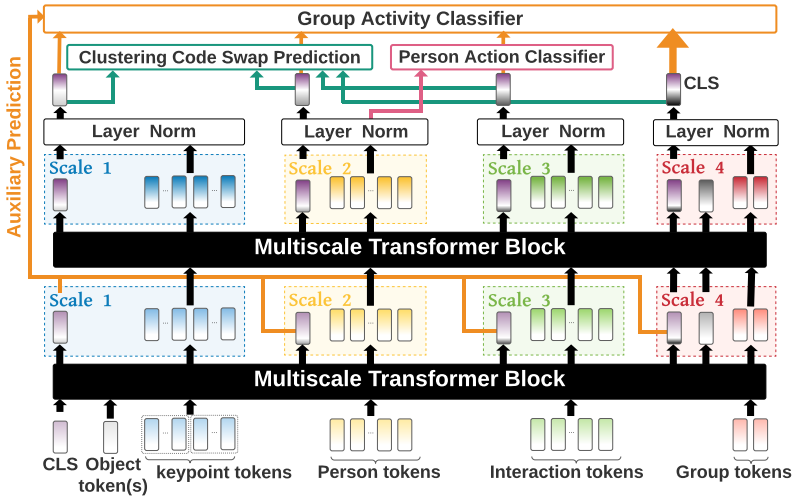

Fig. 3. COMPOSER. Given tokens that represent the multiscale semantic concepts (Fig. 1) in the human-centered video, COMPOSER jointly learns group activity, individual actions and contrastive clustering assignments of clips. Auxiliary predictions are enforced to aid training (Sec. 3.5).

A few recent works replace or augment these features with keypoints/poses of the actors [46,39,17,56]. In this paper, we use only the light-weight coordinate-based keypoint representation. We propose a Multiscale Transformer block to hierarchically reason about entities at different semantic scales and we aid learning group activities by improving the musicale representations. Please see an in-depth discussion on related works in Appendix G.

3 Methodology

We present COMPOSER (Fig. 3), a novel Multiscale Transformer based architecture for GAR. In Sec. 3.1, we describe the multi-scale semantic tokens representing a video with group activities. We introduce COMPOSER and especially its reasoning module Multiscale Transformer in Sec. 3.2. We describe data augmentations in Sec. 3.4 and the exact formulation of auxiliary prediction in Sec. 3.5.

3.1 Tokenizing a Video as Hierarchical Semantic Entities

We model a video as semantic tokens that allow our method easily adaptable to understanding any videos with multi-actor multi-object interactions [34].
• **Person Keypoint**. We define a person keypoint token, $\mathbf{k}_p^j \in \mathbb{R}^d$ that represents a keypoint joint j $(j = 1, \ldots, j')$ of person p $(p = 1, \ldots, p')$ in all timestamps, where j' is the number of joint types and p' is the number of actors. The initial d-dimensional person keypoint token is learned by encoding the numerical

coordinates (in the image space) of a certain keypoint track[3]. The procedure of encoding includes coordinate embedding, time positional embedding, keypoint type embedding, and OKS-based feature embedding [45] to mitigate the issue of noisy estimated keypoints. Details are available in Appendix.

• **Person**. A person token is defined as $\mathbf{p}_p \in \mathbb{R}^d$, initially obtained by aggregating the standardized keypoint coordinates of person p over time through concatenation and FFN-based transformation.

• **Person-to-Person Interaction**. Modeling the person-to-person interactions is critical for GAR [50]. Unlike existing works that typically consider an interaction as an *edge* connecting two person nodes and learn a scalar to depict its importance [54], we model interaction as *nodes* (tokens) to allow for the modeling of complex higher-order interactions [34]. The person-to-person interaction token is defined as $\mathbf{i}_i \in \mathbb{R}^d$ where $i = 1, \ldots, p' \times (p'-1)$ (bi-directed interactions). Initial representation of the interaction between person p and q is learned from concatenation of $\mathbf{p}_p$ and $\mathbf{p}_q$, followed by FFN-based transformation.

• **Person Group**. We define the group token $\mathbf{g}_g \in \mathbb{R}^d$ where $g = 1, \ldots, g'$ for videos where sub-groups are often separable. g' denotes the num. of sub-groups in the video. Given the person-to-group mapping which can be obtained through various mechanisms (e.g., heuristics [39], k-means [31], etc [16,28].), representation of a group is an aggregate over representations of persons in the group similarly through concatenation and FFN.

• **Clip**. The special [`CLS`] token ($\in \mathbb{R}^d$) is a learnable embedding vector and is considered as the clip representation. `CLS` stands for classification and is often used in Transformers to "summarize" the task-related representative information from all tokens in the input sequence [15].

