COMPOSER: Compositional Learning of Group Activity in Videos

Honglu Zhou\textsuperscript{1,}\footnote{Work done as a NEC Labs intern.}, Asim Kadav\textsuperscript{2}, Aviv Shamsian\textsuperscript{3}, Shijie Geng\textsuperscript{1}, Farley Lai\textsuperscript{2}, Long Zhao\textsuperscript{1}
Ting Liu\textsuperscript{1}, Mubbasir Kapadia\textsuperscript{1}, Hans Peter Graf\textsuperscript{2}
\textsuperscript{1} Department of Computer Science, Rutgers University, Piscataway, NJ, USA
\textsuperscript{2} NEC Laboratories America, Inc., San Jose, CA, USA
\textsuperscript{3} Bar-Ilan University, Israel
\textsuperscript{4} Google LLC, Menlo Park, CA, USA
\{hz289, sgl1309, lz311, mk1353\}@cs.rutgers.edu
\{asim, farleylai, hpg\}@nec-labs.com
\{aviv.shamsian\}@biu.ac.il, \{liuti\}@google.com

Abstract

Group Activity Recognition (GAR) detects the activity performed by a group of actors in a short video clip. The task requires the compositional understanding of scene entities and relational reasoning between them. We approach GAR by modeling the video as a series of 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, we only use the keypoint modality which reduces scene biases and improves the generalization ability of the model. We improve the multi-scale 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 novel data augmentations (e.g., Actor Dropout) to aid model training. We demonstrate the model’s strength and interpretability on the challenging Volleyball dataset. COMPOSER achieves a new state-of-the-art 94.5\% accuracy with the keypoint-only modality. COMPOSER outperforms the latest GAR methods that rely on RGB signals, and performs favorably compared against methods that exploit multiple modalities. Our code will be available.

1. Introduction

The Group Activity Recognition (GAR) task detects the activity performed by a group of actors interacting with one other in a short video clip [16, 97]. GAR has wide-spread applications in sports analytics, robot-human interaction, and crowd behavior analysis [24, 26, 72, 104]. Compared to the single-person atomic action recognition task, GAR requires addressing two additional challenges. First, GAR requires a compositional understanding of the scene [1]. Because of the crowded scene, it is challenging to learn meaningful representations for GAR over the entire scene [97].
Since group activity often consists of one or more sub-groups of actors and scene objects, the final action label depends on a compositional understanding of these entities [97, 103]. Second, GAR benefits from relational reasoning over scene elements to understand the relative importance of entities and their interactions [36, 101]. For example, in a volleyball game of multiple actors, 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 [3, 6, 39, 41, 73, 79] or person sub-groups [24, 57, 67] for a compositional understanding of the group activity. Meanwhile, graph [36, 40, 96, 104] and transformer [26, 57] based models have been proposed for relational reasoning over scene entities. However, these methods fail to elaborately design compositional learning for GAR. They 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 [26, 36, 96, 104]) or two scales (e.g., person and sub-group [24, 57, 67]), or keypoint and person [71]). In addition, explicit multiscale modeling is neglected, lacking consistent compositional representations for the group action tasks. Furthermore, they often combine RGB, keypoint and optical flow modalities in order to reach competitive performance (Table 2). However, more modalities in use indicates a heavier burden for real-time applications [82]. More importantly, simply relying on the RGB modality hinders the model’s robustness to changes in background, lighting conditions or textures (Table 4), and often results in poor generalization due to scene biases [14, 83].

In this paper, we present COMPOSER (Fig. 3) 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 and reasoning their interactions to turn sensory signals into a high-level knowledge [37, 52], we approach GAR by modeling a video as tokens that represent the multi-scale semantic concepts in the video (Fig. 1). 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 the Multi-scale Transformer (Fig. 3), 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. 2D keypoints allow the model to focus on the action-specific cues and are more invariant to the scene biases. COMPOSER generalizes better to testing data with different scene background characteristics (Sec. 4.4).

Notably, COMPOSER learns consistent multiscale representations which boost the performance for GAR. 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 (Sec. 3.3). However, distinct from [8] and related work [2, 10], which use information from multiple views, we use information from multiple scales. (Fig. 2). Finally, we use techniques such as auxiliary prediction at each scale (Sec. 3.5) and data augmentation methods (Sec. 3.4) such as Actor Dropout to aid training.

Our contributions are summarized as follows:

- We present COMPOSER, a novel Multiscale Transformer based architecture for compositional learning 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.
- We only use the keypoint information that allows COMPOSER to be robust to changes in background, with auxiliary prediction and data augmentations to aid
learning group activity from the keypoint modality.

- We achieve a new state-of-the-art accuracy (94.5%) on the Volleyball dataset with the keypoint-only information. COMPOSER outperforms the latest methods that use RGB signals and performs favorably compared with methods that exploit multiple modalities.

2. Related Work

2.1. Group Activity Recognition

Early work on GAR relies on handcrafted features [13, 15, 17, 18, 31, 53, 66]; yet notable progress has been made in recent years by deep-learning (DL) based approaches [22, 41]. We review DL-based methods and refer readers to the comprehensive review of GAR presented in [97]. Early DL-based methods use Convolutional Neural Networks (CNNs) to extract features and then apply recurrent neural networks for temporal modeling [46, 58, 80, 95]. Since learning inter-person interactions is essential for GAR [97], much of the research explores how to capture the actor relations [4, 36, 40, 72, 96]. Several works tackle this problem from a graph-based perspective [40, 63, 100, 101] such as applying Graph Convolutional Networks (GCNs) [49, 96]. More recent works utilize attention modeling [63, 73, 98, 103] including using Transformers [26, 57] with a focus on determining the most critical persons [26, 72, 96, 103], groups [24, 57], or interactions [101]. Existing works have primarily use RGB- and optical-flow-based features with RoIAlign [33] to represent individuals [5, 73, 96, 100]. Recent works replace or augment these features with pose information from the individuals [11, 57, 62, 91]. Some only use numerical coordinate-based keypoint representation [71, 72, 91, 106] while others use a high-dimensional vector representation from a deep pose backbone [26, 103] which is not as efficient. In this paper, we use Transformers [92] for higher-order relationship modeling and use only the light-weight 2D keypoint representation. Our work differs from existing methods in that we propose a Multiscale Transformer block to hierarchically reason about entities at different semantic scales. In addition, we improve the musicale representations with a contrastive clustering based learning objective.

2.2. Action Recognition and Keypoint-based Prediction

Action Recognition is one of the primary tasks in video understanding. There has been rapid progress in recent years, starting from recognition of the low-level atomic actions performed by a single individual (e.g., hand-waving), to paired-actions being acted by two persons [50, 70, 81, 105] (e.g., shaking hands), towards group activities that encompass many actors at once [17, 41, 71, 75] (e.g., pedestrians queuing, attack and defense in a sports game). In addition, skeleton-based action recognition has drawn much attention [56, 69, 102, 108, 109], because dynamics of the human body skeleton convey critical information about human poses for human action recognition. Keypoint-based representation can be regarded as a high-level representation for human behavior, and is preferred due to additional benefits such as being compact and robust to variations of viewpoints, appearances, and surrounding distractions [20, 60]. We study keypoint-based group activity recognition. We propose to use multiple techniques including auxiliary prediction and data augmentations that can aid learning group activity from the keypoint modality.

