SPARTAN: Self-supervised Spatiotemporal Transformers Approach to Group Activity Recognition

Naga VS Raviteja Chappa[1], Pha Nguyen[1], Alexander H Nelson[1], Han-Seok Seo[1], Xin Li[5]
Page Daniel Dobbs[1],Khoa Luu[1]
[1]University of Arkansas [5]West Virginia University

{nchappa, panguyen, ahnelson, hanseok, pdobbs, khoaluu}@uark.edu, Xin.Li@mail.wvu.edu
https://uark-cviu.github.io

Abstract

In this paper, we propose a new, simple, and effective Self-supervised Spatio-temporal Transformers (SPARTAN) approach to Group Activity Recognition (GAR) using unlabeled video data. Given a video, we create local and global Spatio-temporal views with varying spatial patch sizes and frame rates. The proposed self-supervised objective aims to match the features of these contrasting views representing the same video to be consistent with the variations in spatiotemporal domains. To the best of our knowledge, the proposed mechanism is one of the first works to alleviate the weakly supervised setting of GAR using the encoders in video transformers. Furthermore, using the advantage of transformer models, our proposed approach supports long-term relationship modeling along spatio-temporal dimensions. The proposed SPARTAN approach performs well on two group activity recognition benchmarks, including NBA and Volleyball datasets, by surpassing the state-of-the-art results by a significant margin in terms of MCA and MPCA metrics[1].

1. Introduction

Group Activity Recognition (GAR) aims to classify the collective actions of individuals in a video clip. This field has gained significant attention due to its diverse applications such as sports video analysis, video monitoring, and interpretation of social situations. Far apart from conventional action recognition methods that focus on understanding individual actions [11, 51, 59, 62], GAR requires a thorough and exact knowledge of interactions between several actors, which poses fundamental challenges such as actor localisation and modelling their spatiotemporal relationships. Most existing methods for GAR [16, 20, 26, 28, 37, 43, 45, 49, 64, 67, 69] require ground-truth bounding boxes of individual actors for training and testing, as well as their
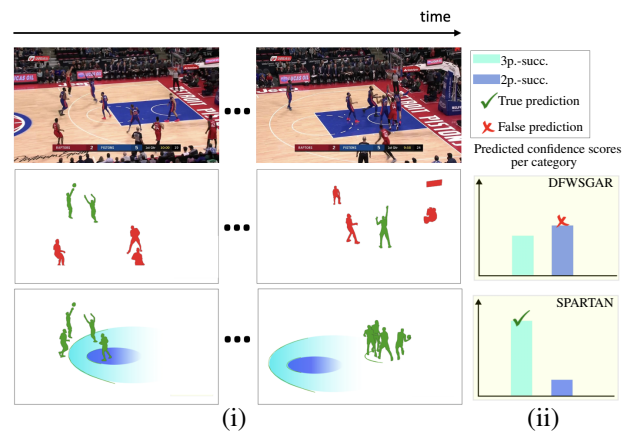

Figure 1. Visualization of attention captured by the model. (i) The attention in this example focuses on how the relationship is established between the actors. Original sequence from NBA dataset [68] (top), Attention captured by DFWSGAR [30] (middle), and SPARTAN model (bottom). Red-colored actors are the irrelevant information to determine the group activity, whereas green-colored actors, including their positions, are the most relevant. (ii) illustrates that DFWSGAR predicts the category wrong due to the effects shown in (i) whereas SPARTAN is more confident in the prediction, which is further justified by the t-SNE plot as shown in Fig. 5.

action class labels for training. The bounding box labels, in particular, are used to extract features of individual actors, such as *RoIPool* [52] and *RoIAlign* [25], and precisely discover their spatio-temporal relations; such actor features are aggregated while considering the relationships between actors to form a group-level video representation, which is then fed to a group activity classifier. Despite the fact that these approaches performed admirably on the difficult task, their reliance on bounding boxes at inference and substantial data labelling annotations makes them unworkable and severely limits their application. To overcome this problem, one approach is to simultaneously train person detection and group activity recognition using bounding box labels [7, 72]. This method estimates the bounding boxes of actors in inference. However, this method calls for individ-

---

[1]The implementation of SPARTAN is available at https://github.com/uark-cviu/SPARTAN

ual actor ground-truth bounding boxes for training videos. Yan *et al*. [68] presented the Weakly Supervised GAR (WS-GAR) learning approach, which does not need actor-level labels in both training and inference, to further lower the annotation cost. They generate actor box suggestions using a detector that has been pre-trained on an external dataset in order to solve the absence of bounding box labels. They then learn to eliminate irrelevant possibilities. Recently, Kim *et al*. [30] introduced a detector-free method for WS-GAR task which captures the actor information using partial contexts of the token embeddings. However, the previous methods [30, 68] have various drawbacks as follows. First, a detector [68] often leads to missing detection of people in case of occlusion, which minimizes overall accuracy. Second, partial contexts [30] can only learn if and only if there is movement in consecutive frames. This can be inferred from the illustration in Fig. 1. Third, the temporal information among the tokens must be consistent, and [30] does not consider different tokens.

In this paper, we introduce a new simple but effective Self-Supervised **Spa**tio-tempo**r**al **Tran**sformers (SPARTAN) approach to the task of Group Action Recognition that is independent of ground-truth bounding boxes, labels during pre-training, and object detector. Our mechanism only exploits the motion as a supervisory signal from the RGB data modality. As seen in Fig. 1 (i), our model captures not only the key actors but also their positions, which shows that our method is more effective in group activity classification than DFWSGAR [30]. Our approach is designed to benefit from varying spatial and temporal details within the same deep network. We use a video transformer [8] based approach to handle varying temporal resolutions within the same architecture. Furthermore, the self-attention mechanism in video transformers can capture local and global long-range dependencies in both space and time, offering much larger receptive fields compared to standard convolutional kernels [42]. The contributions of this work can be summarized as follows.