• **Object**. Scene objects can play a crucial role in videos where human(s) interact with object(s). E.g., in a volleyball game where one person is spiking and multiple nearby actors are all jumping with arms up, it can be difficult to tell which person is the key person with information of just the person keypoints due to their similar poses. The ball keypoints can help to distinguish the key person. Object keypoints can be used to represent an object in the scene with similar benefits of person keypoints (e.g., to boost model robustness [26]). Object keypoint detection [7,32] benefits downstream tasks such as human action recognition [21], object detection [26,55], tracking [36], etc [29]. Thus, we use object keypoints to represent each object for GAR. We denote object token $\mathbf{e}_e \in \mathbb{R}^d$ where $e = 1, \ldots, e'$ and e' is the maximal number of objects a video might have. Similar to person tokens, the initial object tokens are learned from aggregating the coordinate-represented object keypoints.

3.2 Multiscale Transformer

Multiscale Transformer takes a sequence of multiple-scale tokens as input, and refines representations of these tokens. Specifically, tokens of the four scales are:

[3] We use track-based representations [46,58,17,31] to represent each token.

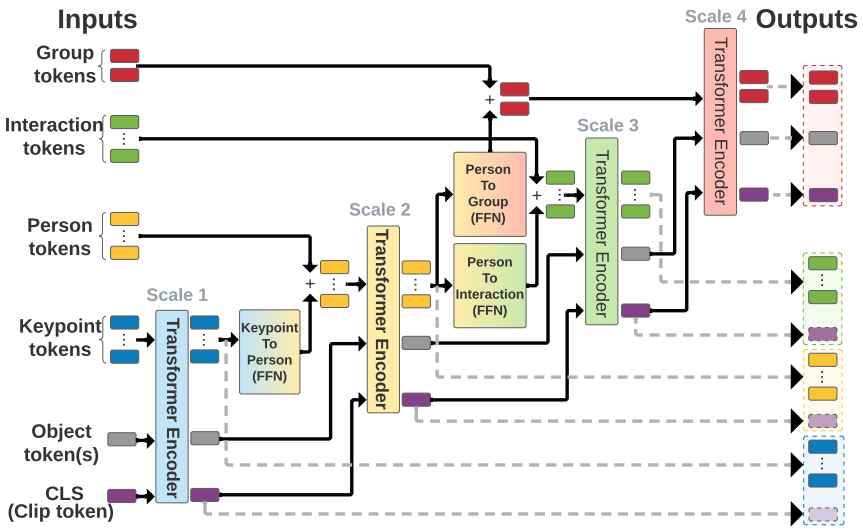

Fig. 4. Multiscale Transformer performs relational reasoning with four Transformer Encoders to operate self-attention on tokens of each scale, while stringing tokens of the four scales together with FFNs and Skip Connections to learn hierarchical representations that make a high-level understanding of group activity possible.

$$
\begin{aligned}
&\textit{Scale 1:} && \left\{[\texttt{CLS}], \mathbf{e}_1, \cdots, \mathbf{e}_{e'}, \mathbf{k}_1^1, \cdots, \mathbf{k}_{p'}^{j'}\right\}, \\
&\textit{Scale 2:} && \left\{[\texttt{CLS}], \mathbf{e}_1, \cdots, \mathbf{e}_{e'}, \mathbf{p}_1, \cdots, \mathbf{p}_{p'}\right\}, \\
&\textit{Scale 3:} && \left\{[\texttt{CLS}], \mathbf{e}_1, \cdots, \mathbf{e}_{e'}, \mathbf{i}_1, \cdots, \mathbf{i}_{p' \times (p'-1)}\right\}, \\
&\textit{Scale 4:} && \left\{[\texttt{CLS}], \mathbf{e}_1, \cdots, \mathbf{e}_{e'}, \mathbf{g}_1, \cdots, \mathbf{g}_{g'}\right\}.
\end{aligned} \tag{1}
$$

We utilize a Transformer encoder [47] at each scale to perform relational reasoning of tokens in that scale. We review details of Transformer in the Appendix.

Hierarchical representations of tokens are maintained in an elaborately designed Multiscale Transformer block (Fig. 4). In the Multiscale Transformer block, operations in the four scales are the same (but with different parameters) to maintain simplicity. Specifically, given a sequence of tokens of scale s (Eq. 1), Transformer encoder outputs refined representations of these tokens. Then, concatenation and FFN are used to aggregate refined representations of *actor-related* tokens, in order to form representations of actor-related tokens in the subsequent coarser scale $s{+}1$. Such learned representations are summed with their initial representations (input to the Multiscale Transformer) (i.e. Skip Connection). The resulting actor-related tokens, as well as scale s updated [CLS] token and object token(s) form the input sequence of the Transformer encoder in the scale $s{+}1$ (see wiring in Fig. 4).

COMPOSER uses the initial representations of the multi-scale semantic tokens (Sec. 3.1) as input, and utilizes multiple blocks of Multiscale Transformer to

perform relational reasoning over these tokens. With refined token representations, COMPOSER *jointly* learns group activity, individual actions and contrastive clustering of clips (the multitask-learning details are in Sec. 3.5).