2.3. Compositional Understanding and Multiscale Learning in Computer Vision (CV)

Compositionality is an active field of research in computer vision [47, 88, 99, 107], natural language processing [21, 42, 85, 93] and machine reasoning [7, 37, 38]. In terms of understanding videos centered on human actions, compositionality can be studied from different lenses, e.g., through formulating an activity in the video as compositions of atomic actions temporally [12, 29, 74] or semantically [77, 89], or decomposing actions by action-based aspects (verbs) and object components (noun) [43, 44, 54, 65] as a series of dynamic human-object interactions. Our paper tackles compositional video understanding by formulating a visual-semantic hierarchy, where each semantic hierarchy is regarded as representation of the video at a particular scale. Such an idea of multiscale learning has been a long-standing topic in CV as well [28, 30, 55, 56, 64, 94]. Recently, researchers have started to introduce the concept of multiscale learning to Transformers [25, 32, 61] by operating self-attention over various scales of resolutions and/or channels, in order to obtain a multiscale pyramid of features often observed in CNNs. Distinct from prior works, we design a Multiscale Transformer block that models semantic scene entities at different hierarchical scales to learn group activities effectively. COMPOSER is the first Transformer-based method with explicit multiscale modeling for GAR.

3. Methodology

We present COMPOSER (Fig. 3), a novel Multiscale Transformer based architecture for GAR. We model a video as tokens that represent the multiscale semantic concepts which include keypoint, person, interaction, group, clip, and object(s) if present (Fig. 1). In Sec. 3.1, we describe the initial encodings of these tokens before sending them into the reasoning module Multiscale Transformer (Sec. 3.2). We describe data augmentations in Sec. 3.4 and the exact formulation of auxiliary prediction in Sec. 3.5.
3.1. Hierarchical and Composite Representation of Semantic Entities

We form the tokens based on track-based representations [26, 57, 91, 106]. We describe initial representations of the tokens in a bottom-up manner, because representations of tokens in coarser scales are aggregated from that of finer scales, in order to form hierarchical representations.

**Keypoint.** Encoding of the keypoint token is crucial because keypoint is the finest-grained entity. We encode the numerical coordinates of keypoints (Fig. 4) [71, 84, 106]. 2D coordinate of a keypoint in frame \( t \) is \((x_t, y_t)\), we apply feature standardization to the raw coordinate and the temporal difference. We also normalize for rotations, translation, and scale differences in a person wise manner [87, 106]. To mitigate the issue of noisy estimated keypoints, we use the temporal Object Keypoint Similarity (OKS) proposed in [84], and use the mean OKS scores of each person as additional features. Inspired by recent advancement in positional encodings (PE), we use the Learned Fourier Positional Encoding [90] that maps each image coordinate into a learned vector. Learned Absolute Positional Encoding is applied to time coordinates. GCN-based Encoding [49] is proposed to encode the keypoint types and the intrinsic connections of keypoints. These features are concatenated to form the composite representation of a keypoint in a frame, and a keypoint track is represented by concatenation and
Feed Forward Network (FFN) based transformation of keypoints’ composite representations in all timestamps (Fig. 5). We define a keypoint token, \( k_p \in \mathbb{R}^d \) 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.

**Person.** A person token is defined as \( p_p \in \mathbb{R}^d \), initially obtained by aggregating the standardized joint coordinates of person \( p \) over time (Fig. 5).

**Person-to-Person Interaction.** Modeling the person-to-person interactions is critical for GAR [97]. Unlike existing works that typically consider an interaction as an edge connecting two person nodes and learn a scalar to depict its importance [101], we model interaction as nodes (tokens). The person-to-person interaction token is defined as \( i_p \in \mathbb{R}^d \) where \( i = 1, \ldots, p' \times (p'-1) \) (bi-directed interactions). Initial representation of the interaction between person \( p \) and \( q \) is learned from \( p_p \) and \( p_q \) (Fig. 5).

**Group.** We define the group token \( g_g \in \mathbb{R}^d \) where \( g = 1, \ldots, g' \) for videos where sub-groups are separable (e.g., clips of volleyball games). \( g' \) denotes the num. of subgroups in the video. Given the person-to-group mapping which can be obtained through various mechanisms (e.g., heuristics, k-means [57, 106], etc [24, 51]), representation of a group is an aggregate over representations of persons in the group similarly through concatenation and FFN (Fig. 5).

**Clip.** The special [CLS] token \( \ell \in \mathbb{R}^d \) 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 [23].

**Object.** Scene objects play a crucial role in GAR. 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 just the keypoint information due to their similar poses. Ball location can help to distinguish the key person. Thus, we use object token \( e_w \in \mathbb{R}^d \) when applicable, \( e = 1, \ldots, e' \) and \( e' \) is the number of special objects. For the Volleyball dataset [41], ball trajectories from [71] are utilized \((e'=1)\). The initial ball token is learned similar to the keypoint token, by concatenating time PE, coordinate PE, standardized ball trajectories from [71] are utilized \((e'=1)\). The initial ball token is learned similar to the keypoint token, by concatenating time PE, coordinate PE, standardized ball 2D coordinate and temporal difference, then aggregate representations of ball locations over time followed by a FFN.

### 3.2. Multiscale Transformer

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

- **Scale 1:** \( \{\text{CLS}, e_1, \ldots, e_{e'}, k_1, \ldots, k_{p'}\} \),
- **Scale 2:** \( \{\text{CLS}, e_1, \ldots, e_{e'}, p_1, \ldots, p_{p'}\} \),
- **Scale 3:** \( \{\text{CLS}, e_1, \ldots, e_{e'}, i_1, \ldots, i_{p' \times (p'-1)}\} \),
- **Scale 4:** \( \{\text{CLS}, e_1, \ldots, e_{e'}, g_1, \ldots, g_{g'}\} \).

We utilize a Transformer encoder [92] at each scale to perform relational reasoning of tokens in that scale (Fig. 3). We review details of the Transformer encoder in the Appendix.

Hierarchical representations of tokens are maintained in an elaborately designed Multiscale Transformer block. In the Multiscale Transformer block, operations in the four scales are the same (but with different parameters) to maintain simplicity and elegance. 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 \( s+1 \).

### 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. 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 (Fig. 3) to capture an agreement of the common semantic information hidden across the scales.

**Preliminaries.** Suppose \( v_{n,s} \in \mathbb{R}^d \) represents the learned representation of clip \( n \) in scale \( s \), where \( s \in \{1, 2, 3, 4\} \).

Following prior works [8, 45], we first project the representation to the unit sphere. We then compute a code (i.e., cluster assignment) \( q_{n,s} \in \mathbb{R}^K \) by mapping \( v_{n,s} \) to a set of \( K \) trainable prototype vectors, \( \{c_1, \ldots, c_K\} \). We denote by \( C \in \mathbb{R}^{K \times d} \) the matrix whose rows are the \( c_1, \ldots, c_K \).