- Instead of considering only motion features across consecutive frames [30], we introduce the first training approach to GAR by exploiting spatial-temporal correspondences. The proposed method varies the space-time features of the inputs to learn long-range dependencies in spatial and temporal domains.

- A new self-supervised learning strategy is performed by jointly learning the inter-frame, i.e., frame-level *temporal*, and intra-frame, i.e., patch-level *spatial*, correspondences further forming into *Inter Teacher-Inter Student loss* and *Inter Teacher-Intra Student loss*. In particular, the spatiotemporal features global, from the entire sequence, and local from the sampled sequence are matched by the learning objectives of the frame

level and the patch level in the latent space.

- With extensive experiments on NBA [68], and Volleyball [28] datasets, the proposed method shows the State-of-the-Art (SOTA) performance results using only RGB inputs.

2. Related Work

2.1. Group Activity Recognition (GAR)

Due to the wide range of applications, GAR has recently gained more attention. The initial approaches in the field utilized probabilistic graphical methods [1–3, 32, 33, 63] and AND-OR grammar methods [4, 54] to process the extracted features. As deep learning evolved over the years, methods involving Convolutional Neural Networks (CNN) [7, 28], Recurrent Neural Networks (RNN) [7, 13, 27, 28, 38, 46, 55, 60, 66] achieved outstanding performance thanks to their learning power of high-level information and temporal context.

Recent methods for identifying group actions [16, 20, 26, 37, 45, 64, 68, 70] typically utilize attention-based models and require explicit character representations to model spatial-temporal relations in group activities. Graph convolution networks, as described in [64, 70], are used to learn spatial and temporal information of actors by constructing relational graphs, while Rui *et al*. [68] suggest building spatial and temporal relation graphs to infer actor links. Kirill *et al*. [20] use a transformer encoder-based technique with different backbone networks to extract features for learning actor interactions from multimodal inputs. Li *et al*. [37] use a clustered attention approach to capture contextual spatial-temporal information. Mingfei *et al*. [23] proposed MAC-Loss which is a combination of spatial and temporal transformers in two complimentary orders to enhance the learning effectiveness of actor interactions and preserve actor consistency at the frame and video levels.

Weakly supervised group activity recognition (GAR). Several techniques have been developed to tackle weakly supervised group activity recognition (GAR) with less supervision, such as using bounding boxes to train built-in detectors or activity maps. One approach is WSGAR, which does not rely on bounding box annotations during training or inference and incorporates an off-the-shelf item detector into the model. Another technique, proposed by Zhang et al. [73], uses activity-specific characteristics to improve WSGAR, but is not specifically designed for GAR. Kim et al. [30] proposed a detector-free method that uses transformer encoders to extract motion features. Our proposed method is a self-supervised training approach dedicated to WSGAR, which does not require actor-level annotations, object detectors, or labels.

Transformers in Vision. Vaswani *et al*. [58] introduced the transformer architecture for sequence-to-sequence machine
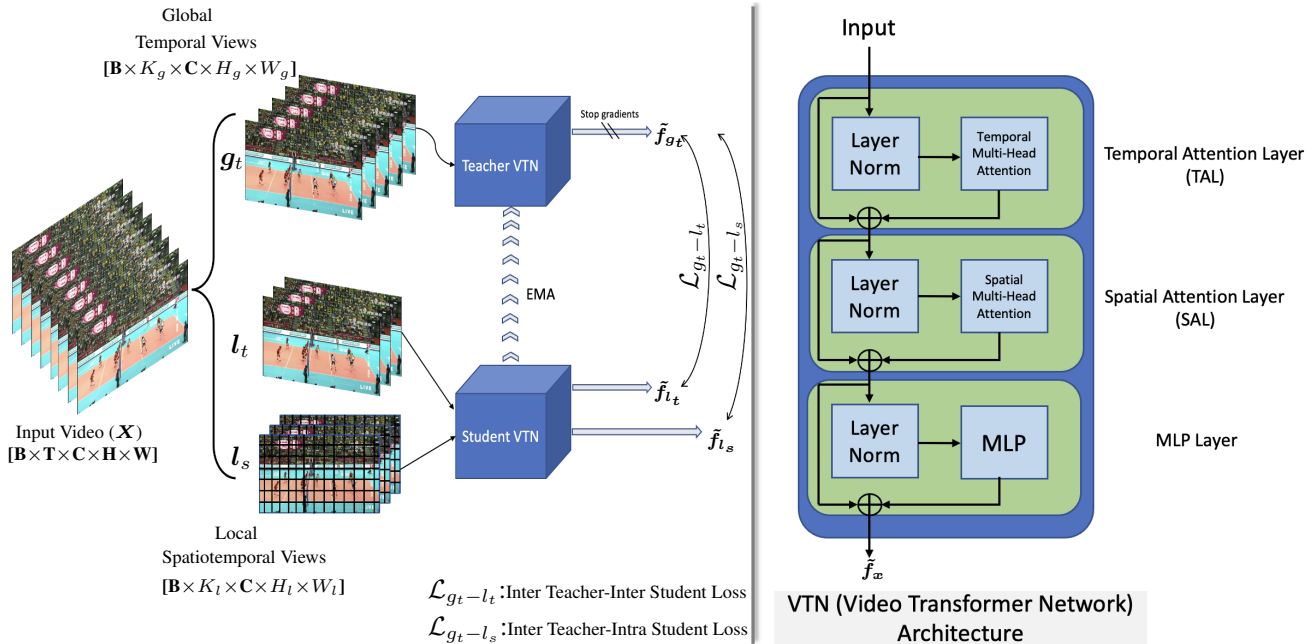
Figure 2. **The proposed SPARTAN Framework** samples gave input video into global and local views. The sampling strategy for video clips results in different frame rates and spatial characteristics between global views and local views, which are subject to spatial augmentations and have limited fields of view. The teacher model processes global views (g_t) to generate a target, while the student model processes local views (l_t & l_s) where $Kl \leq K_g$. The network weights are updated by matching the online student local views to the target teacher global views, which involves *cross-view correspondences* and *motion correspondences*. Our approach utilizes a standard ViT backbone with separate space-time attention [8] and an MLP to predict target features from online features.