3.3 Contrastive Clustering for Scale Agreement

We consider the clip tokens learned at different scales as representations of different *views* of the clip instance. Then, we cluster clip representations learned in all scales while enforcing consistency between cluster assignments produced from different scales of the clip. This can act as regularization of the embedding space during training (Fig. 2). To enforce consistency, we use a swapped prediction mechanism [8] where we predict the cluster assignment of a scale from the representation of another scale. COMPOSER jointly learns GAR and the swapped prediction task to capture an agreement of the common semantic information hidden across the scales.

Preliminaries. Suppose $\mathbf{v}_{n,s} \in \mathbb{R}^d$ represents the learned representation of clip n in scale s, where $s \in \{1,2,3,4\}$. Following prior works [8,27], we first project the representation to the unit sphere. We then compute a code (i.e., cluster assignment) $\mathbf{q}_{n,s} \in \mathbb{R}^K$ by mapping $\mathbf{v}_{n,s}$ to a set of K *trainable* prototype vectors, $\{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$. We denote by $C \in \mathbb{R}^{K \times d}$ the matrix whose rows are the $\mathbf{c}_1, \ldots, \mathbf{c}_K$.

Swapped Prediction. Suppose s and w denote 2 different scales from the four representation scales. The swapped prediction problem aims to predict the code $\mathbf{q}_{n,s}$ from $\mathbf{v}_{n,w}$, and $\mathbf{q}_{n,w}$ from $\mathbf{v}_{n,s}$, with the following loss function:

$$\mathcal{L}_{\text{swap}}\left(\mathbf{v}_{n,w}, \mathbf{v}_{n,s}\right) = \ell\left(\mathbf{v}_{n,w}, \mathbf{q}_{n,s}\right) + \ell\left(\mathbf{v}_{n,s}, \mathbf{q}_{n,w}\right) \tag{2}$$

where $\ell\left(\mathbf{v}_{n,w}, \mathbf{q}_{n,s}\right)$ measures the fit between $\mathbf{v}_{n,w}$ and $\mathbf{q}_{n,s}$. $\ell\left(\mathbf{v}_{n,w}, \mathbf{q}_{n,s}\right)$ is the cross entropy loss between $\mathbf{q}_{n,s}$ and the probability obtained by taking a softmax of the dot products of $\mathbf{v}_{n,w}$ and prototypes in C:

$$\ell\left(\mathbf{v}_{n,w}, \mathbf{q}_{n,s}\right) = -\sum_{k=1}^{K} \mathbf{q}_{n,s}^{(k)} \log \frac{\exp\left(\frac{1}{\tau}\mathbf{v}_{n,w}\mathbf{c}_k^\top\right)}{\sum_{k'=1}^{K} \exp\left(\frac{1}{\tau}\mathbf{v}_{n,w}\mathbf{c}_{k'}^\top\right)} \tag{3}$$

where τ is a temperature parameter. The total loss of the swapped prediction problem is taking Eq. (2) computed over all pairs of scales and all N clips,

$$\mathcal{L}_{\text{cluster}} = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{w,s \in \{1,2,3,4\} \& w \neq s} \mathcal{L}_{\text{swap}}\left(\mathbf{v}_{n,w}, \mathbf{v}_{n,s}\right)\right) \tag{4}$$

Online Clustering. This step produces the cluster assignments using the learned prototypes C and the learned clip representations *only within a batch*, $V \in \mathbb{R}^{B \times d}$ where B denotes the batch size. We perform the clustering in an online fashion for faster training and use the method proposed in [8]. Specifically, online clustering yields the codes $Q \in \mathbb{R}^{B \times K}$. We compute codes Q such that all examples in a batch are equally partitioned by the prototypes (which prevents the trivial

solution where every clip has the same code). Q is optimized to maximize the similarity between the learned clip representations and the prototypes,

$$\max_{Q \in \mathcal{Q}} \operatorname{Tr}\left(QCV^{\top}\right) + \varepsilon H(Q), \tag{5}$$

$$\mathcal{Q} = \left\{ Q \in \mathbb{R}_{+}^{B \times K} \mid \mathbf{1}_B Q = \frac{1}{K}\mathbf{1}_K, Q\mathbf{1}_K^{\top} = \frac{1}{B}\mathbf{1}_B^{\top} \right\}$$

where the trace Tr is the sum of the elements on the main diagonal, H is the entropy function, and ε is a parameter that controls the smoothness of the mapping. $\mathbf{1}_K \in \mathbb{R}^K$ and $\mathbf{1}_B \in \mathbb{R}^B$ are a vector of ones to enforce the equipartition constraint. The continuous solution Q^* of Eq. (5) is computed with the iterative Sinkhorn-Knopp algorithm [14,8].

3.4 Data Augmentation for Keypoint Modality

We use the following data augmentations to aid training and improve generalization ability of the model learned from the *keypoint* modality.

Actor Dropout is performed by removing a random actor in a random frame, inspired by [37] that masks agents with probabilities to predict agent behaviors for autonomous driving. We remove actors by replacing the representation of the actor with a zero vector.