**Swapped Prediction.** Suppose \( s \) and \( w \) denote 2 different scales from the four representation scales. The swapped prediction problem aims to predict the code \( q_{n,s} \) from \( v_{n,w} \), and \( q_{n,w} \) from \( v_{n,s} \), with the following loss function:

\[
\mathcal{L}_{\text{swap}}(v_{n,w}, v_{n,s}) = \ell(v_{n,w}, q_{n,s}) + \ell(v_{n,s}, q_{n,w}) \tag{2}
\]

where \( \ell(v_{n,w}, q_{n,s}) \) measures the fit between \( v_{n,w} \) and \( q_{n,s} \). Specifically, \( \ell(v_{n,w}, q_{n,s}) \) is the cross entropy loss between \( q_{n,s} \) and the probability obtained by taking a softmax of the dot products of \( v_{n,w} \) and prototypes in \( C \):

\[
\ell(v_{n,w}, q_{n,s}) = -\sum_{k=1}^{K} q_{n,s}^{(k)} \log \frac{\exp\left(\frac{1}{\tau} v_{n,w}^{(k)} c_k\right)}{\sum_{k'=1}^{K} \exp\left(\frac{1}{\tau} v_{n,w}^{(k')} c_{k'}\right)} \tag{3}
\]

where \( \tau \) 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:

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

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

where the trace $\Tr$ is the sum of the elements on the main diagonal, $H$ is the entropy function, and $\varepsilon$ is a parameter that controls the smoothness of the mapping. $1_K \in \mathbb{R}^K$ and $1_B \in \mathbb{R}^B$ are a vector of ones to enforce the equi-partition constraint. The continuous solution $Q^*$ of Eq.(5) is computed with the iterative Sinkhorn-Knopp algorithm [8, 19].

3.4. Data Augmentation

We use data augmentations to 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 [68] 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 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.

3.5. Auxiliary Prediction

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:

$$
L_{\text{total}} = \sum_{m=1}^{M-1} \lambda (L_{\text{group}} + L_{\text{person}} + L_{\text{cluster}})
$$

where $L_{\text{group}}$ and $L_{\text{person}}$ are the cross entropy loss, $m$ is the index of the Multiscale Transformer block, $M$ is the total number of the Multiscale Transformer blocks, and $\lambda$ 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 [41] is a commonly-used GAR dataset, consisting of clips from 55 videos of volleyball games (39 for training and 16 for testing). 3,493 training clips and 1,337 testing clips. Each clip has 41 frames. Annotations include the group activity, players’ bounding boxes and their actions provided only for the middle frame of the clip. We use the common approach of using the person tracklet data provided by [76]. 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 individual actions: blocking, digging, falling, jumping, moving, setting, spiking, standing and waiting. Please refer to Appendix for implementation details to reproduce our results.

4.2. Comparison with State-of-the-Art (SOTA) Keypoint-only Modality. We compare COMPOSER with SOTA methods that use only the keypoint input modality in Table 1. COMPOSER outperforms all of these methods, achieving a new SOTA 94.5% accuracy. Jointly learning group activity with person actions helps, and the additional information of the ball keypoint can greatly boost the performance of methods that do not rely on image features.

| Method | Pose | Ball | Person Label | Keypoint Estimator | Acc. ↑ |
|--------|------|------|--------------|--------------------|-------|
| Zappardino et al. [106] | | | | OpenPose | 91.0 |
| GIRN [71] | | | | OpenPose | 88.4 |
| Actor-Transformer [30] | | | | HRNet | 92.3 |
| POGARS [91] | | | | Hourglass | 92.1 |
| COMPOSER (ours) | | | | HRNet | 92.4 |
| | | | | HRNet | 93.4 |
| | | | | HRNet | 94.5 |

Table 1. Comparisons with SOTA techniques that leverage only keypoint information. “Ball” denotes ball keypoints and “Person Label” means training with person action labels. COMPOSER outperforms existing methods and achieves a new SOTA result.
performing methods that use RGB signals, and performing favorably compared with methods that exploit multiple expensive input modalities (RGB, optical flow, keypoint, etc.).

| Method      | Pose | Ball | RGB | Flow | Scene | Acc. † |
|-------------|------|------|-----|------|-------|--------|
| SBGAR [59]  | ✓    | ✓    | ✓   | ✓    |       | 66.9   |
| HDTM [41]   | ✓    |      | ✓   |      |       | 81.9   |
| CERN [79]   |       |      |     |      |       | 83.3   |
| stagNet [73]| ✓    |      |     |      |       | 89.3   |
| RCRG [39]   | ✓    |      |     |      |       | 89.5   |
| SSU [6]     | ✓    |      |     |      |       | 89.9   |
| PRL [36]    | ✓    |      |     |      |       | 91.4   |
| HGCN [100]  | ✓    | ✓    |     | ✓    |       | 91.5   |
| ARG [91]    | ✓    |      | ✓   |      |       | 92.6   |
| CRM [3]     | ✓    |      | ✓   |      |       | 93.0   |
| Ebsanpour et al. [24] | ✓ |      | ✓   |      |       | 93.1   |
| DIN [104]   | ✓    | ✓    |     |      |       | 93.6   |
| GIRN [71]   | ✓    |      | ✓   |      |       | 94.0   |
| Actor-Transformer [26] | ✓ | ✓    |     |      |       | 94.4   |
| TCE+STBiP [103] | ✓ | ✓    |     |      | ✓     | 94.7   |
| Pramono et al. [72] | ✓ | ✓    |     |      | ✓     | 95.0   |
| GroupFormer [57] | ✓ | ✓    |     |      | ✓     | 95.7   |
| COMPOSER (ours) | ✓ | ✓    | ✓   | ✓    |       | 94.5   |

Table 2. Comparisons with SOTA methods that use RGB or multiple modalities. “Flow” denotes optical flow input, and “Scene” denotes features of the entire scene/frame. Fewer modalities indicates a stronger capability of the model itself (fewer checks are better). COMPOSER, with only the keypoint modality, outperforms the latest methods that use RGB signals, and performs favorably compared with methods that exploit multiple modalities.

4.3. Ablation Study

We conduct ablation experiments to verify the effectiveness of proposed techniques; results are in Table 3.

Contrastive clustering and scale agreement regularize the multiscale representations. As demonstrated in Table 3, performance drops without the contrastive clustering learning. The swapped prediction setup helps the model to maintain consistency across representations of the multiple scales of the same clip, which regularizes the intermediate representations. We also experiment with the ‘Label Consistency’ method [78] that minimizes the $L_2$ distance between 2 views of an instance in the logit space. Replacing contrastive clustering with Label Consistency for scale agreement, result is better than the previous ablation, but worse than COMPOSER. Better performance of COMPOSER can be attributed to the additional benefits of the clustering loss, which draws clips that are semantically related close together by comparing with the prototypes. Both experiments indicate that encouraging scale agreement can bring benefits for the multi-scale learning models.

Greater number of scales yields more information and an effective scale agreement. We find that increasing the number of scales leads to a higher accuracy. More scales indicates more information about the entities in the scene.

Besides, given more scales, hierarchical representations are better maintained, and techniques such as contrastive clustering and auxiliary predictions are more effective.

Data augmentations increase the training data size and inject benign noises, leading to generalization. Results in Table 3 show the gains brought by each data augmentation technique described in Sec. 3.4. Among the 4 types of data augmentation, Horizontal Flip (which is commonly used by existing works [91, 106]) and Actor Dropout are the most critical ones. Even though data augmentation is less effective than other techniques proposed, it increases the training data size and injects noises that help model to generalize.