translation. This architecture has since been widely adopted to many various natural processing tasks. Dosovitskiy *et al*. [15] introduced a transformer architecture that is not based on convolution for image recognition tasks. For different downstream computer vision tasks, these works [36, 40, 61, 71] used transformer architecture as a general backbone to make exceptional performance progress. In the video domain, many works [5, 9, 17, 22, 34, 35, 44, 48, 50] exploited spatial and temporal self-attention to learn video representation efficiently. Patrick *et al*. [44] introduce a self-attention block that focuses on the trajectory, which tracks the patches of space and time in a video transformer.

3. SPARTAN

The proposed method aims to recognize a group activity in a given video without using person-bounding boxes or a detector. The general architecture of our self-supervised training within the teacher-student framework for group activity recognition is illustrated in Fig. 2. Unlike the other contrastive learning methods, we process two clips from the same video by changing their spatial-temporal characteristics, which do not rely on the memory banks. The proposed loss formulation matches the features of the two dissimilar clips to impose consistency in motion and spatial changes in the same video. The proposed SPARTAN framework will

be discussed further in the following sections.

3.1. Self-Supervised Training

Given the high temporal dimensionality of videos, motion and spatial characteristics of the group activity will be learned, such as 3p.-succ. (from NBA dataset [68]) or l-spike (from Volleyball dataset [28]) during the video. Thus, several video clips with different motion characteristics can be sampled from a single video. A key novelty of the proposed approach involves predicting these different video clips with varying temporal characteristics from each other in the feature space. It leads to learning contextual information that defines the underlying distribution of videos and makes the network invariant to motion, scale, and viewpoint variations. Thus, self-supervision for video representation learning is formulated as a motion prediction problem that has three key components: **a)** We generate multiple temporal views consisting of different numbers of clips with varying motion characteristics from the same video as in Sec. 3.1.1, **b)** In addition to motion, we vary spatial characteristics of these views as well by generating local, i.e., smaller spatial field, and global, i.e., higher spatial field, of the sampled clips as in Sec. 3.1.2, and **c)** We introduce a loss function in Sec. 3.2 that matches the varying views across spatial and temporal dimensions in the latent space.

3.1.1 Motion Prediction as Self-supervision Learning

The frame rate is a crucial aspect of a video as it can significantly alter the motion context of the content. For instance, the frame rate can affect the perception of actions, such as walking slowly versus walking quickly, and can capture subtle nuances, such as the slight body movements in walking. Traditionally, video clips are sampled at a fixed frame rate [47, 65]. However, when comparing views with different frame rates, i.e., varying numbers of clips, predicting one view from another in feature space requires explicitly modeling object motion across clips. Furthermore, predicting subtle movements captured at high frame rates compels the model to learn contextual information about motion from a low frame rate input.

Temporal Views: We refer to a collection of clips sampled at a specific video frame rate as a temporal view. We generate different views by sampling at different frame rates, producing temporal views with varying resolutions. The number of temporal tokens (T) input to ViT varies in different views. Our proposed method enforces the correspondences between such views, which allows for capturing different motion characteristics of the same action. We randomly sampled these views to create motion differences among them. Our ViT models process these views, and we predict one view from the other in the latent space. In addition to varying temporal resolution, we also vary the resolution of clips across the spatial dimension within these views. It means that the spatial size of a clip can be lower than the maximum spatial size (224), which can also decrease the number of spatial tokens. Similar sampling strategies have been used [18, 29] but under multi-network settings, while our approach handles such variability in temporal resolutions with a single ViT model by using vanilla positional encoding [58].

3.1.2 Cross-View Correspondences

Our training strategy aims to learn the relationships between a given video's temporal and spatial dimensions. To this end, we propose novel cross-view correspondences by altering the field of view during sampling. We generated global and local temporal views from a given video to achieve this.

Global Temporal Views (g_t): We randomly sample K_g (is equal to T) frames from a video clip with spatial size fixed to W_{global} and H_{global}. These views are fed into the teacher network which yields an output denoted by $\tilde{f}_{g_t}$.

Local Spatiotemporal Views (l_t and l_s): Local views cover a limited portion of the video along both spatial and temporal dimensions. We generate local temporal views by randomly sampling several frames K_l ($\leq K_g$) with a spatial size fixed to W_{local} and H_{local}. These views are fed into the student network which yields two outputs denoted by $\tilde{f}_{l_t}$ and $\tilde{f}_{l_s}$ respectively.

Augmentations: We apply different data augmentation techniques to the spatial dimension, that is, to the clips sampled for each view. Specifically, we apply color jittering and gray scaling with probability 0.8 and 0.2, respectively, to all temporal views. We apply Gaussian blur and solarization with probability 0.1 and 0.2, respectively, to global temporal views.