Horizontal Flip is often used by existing GAR methods [58,46,39], which is performed on the video frame level. This augmentation causes the pose of each person and positions of (left and right) sub-groups flipped horizontally. We add a small random perturbation on each flipped keypoint.

Horizontal Move means we horizontally move all keypoints in the clip by a certain number of pixel locations, which is randomly determined per video and bounded by a pre-defined number (i.e., 10). Similarly, afterwards a small random perturbation is applied on each keypoint.

Vertical Move is done similar to the Horizontal Move, except we move the keypoints in the vertical direction.

Novel practices like Actor Dropout, Horizontal/Vertical Move and random perturbations help the model to perform GAR from noisy estimated keypoints.

3.5 Auxiliary Prediction and Multitask Learning

We take the learned representation of the clip at *each* scale of *each* Multiscale Transformer block, and perform *auxiliary group activity predictions* (Fig. 3). Specifically, each of the clip representations learned at each scale of each block is sent as input to the group activity classifier to produce one GAR result. In addition, person representation from the last Multiscale Transformer block is the input to a person action classifier. Meanwhile, the loss of the swapped prediction problem is computed given the learned representations of the clip of all 4 scales from the last Multiscale Transformer block. The total loss is:

$$\mathcal{L}_{\text{total}} = \sum_{m=1}^{M-1} \mathcal{L}_{\text{groupAux}} + \lambda \left(\mathcal{L}_{\text{groupLast}} + \mathcal{L}_{\text{person}} + \mathcal{L}_{\text{cluster}} \right) \tag{6}$$

where $\mathcal{L}_{\text{groupAux}}$ represents the loss from Auxiliary Prediction incurred by clip representations at different scales and early blocks of the Multiscale Transformer, $\mathcal{L}_{\text{groupLast}}$ is from the last Multiscale Transformer block, $\mathcal{L}_{\text{person}}$ is the person action classification loss, and $\mathcal{L}_{\text{cluster}}$ is the contrastive clustering loss (Eq. 4). m denotes the index of the Multiscale Transformer block, M is the total number of the Multiscale Transformer blocks, and λ is a hyper-parameter that weights the importance of predictions from the last block. For metric evaluation, we use the clip token from the last scale in the last Multiscale Transformer as input to the group activity classifier.

4 Experimental Evaluation

4.1 Dataset

The Volleyball dataset [25] (VD) comprises 4,830 clips from 55 videos. The group activity labels include 8 activities: 4 main activities (*set*, *spike*, *pass*, *winpoint*) which are divided into two subgroups, *left* and *right*. Each player can perform one of the 9 actions: *blocking*, *digging*, *falling*, *jumping*, *moving*, *setting*, *spiking*, *standing* and *waiting*. The dataset has a default '**Original**' split in which train/test videos were randomly splitted (39 train and 16 test videos). A skewed '**Olympic**' split [46] was later released in which train/test videos are splitted according to the match venues: 29 train videos are from the same 2012 London Olympics venue, while the rest 26 test videos are from numerous venues, and thus largely differs from the train videos w.r.t. the scene background.
The Collective Activity dataset [13] (CAD) is a dataset with 44 real-life videos [50]. The group activity labels are *crossing*, *waiting*, *queueing*, *walking* and *talking* (person action labels have an additional 'N/A' class). We follow prior works to merge the class *crossing* and *walking* into *moving* [57,48,52,53], and use the same train-test split [57,49,41] and actor tracklets [57,6]. Please refer to Appendix for implementation details on both datasets.

4.2 Comparison with State-of-the-Arts

Scene Generalization for Keypoint-only Setup To support the keypoint-only setup for GAR, we first compare the generalization capability of models using either RGB or the keypoint modality. In Table 1, I3D and VGG-16 are two commonly-used image backbone by prior RGB-based GAR methods; the rest are all GAR models (all use VGG-16 as the backbone).

On VD Olympic split, the best prior RGB-based method is DIN [57] in Table 1. We substitute DIN with a `COMPOSER` variant [4] (Sec. 1) that also consumes RGB input instead of keypoint, and the result is 81.1% which is 2% higher than DIN, suggesting the stronger reasoning strength of `COMPOSER`, but the accuracy is still low due to the RGB signals. POGARS [46] uses the keypoint

[4] This `COMPOSER` variant consumes RGB-based ROI-aligned person features as input, and thus only models 3 scales: person, interaction, and the group scale.

Table 1. Test accuracy on VD **under different train/test splits**. Yellow shaded rows highlight the methods use RGB input, and blue for keypoint

Model	**VD Acc. (%) ↑**	
	Olympic	Original
I3D [9]	73.9	84.6
VGG-16 [43]	76.4	91.6
PCTDM [52]	75.2	91.7
SACRF [40]	71.1	91.8
AT [17]	76.9	93.0
ARG [49]	77.8	93.3
TCE-STBiP [56]	78.5	93.5
DIN [57]	79.1	93.6
POGARS [46]	89.7	93.2
`COMPOSER` (ours)	**95.1**	**93.7**
Improvement	+5.4%	+0.1%

*Note: Keypoint-based methods do NOT use ball keypoint in this table in order to have a rigorous comparison because RGB-based methods are unaware of such info.