Auxiliary prediction aids learning the intermediate representations. For the ‘No Auxiliary Prediction’ ablation, the loss of group activity is computed from the clip token from the last scale of the last Multiscale Transformer (note that person action loss and clustering loss are still in use). Performance of this ablation largely drops, indicating auxiliary prediction is a simple yet effective technique.

Multiscale Transformer learns higher-level knowledge of the video by compositionally reasoning over concepts from finest-grained to coarsest-grained. We design an ablation to remove Multiscale Transformer. Specifically, the group activity classifier directly takes features of the ball token and person tokens (learned from features of keypoint tokens) as an input, and the person action classifier takes the features of person tokens as an input. The outcome of this ablation is worse performance than the 1-scale ablation that does not even use person features and person action labels – which indicates the importance of relational reasoning. Given that this ablation has a reasonably high accuracy without any reasoning module, it suggests that our initial keypoint representations are of high quality (Sec. 3.1).

4.4. Generalization Capability of COMPOSER

Olympic Split. Thilakarathne et al. [91] released a skewed Olympic split of the Volleyball dataset 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 largely differs from the train videos in scene background.

We examine the generalization capability of COMPOSER on this Olympic split, and compare our result with the ones reported in [91] in Table 4. The accuracy of COMPOSER on this skewed split is higher than the accuracy on the original split. In contrast, the 13D model [9] trained using RGB images has a very large drop in accuracy. POGARS [91] is a method that uses only the keypoint information; result of POGARS only dropped slightly on the skewed data. Our results imply that training a model using only keypoint information can reduce scene biases, and generalize better than RGB-based approaches to testing data with different characteristics by ignoring background-related factors.

### 4.5. Analysis

**Confusion Matrix.** We present the confusion matrix on the Volleyball dataset (original split) in Fig. 6. For each class COMPOSER achieves an accuracy over 90% with the lowest accuracy for the right set class. Most failures emerge from distinguishing the set, pass, and spike activities which can be a result of highly similar actions or positions of the key player in some clips. Occasionally, the model struggles to distinguish which team (left or right) performs the activity.

**Attention Matrix.** We also visualize the attention weights from the last Multiscale Transformer block in COMPOSER in Fig. 6. Observations are: Scale 1: tokens mostly attend to wrist, knee, and ankle of actor 0 who is the key player performing setting in the right team. Lower body of actor 0 is most attended to might because of less accurate estimation of the wrist and elbow keypoint due to quick motion and blurry image. Scale 2: CLS, ball, and actors in the right team mostly attend to actor 0, whereas actors in the left team mostly attend to CLS and ball. Scale 3: tokens mostly attend to interactions leading by actor 0 to the nearby actors in the same team. Scale 4: CLS and Ball mostly attend to each other, and two teams mostly attend to each other. COMPOSER is able to attend to relevant information across different scales, and it can produce interpretable results.
5. Conclusion

We propose COMPOSER that uses a Multiscale Transformer to build compositional and relational reasoning at different scales. We improve the intermediate representations using contrastive clustering, auxiliary prediction, and data augmentation techniques. We find that COMPOSER is able to attend to relevant information across different scales and achieves SOTA performance on the Volleyball dataset using only keypoint input modality.

In the future, we plan to expand our methods to more complex scenarios, such as crowd understanding that may require modeling additional hierarchical scales. We also plan to use additional modalities like RGB which can be beneficial for activities which involve significant interaction with objects or the background scene.

References

[1] Amin B Abkenar, Seng W Loke, Arkady Zaslavsky, and Wenny Rahayu. Groupsense: recognizing and understanding group physical activities using multi-device embedded sensing. ACM Transactions on Embedded Computing Systems (TECS), 17(6):1–26, 2019.

[2] Yuki M Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. arXiv preprint arXiv:2006.13662, 2020.

[3] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional relational machine for group activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7892–7901, 2019.

[4] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional relational machine for group activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7892–7901, 2019.

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

[6] Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4315–4324, 2017.

[7] Léon Bottou. From machine learning to machine reasoning. Machine learning, 94(2):133–149, 2014.

[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS), 2020.

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6299–6308, 2017.

[10] Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Anggie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, et al. Multimodal clustering networks for self-supervised learning from unlabeled videos. arXiv preprint arXiv:2104.12671, 2021.

[11] Hsing-Yu Chen and Shang-Hong Lai. Group activity recognition via computing human pose motion history and collective map from video. In Asian Conference on Pattern Recognition, pages 705–718. Springer, 2019.

[12] Junwen Chen, Haiting Hao, Hanbin Hong, and Yu Kong. Rit-18: A novel dataset for compositional group activity understanding. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, pages 853–865, 2019.

[13] Zhongwei Cheng, Lei Qin, Qingming Huang, Shuqiang Jiang, and Qi Tian. Group activity recognition by gaussian processes estimation. In 2010 20th International Conference on Pattern Recognition, pages 3228–3231. IEEE, 2010.

[14] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. 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, pages 853–865, 2019.

[15] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In European Conference on Computer Vision, pages 215–230. Springer, 2012.

[16] Wongun Choi and Silvio Savarese. Understanding collective activities of people from videos. IEEE transactions on pattern analysis and machine intelligence, 36(6):1242–1257, 2013.

[17] Wongun Choi, Khuram Shahid, and Silvio Savarese. 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, pages 1282–1289. IEEE, 2009.

[18] Wongun Choi, Khuram Shahid, and Silvio Savarese. Learning context for collective activity recognition. In CVPR 2011, pages 3273–3280. IEEE, 2011.

[19] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems, 26:2292–2300, 2013.

[20] L Minh Dang, Kyungbok Min, Hanxiang Wang, Md Jail Piran, Cheol Hee Lee, and Hyeonjoon Moon. Sensor-based and vision-based human activity recognition: A comprehensive survey. Pattern Recognition, 108:107561, 2020.

[21] Verna Dankers, Elia Bruni, and Diewukw Hupkes. The paradox of the compositionality of natural language: a neural machine translation case study. arXiv preprint arXiv:2108.05885, 2021.

[22] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4772–4781, 2016.
[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018. 5

[24] Mahsa Ehsanpour, Alireza Abedin, Fatemeh Saleh, Javen Shi, Ian Reid, and Hamid Rezatofighi. 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, pages 177–195. Springer, 2020. 1, 2, 3, 5, 7

[25] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. arXiv preprint arXiv:2104.11227, 2021. 3

[26] Kirill Gavriluk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 839–848, 2020. 1, 2, 3, 4, 6, 7, 21

[27] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Proceedings of the fourteenth international conference on artificial intelligence and statistics, pages 315–323. JMLR Workshop and Conference Proceedings, 2011. 20

[28] Zhiqiang Gong, Ping Zhong, Yang Yu, Weidong Hu, and Shutoo Li. A cnn with multiscale convolution and diversified metric for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 57(6):3599–3618, 2019. 3

[29] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agga: A benchmark for compositional spatio-temporal reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11287–11297, 2021. 3

[30] Eldad Haber, Lars Ruthotto, Elliot Holtham, and Seong-Hwan Jun. Learning across scales—multiscale methods for convolution neural networks. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018. 3

[31] Hossein Hajimirsadeghi, Wang Yan, Arash Vahdat, and Greg Mori. Visual recognition by counting instances: A multi-instance cardinality potential kernel. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2596–2605, 2015. 3

[32] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. arXiv preprint arXiv:2103.00112, 2021. 3, 19