Our approach is based on the intuition that learning to predict a global temporal view of a video from a local temporal view in the latent space can help the model capture high-level contextual information. Specifically, our method encourages the model to model both spatial and temporal context, where the spatial context refers to the possibilities surrounding a given spatial crop and the temporal context refers to possible previous or future clips from a given temporal crop. It is important to note that spatial correspondences also involve a temporal component, as our approach attempts to predict a global view at timestamp $t = j$ from a local view at timestamp $t = i$. To enforce these cross-view correspondences, we use a similarity objective that predicts different views from each other.

3.2. The Proposed Objective Function

Our model is trained with an objective function that predicts different views from each other. These views represent different spatial-temporal variations that belong to the same video.

Given a video $\boldsymbol{X} = \{\boldsymbol{x}_t\}_{t=1}^T$, where T represents the number of frames, let $\boldsymbol{g}_t$, $\boldsymbol{l}_t$ and $\boldsymbol{l}_s$ represent global temporal views, local temporal and spatial views such that $\boldsymbol{g}_t = \{\boldsymbol{x}_t\}_{t=1}^{K_g}$ and $\boldsymbol{l}_t = \boldsymbol{l}_s = \{\boldsymbol{x}_t\}_{t=1}^{K_l}$, where $\boldsymbol{g}_t$, $\boldsymbol{l}_t$ and $\boldsymbol{l}_s$ are subsets of video $\boldsymbol{X}$ and $K_l \leq K_g$ where K_g and K_l are the number of frames for teacher and student (global and local) inputs. We randomly sample K_g global and K_l local temporal views as in Sec. 3.1.2. These temporal views are passed through the student and teacher models to get the corresponding class tokens or feature $\boldsymbol{f}_g$ and $\boldsymbol{f}_l$. These class tokens are normalized as follows.

$$\tilde{\boldsymbol{f}}^{(i)} = \frac{\exp(\boldsymbol{f}^{(i)})/\tau}{\sum_{i=1}^n \exp(\boldsymbol{f}^{(i)})/\tau}, \qquad (1)$$

where τ is a temperature parameter used to control the sharpness of the exponential function [10] and $\boldsymbol{f}^{(i)}$ is each element in $\tilde{\boldsymbol{f}}^{(i)} \in \mathbb{R}^n$.

Inter Teacher-Inter Student Loss: Our $\boldsymbol{g}_t$ have the same spatial size but differ in temporal content because the number of clips/frames is randomly sampled for each view. One of the $\boldsymbol{g}_t$ always passes through the teacher model that serves as the target label. We map the student's $\boldsymbol{l}_t$ with the teacher's $\boldsymbol{g}_t$ to create a global-to-local temporal loss as in Eqn. (2).

$$\mathcal{L}_{g_t-l_t} = -\tilde{\boldsymbol{f}}_{g_t} * log(\tilde{\boldsymbol{f}}_{l_t}), \qquad (2)$$
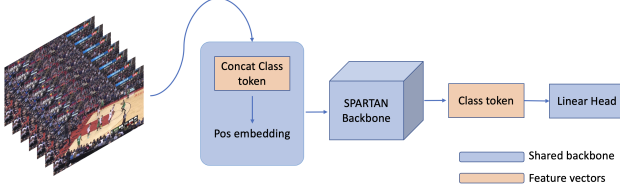
Figure 3. **Inference**. We uniformly sample the video clip and pass it through a shared network and generate feature vectors (class tokens). These vectors are fed to the downstream task classifier.

where $\tilde{f}_{g_t}$ and $\tilde{f}_{l_t}$ are the tokens of the class for g_t and l_t produced by the teacher and student, respectively.

Inter Teacher-Intra Student Loss: Our l_t have a limited field of vision along the spatial and temporal dimensions compared to the g_t. However, the number of local views is four times higher than that of global views. All l_s are passed through the student model and mapped to g_t from the teacher model to create the loss function as in Eq. (3).

$$\mathcal{L}_{g_t - l_s} = \sum_{n=1}^{q} -\tilde{f}_{g_t} * log(\tilde{f}_{l_s}^{(n)}), \tag{3}$$

where $\tilde{f}_{l_s}$ are the tokens of the class for l_s produced by the student and q represents the number of local temporal views set to sixteen in all our experiments. The overall loss to train our model is simply a linear combination of both losses, as in Eqn. (2) and Eqn. (3), given as in Eqn. (4).

$$\mathcal{L} = \mathcal{L}_{g_t - l_t} + \mathcal{L}_{g_t - l_s} \tag{4}$$

3.3. Inference

Fig. 3 illustrates our inference framework. During this stage, fine-tuning of the trained self-supervised model is performed. We use the pre-trained SPARTAN model and fine-tune the model with the available labels, followed by a linear classifier. We use this on downstream tasks to improve performance.

4. Experiments

4.1. Datasets

Volleyball Dataset [28] comprises 3,493 training and 1,337 testing clips, totaling 4,830 labeled clips, from 55 videos. The dataset contains annotations for eight group activity categories and nine individual action labels with corresponding bounding boxes. However, in our WSGAR experiments, we only use the group activity labels and ignore the individual action annotations. For evaluation, we use Multi-class Classification Accuracy (MCA) and Merged MCA metrics, where the latter merges the right set and right pass classes into right pass-set and the left set and left pass classes into left pass-set, as in previous works such as

SAM [68] and DFWSGAR [30]. This is done to ensure a fair comparison with existing methods.