Table 2. Comparisons with state-of-the-art (SOTA) methods that leverage **only keypoint information** on the **VD Original split**. `COMPOSER` outperforms existing methods and achieves a new highest record (+0.7% improvement)

Model	**Keypoint**		Acc.
	Actor	Object	
Zappardino *et al.* [58]	✔		91.0
GIRN [39]	✔		88.4
	✔	✔	92.2
AT [17]	✔		92.3
	✔	✔	92.8
POGARS [46]	✔		93.2
	✔	✔	93.9
`COMPOSER` (ours)	✔		93.7
	✔	✔	**94.6**

modality and has an accuracy of 89.7%, higher than all RGB-based methods. `COMPOSER` with the keypoint-only modality obtains 95.1% accuracy and *significantly outperforms* prior methods, yielding +**5.4**% improvement. These results imply that the keypoint-only setup can reduce scene biases, and generalize better than approaches relying on the RGB modality to testing data with different visual characteristics from training.

We also report the results of these methods that we obtained on VD Original split in Table 1. From this side-by-side comparison, the difference between the Olympic and Original split is vivid. Current GAR methods have quite saturated performances on the Original split of VD and the results are all very high (more evidence later). Therefore, we recommend readers using the more challenging VD Olympic split for future research on GAR. Note that the `COMPOSER` that outperforms prior methods in Table 1 is only an ablated version of ours in that not using the object token(s). In addition, GroupFormer [31] is currently the best-performing method (Table 4 in Appendix) and its RGB-only variant has the result of 94.1% accuracy on VD Original split. However, GroupFormer uses additional scene features with the Inception-v3 backbone.

Comparisons of Methods Using Keypoint-only Modality In Table 2, we compare `COMPOSER` with more GAR methods that use only the keypoint modality on VD Original split following conventions. `COMPOSER` achieves a new SOTA 94.6% accuracy with +**0.7**% improvement.

Table 3. Comparisons with SOTA methods that use **a single or multiple modalities** on the original split of VD and CAD. "Flow" denotes optical flow input, and "Scene" denotes features of the entire frames. Fewer modalities indicates a stronger capability of the model itself (*fewer checks are better*). The top 3 performance scores are highlighted as: **First**, *Second**, *Third*. `COMPOSER` outperforms the latest GAR methods that use a single modality (+0.7% improvement on VD and +2.8% improvement on CAD), and performs favorably compared with methods that exploit multiple expensive modalities

Model	Modality				Dataset	
	Keypoint	RGB	Flow	Scene	VD	CAD
HDTM [25]		✔			81.9	81.5
CERN [42]		✔			83.3	87.2
stagNet [41]		✔			89.3	89.1
RCRG [23]		✔			89.5	N/A
SSU [6]		✔			90.6	N/A
PRL [20]		✔			91.4	N/A
ARG [49]		✔			92.5	91.0
HiGCIN [53]		✔			91.5	93.4
DIN [57]		✔			93.6	N/A
Zappardino *et al.* [58]	✔				91.0	N/A
GIRN [39]	✔				92.2	N/A
AT [17]	✔				92.3	N/A
POGARS [46]	✔				93.9	N/A
CRM [4]		✔	✔		93.0	85.8
AT [17]		✔	✔		93.0	92.8
Ehsanpour *et al.* [16]		✔		✔	93.1	89.4
GIRN [39]	✔	✔	✔		94.0	N/A
TCE+STBiP [56]	✔	✔		✔	*94.7*	N/A
SACRF [40]	✔	✔	✔	✔	*95.0**	*95.2*
GroupFormer [31]	✔	✔	✔	✔	**95.7**	**96.3**
`COMPOSER` (ours)	✔				94.6	*96.2**

*Note: The best results of each method that were reported by the method authors are listed in the table in order to be compared with ours most rigidly. 'N/A' stands for 'not available'. Yellow shaded rows highlight that the methods use just the RGB-based input, whereas blue for just keypoint.

Among these methods, Zappardino *et al.* [58] use CNNs to learn group activity in Volleyball games, given sequence of person keypoint coordinates, their temporal differences, and keypoint differences from each actor to the pivot-actor that is selected by the model. The model does not model human-object interactions. AT [17] does not consider human-object interactions either, but because AT is also a Transformer-based model like ours, we can easily improve it by feeding our object tokens as additional inputs to AT. Moreoever, GIRN [39] and POGARS [46] are designed to leverage ball trajectory for learning group activity in videos of Volleyball games. As shown in Table 2, the object keypoint information can greatly boost the performance by providing additional context. GIRN models interactions between joints within an actor and across actors, as well as joint-object interactions. POGARS uses 1D CNNs to learn spatiotemporal dynamics of actors. AT, GIRN, and POGARS all use dot-product-based attention mechanisms similar to ours, however, they fail to fully model the hier-