[33] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017. 3

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 19

[35] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415, 2016. 20

[36] Guyue Hu, Bo Cui, Yuan He, and Shan Yu. Progressive relation learning for group activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 980–989, 2020. 2, 3, 7

[37] Drew Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. Advances in Neural Information Processing Systems, 32:5903–5916, 2019. 2, 3

[38] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In International Conference on Learning Representations, 2018. 3

[39] Mostafa S Ibrahim and Greg Mori. Hierarchical relational networks for group activity recognition and retrieval. In Proceedings of the European conference on computer vision (ECCV), pages 721–736, 2018. 2, 7

[40] Mostafa S Ibrahim and Greg Mori. Hierarchical relational networks for group activity recognition and retrieval. In Proceedings of the European conference on computer vision (ECCV), pages 721–736, 2018. 2, 3

[41] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1971–1980, 2016. 2, 3, 5, 6, 7

[42] Ozan Irsoy and Claire Cardie. Deep recursive neural networks for compositionality in language. In Advances in neural information processing systems, pages 2096–2104, 2014. 3

[43] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10236–10247, 2020. 3

[44] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-chun Zhu. Lemma: A multi-view dataset for learning multi-agent multi-task activities. In European Conference on Computer Vision, pages 767–786. Springer, 2020. 3

[45] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Philipp Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. arXiv preprint arXiv:2004.11362, 2020. 5

[46] Pil-Soo Kim, Dong-Gyu Lee, and Seong-Whan Lee. Discriminative context learning with gated recurrent unit for group activity recognition. Pattern Recognition, 76:149–161, 2018. 3

[47] Tae Soo Kim and Gregory D Hager. Saffcar: Structured attention fusion for compositional action recognition. arXiv preprint arXiv:2012.02109, 2020. 3

[48] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 22

[49] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016. 3, 4

[50] Yu Kong, Yunde Jia, and Yun Fu. Learning human interaction by interactive phrases. In European conference on computer vision, pages 300–313. Springer, 2012. 3
[51] Maria Koshkina, Hemanth Pidaparthy, and James H Elder. Contrastive learning for sports video: Unsupervised player classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4528–4536, 2021. 5

[52] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. Behavioral and brain sciences, 40, 2017. 2

[53] Tian Lan, Yang Wang, Weilong Yang, Stephen N Robinvitch, and Greg Mori. Discriminative latent models for recognizing contextual group activities. IEEE transactions on pattern analysis and machine intelligence, 34(8):1549–1562, 2011. 3

[54] Hung Le, Nancy F Chen, and Steven CH Hoi. C³: Compositional counterfactual constrastive learning for video-grounded dialogues. arXiv preprint arXiv:2106.08914, 2021. 3

[55] Guanbin Li and Yizhou Yu. Visual saliency detection based on multiscale deep cnn features. IEEE transactions on image processing, 25(11):5012–5024, 2016. 3

[56] Maosen Li, Siheng Chen, Yangzheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 214–223, 2020. 3

[57] Shuaicheng Li, Qianggang Cao, Lingbo Liu, Kunlin Yang, Shihan Liu, Jun Hou, and Shuai Yi. Groupformer: Group activity recognition with clustered spatial-temporal transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 13668–13677, 2021. 2, 3, 4, 5, 7

[58] Xin Li and Mooi Choo Chua. Sbgar: Semantics based group activity recognition. In Proceedings of the IEEE international conference on computer vision, pages 2876–2885, 2017. 3

[59] Xin Li and Mooi Choo Chua. Sbgar: Semantics based group activity recognition. In Proceedings of the IEEE international conference on computer vision, pages 2876–2885, 2017. 7

[60] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In Proceedings of the 28th ACM International Conference on Multimedia, pages 2490–2498, 2020. 3

[61] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030, 2021. 3

[62] Lihua Lu, Huijun Di, Yao Lu, Lin Zhang, and Shunzhou Wang. Spatio-temporal attention mechanisms based model for collective activity recognition. Signal Processing: Image Communication, 74:162–174, 2019. 3

[63] Lihua Lu, Yao Lu, Ruizhe Yu, Huijun Di, Lin Zhang, and Shunzhou Wang. Gain: Graph attention interaction model for collective activity recognition. IEEE Transactions on Multimedia, 22(2):524–539, 2019. 3

[64] Qianli Ma, Lifeng Shen, Enhuan Chen, Shuai Tian, Jiabing Wang, and Garrison W Cottrell. Walking walking walking: Action recognition from action echoes. In IJCAI, pages 2457–2463, 2017. 3

[65] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something–else: Compositional action recognition with spatial-temporal interaction networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1049–1059, 2020. 3

[66] Moin Nabi, Alessio Bue, and Vittorio Murino. Temporal posets for collective activity detection and recognition. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 500–507, 2013. 3

[67] Chihiro Nakatani, Kohei Sendo, and Norimichi Urita. Group activity recognition using joint learning of individual action recognition and people grouping. In 2021 17th International Conference on Machine Vision and Applications (MVA), pages 1–3. IEEE, 2021. 2

[68] Jiquan Ngiam, Benjamin Caine, Vijay Vasudevan, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Vengopal, et al. Scene transformer: A unified architecture for predicting multiple agent trajectories. arXiv preprint arXiv:2106.08417, 2021. 6

[69] Xuan Son Nguyen. Geomnet: A neural network based on riemannian geometries of spd matrix space and cholesky space for 3d skeleton-based interaction recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 13379–13389, 2021. 3

[70] Mauricio Perez, Jun Liu, and Alex C Kot. Interaction residual network for mutual action recognition. IEEE Transactions on Multimedia, 2021. 3

[71] Mauricio Perez, Jun Liu, and Alex C Kot. Skeleton-based relational reasoning for group activity analysis. Pattern Recognition, page 108360, 2021. 2, 3, 4, 5, 6, 7, 21

[72] 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 European Conference on Computer Vision, pages 71–90. Springer, 2020. 1, 3, 7

[73] Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. Stagnet: An attentive semantic rnn for group activity recognition. In Proceedings of the European Conference on Computer Vision (ECCV), pages 101–117, 2018. 2, 3, 7

[74] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11184–11193, 2021. 3

[75] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gornb, Kevin Murphy, and Li Fei-Fei. Detecting events and key actors in multi-person videos. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3043–3053, 2016. 3
[76] Kohei Sendo and Norimichi Ukitani. Heatmapping of people involved in group activities. In 2019 16th International Conference on Machine Vision Applications (MVA), pages 1–6. IEEE, 2019.

[77] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2616–2625, 2020.

[78] Changhao Shi, Chester Holtz, and Gal Mishne. Online adversarial purification based on self-supervised learning. In International Conference on Learning Representations, 2020.

[79] Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. Cern: confidence-energy recurrent network for group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5523–5531, 2017.

[80] Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. Cern: confidence-energy recurrent network for group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5523–5531, 2017.

[81] Xiangbo Shu, Jinhui Tang, Guojun Qi, and Tianmin Shu. Social adaptive module for weakly-supervised group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5998–6008, 2017.

[82] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv:1406.2199, 2014.

[83] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepthi Ghadiyaram. 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, pages 11070–11078, 2020.

[84] Michael Snover, Asim Kadav, Farley Lai, and Hans Peter Graf. 15 keypoints is all you need. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6738–6748, 2020.