NBA Dataset [68] in our experiment comprises a total of 9,172 labeled clips from 181 NBA videos, with 7,624 clips used for training and 1,548 for testing. Each clip is annotated with one of nine group activities, but there is no information on individual actions or bounding boxes. In evaluating the model, we use the Multi-class Classification Accuracy (MCA) and Mean Per Class Accuracy (MPCA) metrics, with MPCA used to address the issue of class imbalance in the dataset.

4.2. Deep Network Architecture

Our video processing approach uses a vision transformer (ViT) [8] to apply individual attention to both the temporal and spatial dimensions of the input video clips. The ViT consists of 12 encoder blocks and can process video clips of size $(B \times T \times C \times W \times H)$, where B and C represent the batch size and the number of color channels, respectively. The maximum spatial and temporal sizes are $W = H = 224$ and $T = 18$, respectively, meaning that we sample 18 frames from each video and rescale them to 224×224. Our network architecture (see Fig. 2) is designed to handle variable input resolution during training, such as differences in frame rate, number of frames in a video clip, and spatial size. However, each ViT encoder block processes a maximum of 196 spatial and 16 temporal tokens, and each token has an embedding dimension of $\mathbb{R}^m$ [15]. Along with these spatial and temporal input tokens, we also use a single classification token as a characteristic vector within the architecture [14]. This classification token represents the standard features learned by the ViT along the spatial and temporal dimensions of a given video. During training, we use variable spatial and temporal resolutions that are $W \leq 224$, $H \leq 224$, and $T \leq 18$, which result in various spatial and temporal tokens. Finally, we apply a projection head to the class token of the final ViT encoder [10, 21].

Self-Distillation. In our approach (shown in Fig. 2), we adopt a teacher-student setup for self-distillation inspired by [10, 21]. The teacher model has the same architecture as the student model, including the ViT backbone and predictor MLP, but it does not undergo direct training. Instead, during each training step of the student model, we update the teacher weights using an exponential moving average (EMA) of the student weights [10]. This approach enables us to use a single shared network to process multiple input clips.

4.3. Implementation Details

For both the NBA and Volleyball datasets, frames are sampled at a rate of T (K_g) using segment-based sampling [59]. The frames are then resized to $W_g = 224$ $H_g = 224$

Method	MCA	MPCA
Video backbone		
TSM [39]	66.6	60.3
VideoSwin [41]	64.3	60.6
GAR model		
ARG [64]	59.0	56.8
AT [20]	47.1	41.5
SACRF [45]	56.3	52.8
DIN [70]	61.6	56.0
SAM [68]	54.3	51.5
DFWSGAR [30]	75.8	71.2
Ours	**82.1**	**72.8**

Table 1. Comparisons with the State-of-the-Art GAR models and video backbones on the NBA dataset [68].

for the teacher input and $W_l = 96$ $H_l = 96$ for the student input, respectively. For the Volleyball dataset, we use $K_g = 5$ ($K_l \in 3, 5$), while for the NBA dataset, we use $K_g = 18$ ($K_l \in 2, 4, 8, 16, 18$). We randomly initialize weights relevant to temporal attention, while spatial attention weights are initialized using a ViT model trained self-supervised over ImageNet-1K [53]. This initialization setup allows us to achieve faster convergence of space-time ViT similar to the supervised setting [8]. We use an Adam optimizer [31] with a learning rate of 5×10^{-4}, scaled using a cosine schedule with a linear warm-up for five epochs [12, 56]. We also use weight decay scaled from 0.04 to 0.1 during training. For the **downstream task**, we train a linear classifier on our pretrained SPARTAN backbone. During training, the backbone is frozen, and the classifier is trained for 100 epochs with a batch size of 32 on a single NVIDIA-V100 GPU using SGD with an initial learning rate of 1e-3 and a cosine decay schedule. We also set the momentum to 0.9.

4.4. Comparison with state-of-the-art methods

NBA dataset We compare our approach to the state-of-the-art in GAR and WSGAR, which leverage bounding box recommendations produced by SAM [68], as well as to current video backbones in the weakly supervised learning environment, using the NBA dataset. We exclusively utilise RGB frames as input for each approach, including the video backbones, to ensure a fair comparison. Table 1 lists the findings. Please take note that the reproduction of the SAM [68] has greater scores than those listed in the original article. With 6.3%p of MCA and 1.6%p of MPCA, the proposed method outperforms existing GAR and WSGAR methods by a significant margin. Additionally, our approach is contrasted with two current video backbones utilised in traditional action detection, ResNet-18 TSM [39] and VideoSwin-T [41]. These strong backbones perform admirably in WSGAR, but ours is the finest.

Volleyball dataset. For the volleyball dataset, we compare our approach to the most recent GAR and WSGAR approaches in two different supervision levels: fully super-

Method	Backbone	MCA	Merged MCA
Fully supervised			
SSU [7]	Inception-v3	89.9	-
PCTDM [66]	ResNet-18	90.3	94.3
StagNet [46]	VGG-16	89.3	-
ARG [64]	ResNet-18	91.1	<u>95.1</u>
CRM [6]	I3D	92.1	-
HiGCIN [67]	ResNet-18	91.4	-
AT [20]	ResNet-18	90.0	94.0
SACRF [45]	ResNet-18	90.7	92.7
DIN [70]	ResNet-18	<u>93.1</u>	**95.6**
TCE+STBiP [69]	VGG-16	**94.1**	-
GroupFormer [37]	Inception-v3	**94.1**	-
Weakly supervised			
PCTDM [66]	ResNet-18	80.5	90.0
ARG [64]	ResNet-18	87.4	92.9
AT [20]	ResNet-18	84.3	89.6
SACRF [45]	ResNet-18	83.3	86.1
DIN [70]	ResNet-18	86.5	93.1
SAM [68]	ResNet-18	86.3	93.1
DFWSGAR [30]	ResNet-18	90.5	94.4
Ours	ViT-Base	**92.9**	**95.6**

Table 2. Comparison with the state-of-the-art methods on the Volleyball dataset. [28]

vised and weakly supervised. The usage of actor-level labels, such as individual action class labels and ground-truth bounding boxes, in training and inference differs across the two settings.