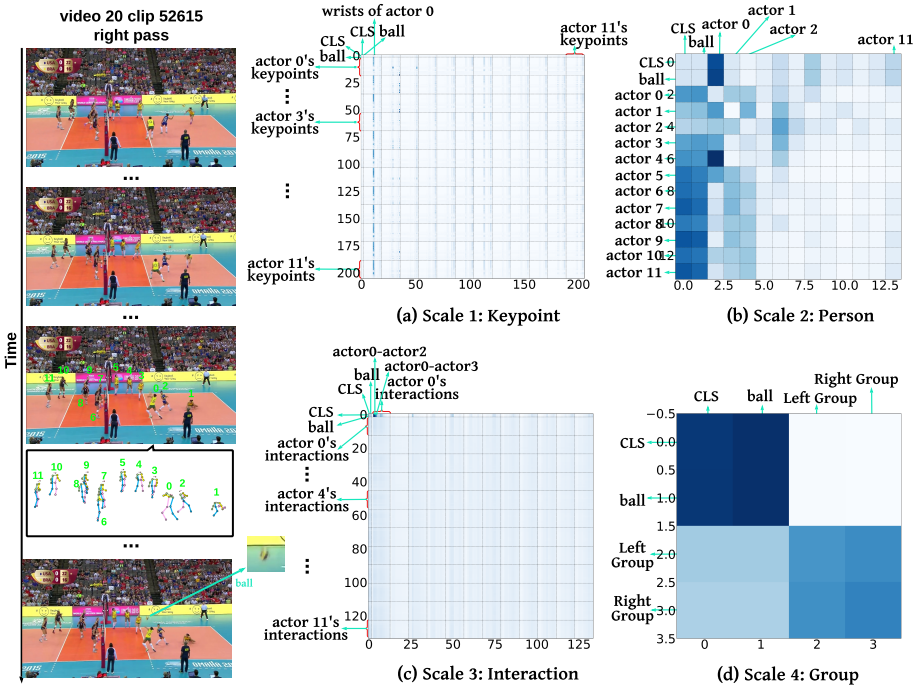

Fig. 5. Qualitative results of COMPOSER on VD – showcasing attention matrices of an instance in the "**right pass**" class (key actor is actor 0).

archical entities in the video (e.g., they all only use attention to learn person-wise importance, and at most consider two scales: keypoint and person), and more importantly, they lack explicit strategy to improve the multiscale representations in order to aid the compositional reasoning of group activity recognition.

Comparisons of Methods Using Other Modalities We compare results of COMPOSER with the best reported results of SOTA methods that use a single or multiple modalities in Table 3 on both VD and CAD. COMPOSER still achieves competitive performance – outperforming methods that use only RGB signals, obtaining +0.7% improvement on VD and +2.8% improvement on CAD if compared with methods that use a single modality (RGB or keypoint), and performing favorably compared with methods that exploit multiple expensive input modalities.

GroupFormer [31] has the highest accuracy on VD and CAD due to learning the representations of the multiscale scene entities (person and person group) with a Clustered Spatial-Temporal Transformer, and leveraging scene context and multiple *expensive* modalities (FLOPs: GroupFormer **595M** v.s. COMPOSER **297M**; details are in Appendix).

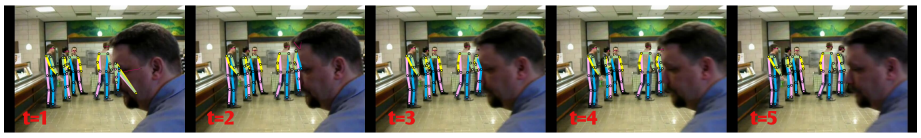

Fig. 6. Qualitative results on CAD (video ID '10'). COMPOSER successfully predicts 'Queueing' even when the input keypoints are partially noisy due to occlusion.

4.3 Qualitative Results

We visualize the attention weights in Fig. 5. We highlight the tokens that the model has mostly attended to at each scale (e.g., wrists of actor 0 at the person keypoint scale). COMPOSER is able to attend to relevant information across different scales, and it can produce interpretable results. In Fig. 6, we visualize the keypoint input to COMPOSER on a CAD instance. COMPOSER implicitly learns the human motion patterns from the keypoint features to handle partial occlusions.

Please check Appendix for more analyses including ablation studies, confusion matrices, parameter sensitivity analyses w.r.t. the number of scales and the number of prototypes, more qualitative results including failure cases, etc.

5 Conclusion

We propose COMPOSER that uses a Multiscale Transformer to learn compositional reasoning at different scales for group activity recognition. We also improve the intermediate representations using contrastive clustering, auxiliary prediction, and data augmentation techniques. We demonstrate the model's strength and interpretability on two widely-used datasets (Volleyball and Collective Activity). COMPOSER achieves up to +5.4% improvement with just the keypoint modality.