[85] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association for Computational Linguistics, 2:207–218, 2014.

[86] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1):1929–1958, 2014.

[87] Jennifer J Sun, Jiapiing Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In European Conference on Computer Vision, pages 53–70. Springer, 2020.

[88] Pengzhan Sun, Bo Wu, Xunsong Li, Wen Li, Lixin Duan, and Chuang Gan. Counterfactual debiasing inference for compositional action recognition. In Proceedings of the 29th ACM International Conference on Multimedia, pages 3220–3228, 2021.

[89] Didac Sursí, Ruoshi Liu, and Carl Vondrick. Learning the predictability of the future. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12607–12617, 2021.

[90] Matthew Tancik, Pratul P Srinivasan, Ben Meldenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. arXiv preprint arXiv:2006.10739, 2020.

[91] Haritha Thilakarathne, Aiden Nibali, Zhen He, and Stuart Morgan. Pose is all you need: The pose only group activity recognition system (pogars). arXiv preprint arXiv:2108.04186, 2021.

[92] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

[93] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. arXiv preprint arXiv:1511.06361, 2015.

[94] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 2020.

[95] Minsi Wang, Binghong Ni, and Xiaokang Yang. Recurrent modeling of interaction context for collective activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3048–3056, 2017.

[96] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9964–9974, 2019.

[97] Li-Fang Wu, Qi Wang, Meng Jian, Yu Qiao, and Bo-Xuan Zhao. A comprehensive review of group activity recognition in videos. International Journal of Automation and Computing, pages 1–17, 2021.

[98] Dezhong Xu, Heng Fu, Lifang Wu, Meng Jian, Dong Wang, and Xu Liu. Group activity recognition by using effective multiple modality relation representation with temporal-spatial attention. IEEE Access, 8:65689–65698, 2020.

[99] Rui Yan, Lingxi Xie, Xiangbo Shu, and Jinhui Tang. Interactive fusion of multi-level features for compositional activity recognition. arXiv preprint arXiv:2012.05689, 2020.

[100] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Higcin: hierarchical graph-based cross inference network for group activity recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[101] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Social adaptive module for weakly-supervised group
activity recognition. In *European Conference on Computer Vision*, pages 208–224. Springer, 2020. 2, 3, 5

[102] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 3

[103] Hangjie Yuan and Dong Ni. Learning visual context for group activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3261–3269, 2021. 2, 3, 7, 21

[104] Hangjie Yuan, Dong Ni, and Mang Wang. Spatio-temporal dynamic inference network for group activity recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7476–7485, 2021. 1, 2, 7, 21

[105] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35. IEEE, 2012. 3

[106] Fabio Zappardino, Tiberio Uricchio, Lorenzo Seidenari, and Alberto Del Bimbo. Learning group activities from skeletons without individual action labels. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10412–10417. IEEE, 2021. 3, 4, 5, 6, 7

[107] Yibing Zhan, Jun Yu, Ting Yu, and Dacheng Tao. Multi-task compositional network for visual relationship detection. *International Journal of Computer Vision*, 128(8):2146–2165, 2020. 3

[108] Long Zhao, Yuxiao Wang, Jiaping Zhao, Liangzhe Yuan, Jennifer J Sun, Florian Schroff, Hartwig Adam, Xi Peng, Dimitris Metaxas, and Ting Liu. Learning view-disentangled human pose representation by contrastive cross-view mutual information maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12793–12802, 2021. 3

[109] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016. 3
Appendix

This appendix is organized as follows:
A. Ablation study with different number of prototypes for the contrastive clustering algorithm.
B. Additional qualitative results.
C. Failure cases of COMPOSER.
D. Discussion of limitations of COMPOSER.
E. Method and implementation details.

A. Experimental Results of COMPOSER with Different Number of Prototypes.

In Table 5, we evaluate the impact of the number of prototypes $K$ (i.e., the number of clip clusters) that is used for contrastive clustering learning on the GAR accuracy of COMPOSER. We use the original split of the Volleyball dataset for this evaluation. We observe that varying the number of prototypes does not affect much the performance. The performance first improves as the number of prototypes increases, then decreases as the number of prototypes keeps increasing. The number of prototypes has little influence as long as it is “enough” [8]. The practice is to set the number of prototypes at least one order of magnitude larger than the true number of classes in the dataset [8].

| Number of Prototypes | 10  | 50  | 100 | 1,000 | 10,000 |
|----------------------|-----|-----|-----|-------|--------|
| GAR Accuracy (%)     | 94.02 | 94.54 | 94.17 | 94.54 | 94.32 |

Table 5. Impact of the number of prototypes. GAR accuracy of COMPOSER on the original split of the Volleyball dataset using different number of prototypes. The number of prototypes has little influence on the performance of COMPOSER as long as it is “enough” [8].

![Figure 7. t-SNE visualizations on the Volleyball dataset show that the clip embedding space learned by COMPOSER using different number of prototypes: (a) 10 prototypes, and (b) 1000 prototypes. Best viewed in color. More number of prototypes can lead to a better separation of the clips in distinct group activity classes.](image)

We visualize the clip embedding space (after t-SNE 2D projection) learned by COMPOSER using 10 prototypes and 1000 prototypes in Fig. 7 (a) and (b), respectively. We take the representation of the CLS token from the last scale and the last Multiscale Transformer block as the representation of the clip to produce the embedding space visualization. In Fig. 7, each dot represents a test clip and the color of the dot indicates the group activity label of the clip. A higher number of prototypes can lead to a better grouping of clips with the same group activity class, as well as a better separation of clips in different group activity classes; this accords with the quantitative results shown in Table 5.

B. Additional Qualitative Results.

In Sec.4.5 of the main paper, we visualize the attention matrices produced by the last Multiscale Transformer block in COMPOSER of a test clip in group activity class “set” (more specifically, “right set”). We provide the same visualization for the other three main group activity classes in the Volleyball dataset, “pass”, “spike”, and “winpoint”, in Fig. 8, Fig. 9 and Fig. 10, respectively. Please refer to the caption of each figure for a detailed explanation of the visualization. Here, we summarize the main findings:

1. At scale 1 (keypoint scale), scale 2 (person scale) and scale 3 (person-to-person interaction scale), actor-related tokens associated with the key person(s) are often identified as the most important tokens for the CLS and ball tokens by COMPOSER.

2. For all test clips from the Volleyball dataset, we find that at scale 4 – the group scale, the CLS and ball tokens mostly attend to each other, and the two teams mostly attend to each other. We posit that this is because, at the previous three scales, the CLS and ball tokens mostly attend to actor-related tokens. The ball token, which contains less correlated information, then becomes as the most important signal at scale 4 for the CLS token in order to correctly recognize the group activity.

3. Wrists, knees and ankles are often identified by COMPOSER as the most relevant person keypoint joint types, among a total of 17 different keypoint types: (nose, left eye, right eye, left ear, right ear, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle and right ankle).