For a fair comparison, we report the results of previous methods [6, 7, 37, 46, 67, 69] using only the RGB input, and the reproduced results [20, 45, 64, 66, 70] using the ResNet-18 backbone. Note that the first is from the original papers, and the second is the MCA values of [70]. We eliminate the individual action classification head and substitute an object detector trained on an external dataset for the ground-truth bounding boxes in the weakly supervised situation. Table 2 presents the results. Results from earlier techniques in fully supervised and weakly supervised environments are displayed in the first and second sections, respectively. In weakly supervised conditions, our technique significantly outperforms all GAR and WSGAR models, outperforming them by 2.4% of MCA and 1.2% of Merged MCA when compared to the models' utilising ViT-Base backbone. Our technique outperforms current GAR methods, such as [7, 20, 45, 46, 66], by employing more thorough actor-level supervision.

4.5. Ablation Study

We perform a comprehensive analysis of the different components that contribute to the effectiveness of our method. Specifically, we evaluate the impact of five individual elements: **a)** various combinations of local and global

$l_t \to g_t$	$l_s \to g_t$	$l_s \to l_t$	$g_t \to l_t$	NBA	Volleyball
✓	✗	✗	✗	61.03	62.70
✗	✓	✗	✗	62.59	65.40
✓	✓	✗	✗	**81.20**	**90.80**
✓	✓	✓	✗	72.11	77.62
✓	✓	✗	✓	78.17	85.88
✗	✗	✓	✓	64.36	71.87

Table 3. **View Correspondences (VC).** The most optimal combination for predicting view correspondences involves predicting local-to-global (temporal) and local-to-global (spatial) views, outperforming other combinations.

Method	NBA	Volleyball
Ours + TIS [47]	78.45	88.11
Ours + MC	**81.20**	**90.80**

Table 5. **Temporal Sampling Strategy** . We evaluate the effectiveness of our proposed temporal sampling strategy, called "*motion correspondences (MC)*" (Sec. 3.1.1), by comparing it with an alternate approach, the "temporal interval sampler (TIS)" [47], used with CNNs under contrastive settings.

Patch size	NBA	Volleyball
8	78.71	87.10
16	**81.20**	**90.80**
32	72.56	79.21

Table 6. **Spatial Augmentations (SA)**: Applying different patch sizes randomly over the spatial dimensions for different views leads to consistent improvements on both NBA and Volleyball datasets.

Multi-view	NBA	Volleyball
✗	76.17	88.35
✓	**81.20**	**90.80**

Table 7. **Inference**: Providing multiple views of different spatiotemporal resolutions to a shared network (multiview) leads to noticeable performance improvements compared to using a single view for both the NBA and Volleyball datasets.

Spatial	Temporal	NBA	Volleyball
✓	✗	69.38	78.59
✗	✓	72.90	81.45
✓	✓	**81.20**	**90.80**

Table 4. **Spatial vs Temporal variations.** The best results are achieved by utilizing cross-view correspondences with varying fields of view along both spatial and temporal dimensions. It is observed that temporal variations between views have a greater impact on performance compared to applying only spatial variation.

which is absent for individual cases. We also observe a consistent performance drop for $l_s \to l_t$ correspondences (no overlap views), consistent with previous findings on the effectiveness of temporally closer positive views for contrastive self-supervised losses [19, 47].

Spatial vs. Temporal Field of View: We determine the optimal combination of spatio-temporal views in Table 3 by varying the field of view (crops) along both spatial and temporal dimensions (as described in Sec.3.1.2). To evaluate the effects of variations along these dimensions, we conduct experiments as presented in Table4. Specifically, we compare the performance of our approach with no variation along the spatial dimension (where all frames have a fixed spatial resolution of 224×224 with no spatial cropping) and with no variation along the temporal dimension (where all frames in our views are sampled from a fixed time-axis region of a video). Our findings show that temporal variations have a significant impact on NBA, while variations in the field of view along both spatial and temporal dimensions lead to the best performance (as shown in Table 4).

Temporal Sampling Strategy: Our investigation examines the possibility of replacing the temporal sampling strategy for motion correspondences (MC) proposed in our study with alternate sampling methods. To evaluate the effectiveness of MC, we replace it with an alternative approach within SPARTAN. Specifically, we test the temporal interval sampling (TIS) strategy introduced in [47], which has achieved state-of-the-art performance in self-supervised contrastive video settings with CNN backbones. Our experiments incorporating TIS in SPARTAN (Table 5) demonstrate that our proposed MC sampling strategy offers superior performance compared to TIS.

Spatial Augmentations: We then investigate the impact of standard spatial augmentations (SA) on video data by experimenting with different patch sizes. Previous studies have shown that varying patch sizes can enhance the performance of CNN-based video self-supervision approaches. In our study, we evaluate the effect of patch size on our approach and present the results in Table 6, indicating that a patch size of 16 yields the best improvements. Based on these findings, we incorporate a patch size of 16 in our SPAR-

view correspondences; **b)** different field of view variations along the temporal and spatial dimensions; **c)** the choice of temporal sampling strategy; **d)** the use of spatial augmentations; and **e)** the inference approach.