One limitation is that videos with severe occlusions remain challenging for COMPOSER like other existing methods, due to errors from detecting keypoints. Adopting 3D keypoints or stronger backbones that estimate keypoints directly from the video [38,1] can help to address the issue. Possible future directions include 1) expanding our methods to more complex scenarios, such as crowd understanding that may require modeling additional hierarchical scales; and 2) exploring effective multimodal fusion methods in order to use additional modalities like RGB but without suffering from scene biases, since RGB can be beneficial for activities that involve significant interaction with the background scene.

Acknowledgments The research was supported in part by NSF awards: IIS-1703883, IIS-1955404, IIS-1955365, RETTL-2119265, and EAGER-2122119. This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 22STESE00001 01 01. Disclaimer: The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

References

1. Mediapipe pose: Ml solution for high-fidelity body pose tracking from rgb video frames. `https://google.github.io/mediapipe/solutions/pose.html`
2. Abkenar, A.B., Loke, S.W., Zaslavsky, A., Rahayu, W.: Groupsense: recognizing and understanding group physical activities using multi-device embedded sensing. ACM Transactions on Embedded Computing Systems (TECS) **17**(6), 1–26 (2019)
3. Asano, Y.M., Patrick, M., Rupprecht, C., Vedaldi, A.: Labelling unlabelled videos from scratch with multi-modal self-supervision. arXiv preprint arXiv:2006.13662 (2020)
4. Azar, S.M., Atigh, M.G., Nickabadi, A., Alahi, A.: Convolutional relational machine for group activity recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7892–7901 (2019)
5. Azar, S.M., Atigh, M.G., Nickabadi, A., Alahi, A.: Convolutional relational machine for group activity recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7892–7901 (2019)
6. Bagautdinov, T., Alahi, A., Fleuret, F., Fua, P., Savarese, S.: Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4315–4324 (2017)
7. Blomqvist, K., Chung, J.J., Ott, L., Siegwart, R.: Semi-automatic 3d object keypoint annotation and detection for the masses. arXiv preprint arXiv:2201.07665 (2022)
8. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS) (2020)
9. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
10. Chen, B., Rouditchenko, A., Duarte, K., Kuehne, H., Thomas, S., Boggust, A., Panda, R., Kingsbury, B., Feris, R., Harwath, D., et al.: Multimodal clustering networks for self-supervised learning from unlabeled videos. arXiv preprint arXiv:2104.12671 (2021)
11. Choi, J., Gao, C., Messou, J.C., Huang, J.B.: Why can't i dance in a mall? learning to mitigate scene bias in action recognition. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. pp. 853–865 (2019)
12. Choi, W., Savarese, S.: Understanding collective activitiesof people from videos. IEEE transactions on pattern analysis and machine intelligence **36**(6), 1242–1257 (2013)
13. Choi, W., Shahid, K., Savarese, S.: What are they doing?: Collective activity classification using spatio-temporal relationship among people. In: 2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops. pp. 1282–1289. IEEE (2009)
14. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26**, 2292–2300 (2013)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

16. Ehsanpour, M., Abedin, A., Saleh, F., Shi, J., Reid, I., Rezatofighi, H.: Joint learning of social groups, individuals action and sub-group activities in videos. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 177–195. Springer (2020)
17. Gavrilyuk, K., Sanford, R., Javan, M., Snoek, C.G.: Actor-transformers for group activity recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 839–848 (2020)
18. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
19. Hinojosa, C., Niebles, J.C., Arguello, H.: Learning privacy-preserving optics for human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2573–2582 (2021)
20. Hu, G., Cui, B., He, Y., Yu, S.: Progressive relation learning for group activity recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 980–989 (2020)
21. Huang, Y., Kadav, A., Lai, F., Patel, D., Graf, H.P.: Learning higher-order object interactions for keypoint-based video understanding (2021)
22. Hudson, D., Manning, C.D.: Learning by abstraction: The neural state machine. Advances in Neural Information Processing Systems **32**, 5903–5916 (2019)
23. Ibrahim, M.S., Mori, G.: Hierarchical relational networks for group activity recognition and retrieval. In: Proceedings of the European conference on computer vision (ECCV). pp. 721–736 (2018)
24. Ibrahim, M.S., Mori, G.: Hierarchical relational networks for group activity recognition and retrieval. In: Proceedings of the European conference on computer vision (ECCV). pp. 721–736 (2018)
25. Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G.: A hierarchical deep temporal model for group activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1971–1980 (2016)
26. Jaiswal, A., Singh, S., Wu, Y., Natarajan, P., Natarajan, P.: Keypoints-aware object detection. In: NeurIPS 2020 Workshop on Pre-registration in Machine Learning. pp. 62–72. PMLR (2021)
27. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. arXiv preprint arXiv:2004.11362 (2020)
28. Koshkina, M., Pidaparthy, H., Elder, J.H.: Contrastive learning for sports video: Unsupervised player classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4528–4536 (2021)
29. Kulkarni, T.D., Gupta, A., Ionescu, C., Borgeaud, S., Reynolds, M., Zisserman, A., Mnih, V.: Unsupervised learning of object keypoints for perception and control. Advances in neural information processing systems **32** (2019)
30. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and brain sciences **40** (2017)
31. Li, S., Cao, Q., Liu, L., Yang, K., Liu, S., Hou, J., Yi, S.: Groupformer: Group activity recognition with clustered spatial-temporal transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13668–13677 (2021)
32. Lu, C., Koniusz, P.: Few-shot keypoint detection with uncertainty learning for unseen species. arXiv preprint arXiv:2112.06183 (2021)
33. Lu, L., Lu, Y., Yu, R., Di, H., Zhang, L., Wang, S.: Gaim: Graph attention interaction model for collective activity recognition. IEEE Transactions on Multimedia **22**(2), 524–539 (2019)