4. At scale 3 – the interaction scale, the pattern of the attention weights of the “winpoint” class is quite different from the pattern observed in the other three main group activity classes (“set”, “pass”, and “spike”). Recognition of the other group activities in the Volleyball dataset (e.g., “spike”), is often determined by a key player who is performing the key action (e.g., “spiking”). In contrast, the “winpoint” group activity is not heavily dependent on one or two key players, instead, it is defined by the overall person-to-person interactions of the team that just scored. COMPOSER learns to recognize the group activities based upon the unique characteristics of each group activity class.
At the keypoint scale, wrists of actor 0 and actor 3 are the keypoints that are most attended to by other tokens. Note that actor 0 is the player who is performing the key action. At the group scale, the CLS and ball tokens mostly attend to actor 0. Other person tokens mostly attend to the CLS, ball, actor 0, 1 and 2. At the interaction scale, the CLS and ball tokens mostly attend to the interactions leading by actor 0 to actor 2 and 3, i.e., interactions (actor 0, actor 2) and (actor 0, actor 3).

Figure 8. Qualitative results of COMPOSER on the Volleyball dataset – showcasing attention matrices (from the last Multiscale Transformer block) of a clip instance in the “right pass” class. Darker color denotes larger attention weights. (a) Scale 1: At the keypoint scale, wrists of actor 0 are the keypoints that are most attended to by other tokens. Note that actor 0 is the player who is performing the key action. (b) Scale 2: At the person scale, the CLS and ball tokens mostly attend to actor 0. Other person tokens mostly attend to the CLS, ball, actor 0, 1 and 2. (c) Scale 3: At the interaction scale, the CLS and ball tokens mostly attend to the interactions leading by actor 0 to actor 2 and 3, i.e., interactions (actor 0, actor 2) and (actor 0, actor 3). (d) Scale 4: At the group scale, the CLS and ball tokens mostly attend to each other, and two teams mostly attend to each other. Please zoom in on the image to appreciate the details. COMPOSER is able to attend to relevant entities across different scales and can produce interpretable results.

Overall, COMPOSER produces interesting and interpretable qualitative results. COMPOSER is able to attend to relevant entities across different scales. We hypothesize that the attention weights of Multiscale Transformer can be utilized to further obtain performance improvements, which we will explore in our future work.

C. Failure Cases.

False failure cases – Mislabeled test clips. “False failure case” means the annotated label is wrong but the prediction from COMPOSER is correct. Through examining failure cases of COMPOSER, we find numerous test clips in the Volleyball dataset having wrong annotated labels. Sometimes the annotation of the team is wrong (e.g., a test clip with the “left set” group activity is annotate as “right set” as shown in Fig. 11), and sometimes the annotated main group activity is wrong (e.g., a test clip with the “right set” group activity is annotate as “right spike” as shown in Fig. 12). Instead, COMPOSER predicts the correct group activities with interpretable attention weights.

True failure cases. “True failure case” means the prediction from COMPOSER is wrong and the annotated label is
correct. We visualize two true failure cases of COMPOSER in Fig. 13 and Fig. 14. As shown in the frames on the left side of Fig. 13, the ground truth label of the test clip is right pass. Prediction from COMPOSER is right set. Note that actor 0 is the one performing the key action. For this failure case, COMPOSER tends to have a discrepancy in what to attend to between the scales. At scale 2 and 3 (person scale and interaction scale, respectively), COMPOSER successfully identifies tokens associated with the key person actor 0 as the most important tokens. However, at scale 1, COMPOSER fails to focus on keypoints of actor 0 and eventually makes a wrong group activity prediction. The failure of COMPOSER can be attributed to the fact that the arms of actor 0 is occluded, which could be relatively rare. In addition, in Fig. 14, COMPOSER fails to identify which team is performing the group activity “spike”. However, we notice a major change in the camera location of this test clip, which causes this test clip possibly to be difficult for hu-
Figure 10. **Qualitative results of COMPOSER on the Volleyball dataset** – showcasing attention matrices (from the last Multiscale Transformer block) of a clip instance in the “right winpoint” class. Darker color denotes larger attention weights. 

(a) **Scale 1: Keypoint**

(b) **Scale 2: Person**

(c) **Scale 3: Interaction**

(d) **Scale 4: Group**

---

A woman gives a man a high five. A man to identify that the right team is performing the group activity “spike”. This can explain why COMPOSER predicts “left spike” instead of “right spike”. **COMPOSER** learns to recognize the “winpoint” group activity based on the characteristics of the “winpoint” class.
D. Discussion of Limitations of COMPOSER.

As the Confusion Matrix (Sec. 4.5 in the main paper) and the failure cases suggest, sometimes it can be challenging for COMPOSER to discriminate the set, pass and spike group activities. We observe errors from our pose estimator, especially when the scenes are cluttered or have rapid motion. We hypothesize that improving the estimated poses or alternatively proposing methods that can deal with noisy input poses may help to address this limitation. Occasionally, COMPOSER struggles to distinguish which team (left or right) performs the group activity. We hypothesize it is because of close positions and actions of multiple players near the net. Hence, further adding object features such as the location of the net to COMPOSER may help the model to better distinguish which team (left or right) performs the group activity. Moreover, currently COMPOSER only performs relational modeling in the spatial domain. Enhancing COMPOSER to perform relational reasoning in both spatial and temporal domains may improve the performance and robustness of COMPOSER by disregarding the noises in certain frames.
E. Method and Implementation Details

E.1 Transformer

We briefly describe the Transformer encoder [92] used in Multiscale Transformer in this subsection. The basic components of the Transformer encoder include 1) Multi-head Self-Attention (MSA), 2) Multi-Layer Perceptron (MLP), and 3) Skip Connection [34], Dropout [86] and Layer Normalization [5] (Add & Dropout & LN).

**MSA.** Central to the Transformer encoder is the self-attention function. In the self-attention function, the input $X \in \mathbb{R}^{n \times d}$ is first linearly transformed to three parts, i.e., query $Q \in \mathbb{R}^{n \times d_{k}}$, key $K \in \mathbb{R}^{n \times d_{k}}$ and value $V \in \mathbb{R}^{n \times d_{v}}$, where $n$ denotes the number of tokens in the input sequence, and $d$, $d_{k}$ and $d_{v}$ are the representation dimensions of the input, query (or key), and value, respectively. The Scaled Dot-Product Attention is applied on $Q$, $K$, and $V$:

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_{k}}} \right) V$$ (7)

Then, a linear layer is used to produce the output. We use $h$ paralleled heads of the Scaled Dot-Product Attention to increase the representation power. Specifically, MSA splits the query, key and value for $h$ times and performs the self-attention function in parallel, and then the output values of each head are concatenated and linearly projected to form the final output [32].
Figure 13. True failure case of COMPOSER on the Volleyball dataset. “True failure case” means the prediction from COMPOSER is wrong and the annotated label is correct. The ground truth label of this test clip is right pass. Prediction from COMPOSER is right set. Actor 0 is the one performing the key action. We also showcase the attention matrices on the right; darker color denotes larger attention weights. At scale 2 and 3 (person scale and interaction scale, respectively), COMPOSER successfully identifies tokens associated with the key person, actor 0, as the most important tokens. However, at scale 1, COMPOSER fails to focus on keypoints of actor 0 and eventually makes a wrong group activity prediction. Please zoom in on the image to appreciate the details.

**MLP.** The MLP is for feature transformation and non-linearity:

$$MLP(X) = \sigma (X W_1 + b_1) W_2 + b_2$$  \hspace{1cm} (8)

where $W_1 \in \mathbb{R}^{d \times d_m}$ and $W_2 \in \mathbb{R}^{d_m \times d}$ are weights of the two fully-connected layers, $b_1 \in \mathbb{R}^{d_m}$ and $b_2 \in \mathbb{R}^{d}$ are the bias terms, and $\sigma(\ldots)$ is the non-linear activation function such as ReLU [27] or GELU [35].