View Correspondences: We propose cross-view correspondences (VC) to learn correspondences between local and global views. To investigate the effect of predicting each type of view from the other, we conduct experiments presented in Table 3. Our results show that jointly predicting $l_t \to g_t$ and $l_s \to g_t$ view correspondences leads to optimal performance. However, predicting $g_t \to l_t$ or $l_s \to l_t$ views results in reduced performance, possibly because joint prediction emphasizes learning rich context,
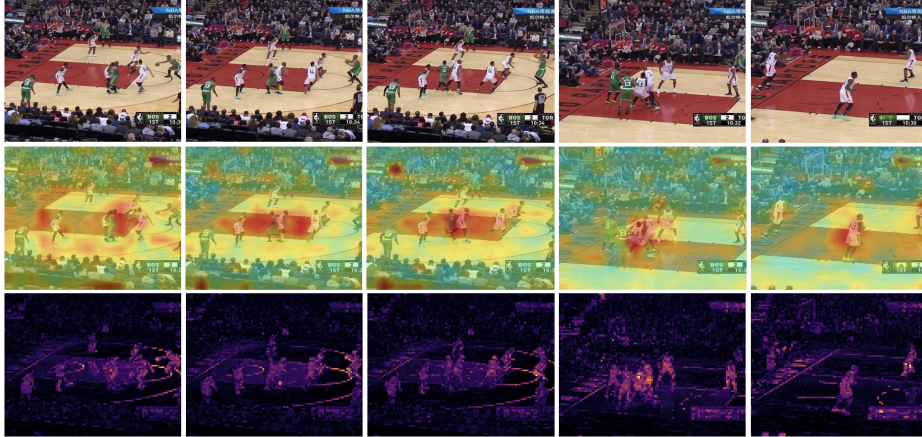
Figure 4. Visualization of the Transformer attention maps for NBA dataset. (top) Original sequence from NBA dataset [68], (middle) Attention maps from DFWSGAR [30] and (bottom) Attention maps from our SPARTAN model.
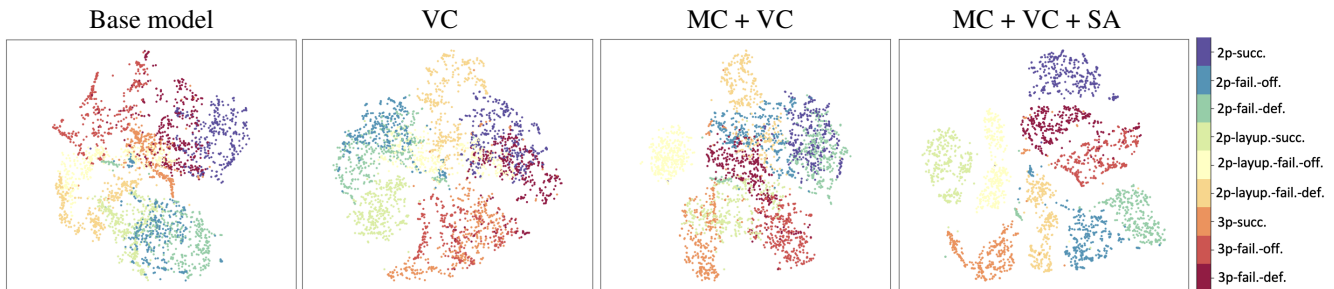

Figure 5. t-SNE [57] visualization of feature embedding learned by different variants of our SPARTAN model in the NBA dataset.

TAN training process.

Inference: To assess the impact of our proposed inference method (Sec. 3.3), we analyze the results presented in Table 7. Our findings demonstrate that our approach yields greater improvements on the NBA [68] and Volleyball [28] datasets, which contain classes that can be more easily distinguished using motion information [24].

4.6. Qualitative Results

We show the attention visualisations derived from the final Transformer encoder layer on the NBA dataset in Fig. 4. The results indicate that the model learnt to pay attention to essential concepts, such as the position of the players, and to follow the activity in a specific video clip. The t-SNE [57] visualisation results of our model and its modifications are shown in Fig. 5. Each model's final group representation on NBA is shown in two-dimensional space. The recommended modules help to clearly separate each class.

5. Conclusion

Our work introduces SPARTAN, a self-supervised video transformer-based model. The approach involves generating multiple spatio-temporally varying views from a single video at different scales and frame rates. Two sets of cor-

respondence learning tasks are then defined to capture the motion properties and cross-view relationships between the sampled clips. The self-supervised objective involves reconstructing one view from the other in the latent space of teacher and student networks. Moreover, our SPARTAN can model long-range spatio-temporal dependencies and perform dynamic inference within a single architecture. We evaluate SPARTAN on two group activity recognition benchmarks and find that it outperforms the current state-of-the-art models.

Limitations: Our paper investigates the application of SPARTAN in the context of the RGB input modality. Currently, we do not utilize the additional supervision provided by alternate modalities in large-scale multimodal video datasets. However, in future work, we plan to explore ways in which we can modify SPARTAN to take advantage of multimodal data sources.