34. Luo, Z., Xie, W., Kapoor, S., Liang, Y., Cooper, M., Niebles, J.C., Adeli, E., Li, F.F.: Moma: Multi-object multi-actor activity parsing. Advances in Neural Information Processing Systems **34** (2021)
35. Nakatani, C., Sendo, K., Ukita, N.: Group activity recognition using joint learning of individual action recognition and people grouping. In: 2021 17th International Conference on Machine Vision and Applications (MVA). pp. 1–5. IEEE (2021)
36. Nebehay, G., Pflugfelder, R.: Consensus-based matching and tracking of keypoints for object tracking. In: IEEE Winter Conference on Applications of Computer Vision. pp. 862–869. IEEE (2014)
37. Ngiam, J., Caine, B., Vasudevan, V., Zhang, Z., Chiang, H.T.L., Ling, J., Roelofs, R., Bewley, A., Liu, C., Venugopal, A., et al.: Scene transformer: A unified architecture for predicting multiple agent trajectories. arXiv preprint arXiv:2106.08417 (2021)
38. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7753–7762 (2019)
39. Perez, M., Liu, J., Kot, A.C.: Skeleton-based relational reasoning for group activity analysis. Pattern Recognition p. 108360 (2021)
40. Pramono, R.R.A., Chen, Y.T., Fang, W.H.: Empowering relational network by self-attention augmented conditional random fields for group activity recognition. In: European Conference on Computer Vision. pp. 71–90. Springer (2020)
41. Qi, M., Qin, J., Li, A., Wang, Y., Luo, J., Van Gool, L.: stagnet: An attentive semantic rnn for group activity recognition. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 101–117 (2018)
42. Shu, T., Todorovic, S., Zhu, S.C.: Cern: confidence-energy recurrent network for group activity recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5523–5531 (2017)
43. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
44. Singh, K.K., Mahajan, D., Grauman, K., Lee, Y.J., Feiszli, M., Ghadiyaram, D.: Don't judge an object by its context: Learning to overcome contextual bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11070–11078 (2020)
45. Snower, M., Kadav, A., Lai, F., Graf, H.P.: 15 keypoints is all you need. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6738–6748 (2020)
46. Thilakarathne, H., Nibali, A., He, Z., Morgan, S.: Pose is all you need: The pose only group activity recognition system (pogars). arXiv preprint arXiv:2108.04186 (2021)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
48. Wang, M., Ni, B., Yang, X.: Recurrent modeling of interaction context for collective activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3048–3056 (2017)
49. Wu, J., Wang, L., Wang, L., Guo, J., Wu, G.: Learning actor relation graphs for group activity recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9964–9974 (2019)

50. Wu, L.F., Wang, Q., Jian, M., Qiao, Y., Zhao, B.X.: A comprehensive review of group activity recognition in videos. International Journal of Automation and Computing pp. 1–17 (2021)
51. Xu, D., Fu, H., Wu, L., Jian, M., Wang, D., Liu, X.: Group activity recognition by using effective multiple modality relation representation with temporal-spatial attention. IEEE Access **8**, 65689–65698 (2020)
52. Yan, R., Tang, J., Shu, X., Li, Z., Tian, Q.: Participation-contributed temporal dynamic model for group activity recognition. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 1292–1300 (2018)
53. Yan, R., Xie, L., Tang, J., Shu, X., Tian, Q.: Higcin: hierarchical graph-based cross inference network for group activity recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
54. Yan, R., Xie, L., Tang, J., Shu, X., Tian, Q.: Social adaptive module for weakly-supervised group activity recognition. In: European Conference on Computer Vision. pp. 208–224. Springer (2020)
55. Yang, Z., Liu, S., Hu, H., Wang, L., Lin, S.: Reppoints: Point set representation for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9657–9666 (2019)
56. Yuan, H., Ni, D.: Learning visual context for group activity recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 3261–3269 (2021)
57. Yuan, H., Ni, D., Wang, M.: Spatio-temporal dynamic inference network for group activity recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7476–7485 (2021)
58. Zappardino, F., Uricchio, T., Seidenari, L., Del Bimbo, A.: Learning group activities from skeletons without individual action labels. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 10412–10417. IEEE (2021)