**Add & Dropout & LN.** The output from MSA (or MLP) is added with the input of MSA (or MLP) to enforce the skip connection. Then, a dropout layer is used, followed by the Layer Normalization (LN) that enables stable training and faster convergence. Layer normalization is applied over each sample $x \in \mathbb{R}^{d}$ as follows:

$$LN(x) = \frac{x - \mu}{\delta} \circ \gamma + \beta$$  \hspace{1cm} (9)

where $\mu \in \mathbb{R}$ and $\delta \in \mathbb{R}$ are the mean and standard deviation of the features, respectively, $\circ$ denotes the element-wise multiplication, $\gamma \in \mathbb{R}^{d}$ and $\beta \in \mathbb{R}^{d}$ are learnable affine transform parameters for scaling and shifting, respectively.

Given the input $X$, the computations in order in the Transformer encoder are: MSA, Add & Dropout & LN, MLP, and Add & Dropout & LN.
E.2 Implementation Details of COMPOSER

In order to make a fair comparison with related works [26, 96, 103, 104], we use \( T = 10 \) frames as the input to our model for both training and testing, which corresponds to 5 frames before the annotated frame and 4 frames after. The frame resolution is 720 x 1280. The maximal number of actors in a video \( p' = 12 \). Actors are grouped into the two sub-groups by heuristics, i.e., according to the horizontal positions of the players, the left most 6 players form a sub-group and the rest players form the other sub-group. We find that using clustering algorithms such as K-means given the coordinates of the players as features can generate similar results as the heuristics on the Volleyball dataset. Hence, we choose to use the heuristics for simplicity. The ball trajectories from [71] are utilized to encode the object (ball) token \( c' = 1 \). We use HRNet [94] as the pose backbone to obtain estimated human skeletons following [26] \( j' = 17 \). The number of GCN layers for keypoint type encoding is set to 3. In data augmentations, the range of random perturbation is set to 1 pixel location. The number of Multiscale Transformer blocks \( M \) is set as 2, hidden dimension \( d = 256 \) (\( d_k, d_v \) and \( d \) are equal), and the number of attention heads of the Transformer encoder at each scale is set to 2, 8, 2 and 2, respectively. The dropout rate of the Transformer encoder at each scale is set to 0.5, 0.2, 0.2 and 0, respectively. We find that a smaller dropout rate in the coarser scales tends to yield a better performance. The dimension of the MLP layer in all Transformer encoders is set to 1024 (i.e., \( d_n = 1024 \)), and the non-linear activation

Figure 14. True failure case of COMPOSER on the Volleyball dataset. “True failure case” means the prediction from COMPOSER is wrong and the annotated label is correct. The ground truth label of this test clip is right spike. Prediction from COMPOSER is left spike. Note that the camera view of this clip is largely different from other clips in the Volleyball dataset, and COMPOSER fails to distinguish which team performs the spike activity. We also showcase the attention matrices on the right; darker color denotes larger attention weights. Please zoom in on the image to appreciate the details.
function is ReLU. We use the Adam optimizer [48]. We train the model for 45 epochs with an initial learning rate 0.0005 and decrease the learning rate to 0.0001 at epoch 40. The weight decay is 0.001, λ is 3, and batch size is 256. Following [8], the temperature parameter τ = 0.1, ε = 0.05, and the number of iterations of the Sinkhorn-Knopp algorithm is set to 3. The number of prototypes K is 1000 for all experiments described in the main paper.

E.3 Implementation Details of Ablation Studies

In this subsection, we describe the methodology of the ablations that we present in Sec. 4.3 of the main paper in details. For all of the ablations, we use the same set of hyper-parameters as our full model.

No Clustering. The only difference from this ablation to our full model COMPOSER is the loss function used in training. Instead of using Equation (6) (in the main paper) as the loss function, this ablation uses the following loss function:

\[ L_{total} = \sum_{m=1}^{M-1} L_{group} + \lambda (L_{group} + L_{person}) \]  

(10)

Label Consistency [78] for Scale Agreement. Similar to the previous ablation, the only difference from this ablation to our full model COMPOSER is the loss function used in training, which is formulated as follows:

\[ L_{total} = \sum_{m=1}^{M-1} L_{group} + \lambda (L_{group} + L_{person} + L_{consistency}) \]  

(11)

where \( L_{consistency} \) represents the loss term that minimizes the \( L2 \) distance between 2 scales of a clip in the logit space. Specifically, for every two pairs of scales, we compute the \( L2 \) loss given the two sets of GAR logits of the two scales. The \( L_{consistency} \) term is the mean of such \( L2 \) losses over all pairs of scales.

1-Scale: Keypoint. For this ablation, there is only one Transformer encoder in the Multiscale Transformer block, and the tokens to the Transformer encoder are:

\[ \{[CLS], e_1, \cdots, e_{e'}, k_1, \cdots, k_{e'}' \} \]  

(12)

Since only the keypoint tokens are refined by the Multiscale Transformer block in this ablation and there is only 1 scale, this ablation uses the following loss function:

\[ L_{total} = \sum_{m=1}^{M-1} L_{group} + \lambda L_{group} \]  

(13)

2-Scale: Keypoint + Person. For this ablation, there are two hierarchical scales in the Multiscale Transformer block, and the tokens to the Transformer encoders are:

\[ \{[CLS], e_1, \cdots, e_{e'}, k_1, \cdots, k_{e'}' \} \]  

\[ \{[CLS], e_1, \cdots, e_{e'}, p_1, \cdots, p_{p'} \} \]  

(14)

This ablation uses the same loss function as our full model except that the number of pairs of scales for swapped prediction is only 1, i.e., pair scale1-scale2.

3-Scale: Keypoint + Person + Interaction. For this ablation, there are three hierarchical scales in the Multiscale Transformer block, and the tokens to the Transformer encoders are:

\[ \text{Scale 1:} \quad \{[CLS], e_1, \cdots, e_{e'}, k_1, \cdots, k_{e'}' \} \]  

\[ \text{Scale 2:} \quad \{[CLS], e_1, \cdots, e_{e'}, p_1, \cdots, p_{p'} \} \]  

\[ \text{Scale 3:} \quad \{[CLS], e_1, \cdots, e_{e'}, i_1, \cdots, i_{p' \times (p'-1)} \} \]  

(15)

This ablation uses the same loss function as our full model except that the number of pairs of scales for swapped prediction is 3, i.e., pair scale1-scale2, pair scale1-scale3, and pair scale2-scale3.

No Auxiliary Prediction. The only difference from this ablation to our full model COMPOSER is the loss function used in training. This ablation uses the following loss function:

\[ L_{total} = L_{group} + L_{person} + L_{cluster} \]  

(16)

No Multiscale Transformer. In this ablation, the group activity classifier simply takes features of the initial ball token and person tokens as inputs, and the person tokens are aggregated from the initial representations of keypoint tokens through concatenation and FFN. In addition, features of the person tokens are the inputs to the person action classifier. No transformers are used in this ablation. Due to the lack of relational reasoning performed at multiple scales, this ablation uses the following loss function:

\[ L_{total} = L_{group} + L_{person} \]  

(17)