References

[1] Mohamed Rabie Amer, Peng Lei, and Sinisa Todorovic. Hirf: Hierarchical random field for collective activity recognition in videos. In *European Conference on Computer Vision*, pages 572–585. Springer, 2014. 2

[2] Mohamed R Amer and Sinisa Todorovic. Sum product networks for activity recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):800–813, 2015. 2

[3] Mohamed R Amer, Sinisa Todorovic, Alan Fern, and Song-Chun Zhu. Monte carlo tree search for scheduling activity recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1353–1360, 2013. 2

[4] Mohamed R Amer, Dan Xie, Mingtian Zhao, Sinisa Todorovic, and Song-Chun Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *European Conference on Computer Vision*, pages 187–200. Springer, 2012. 2

[5] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021. 3

[6] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional relational machine for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7892–7901, 2019. 6

[7] 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. 1, 2, 6

[8] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning,*, July 2021. 2, 3, 5, 6

[9] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning,*, 2021. 3

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE international conference on computer vision*, 2021. 4, 5

[11] 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. 1

[12] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. 2021. 6

[13] 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. 2

[14] 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

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 5

[16] 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 *European Conference on Computer Vision*, pages 177–195. Springer, 2020. 1, 2

[17] 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

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211, 2019. 4

[19] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. 2021. 7

[20] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 839–848, 2020. 1, 2, 6

[21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in neural information processing systems*, 2020. 5

[22] Mingfei Han, Yali Wang, Xiaojun Chang, and Yu Qiao. Mining inter-video proposal relations for video object detection. In *European conference on computer vision*, pages 431–446. Springer, 2020. 3

[23] Mingfei Han, David Junhao Zhang, Yali Wang, Rui Yan, Lina Yao, Xiaojun Chang, and Yu Qiao. Dual-ai: Dual-path actor interaction learning for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2990–2999, 2022. 2

[24] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in neural information processing systems*, 2020. 8

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2961–2969, 2017. 1

[26] Guyue Hu, Bo Cui, Yuan He, and Shan Yu. Progressive relation learning for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 980–989, 2020. 1, 2

[27] Mostafa S Ibrahim and Greg Mori. Hierarchical relational networks for group activity recognition and retrieval. In

European Conference on Computer Vision, pages 721–736, 2018. 2

[28] 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. 1, 2, 3, 5, 6, 8

[29] Kumara Kahatapitiya and Michael S Ryoo. Coarse-fine networks for temporal activity detection in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021. 4

[30] Dongkeun Kim, Jinsung Lee, Minsu Cho, and Suha Kwak. Detector-free weakly supervised group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20083–20093, 2022. 1, 2, 5, 6, 8

[31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Int. Conf. Learn. Represent.*, 2015. 6

[32] Tian Lan, Leonid Sigal, and Greg Mori. Social roles in hierarchical models for human activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1354–1361. IEEE, 2012. 2

[33] Tian Lan, Yang Wang, Weilong Yang, Stephen N Robinovitch, and Greg Mori. Discriminative latent models for recognizing contextual group activities. *IEEE transactions on pattern analysis and machine intelligence*, 34(8):1549–1562, 2011. 2

[34] Yunjia Lei, Son Lam Phung, Abdesselam Bouzerdoum, Hoang Thanh Le, and Khoa Luu. Pedestrian lane detection for assistive navigation of vision-impaired people: Survey and experimental evaluation. *IEEE Access*, 10:101071–101089, 2022. 3

[35] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition, 2022. 3

[36] Mingjie Li, Wenjia Cai, Rui Liu, Yuetian Weng, Xiaoyun Zhao, Cong Wang, Xin Chen, Zhong Liu, Caineng Pan, Mengke Li, et al. Ffa-ir: Towards an explainable and reliable medical report generation benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 3

[37] Shuaicheng Li, Qianggang Cao, Lingbo Liu, Kunlin Yang, Shinan Liu, Jun Hou, and Shuai Yi. Groupformer: Group activity recognition with clustered spatial-temporal transformer. *Proceedings of the IEEE international conference on computer vision*, 2021. 1, 2, 6

[38] Xin Li and Mooi Choo Chuah. Sbgar: Semantics based group activity recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2876–2885, 2017. 2

[39] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 6

[40] 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. In *Proceedings of the IEEE international conference on computer vision*, 2021. 3

[41] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. 2021. 6

[42] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. 2021. 2

[43] Pha Nguyen, Thanh-Dat Truong, Miaoqing Huang, Yi Liang, Ngan Le, and Khoa Luu. Self-supervised domain adaptation in crowd counting. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2786–2790, 2022. 1

[44] Mandela Patrick, Dylan Campbell, Yuki M Asano, Ishan Misra Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, Jo Henriques, et al. Keeping your eye on the ball: Trajectory attention in video transformers. In *NeurIPS*, 2021. 3

[45] 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, 2, 6

[46] Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. stagnet: An attentive semantic rnn for group activity recognition. In *European Conference on Computer Vision*, pages 101–117, 2018. 2, 6

[47] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021. 4, 7

[48] Kha Gia Quach, Huu Le, Pha Nguyen, Chi Nhan Duong, Tien Dai Bui, and Khoa Luu. Depth perspective-aware multiple object tracking. *arXiv preprint arXiv:2207.04551*, 2022. 3

[49] Kha Gia Quach, Ngan Le, Chi Nhan Duong, Ibsa Jalata, Kaushik Roy, and Khoa Luu. Non-volume preserving-based fusion to group-level emotion recognition on crowd videos. *Pattern Recognition*, Volume 128:Article 108646, August, 2022. 1

[50] Kha Gia Quach, Pha Nguyen, Huu Le, Thanh-Dat Truong, Chi Nhan Duong, Minh-Triet Tran, and Khoa Luu. Dyglip: A dynamic graph model with link prediction for accurate multi-camera multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13784–13793, June 2021. 3

[51] Kanchana Ranasinghe, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Michael S. Ryoo. Self-supervised video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2874–2884, June 2022. 1

[52] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 1

[53] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy,

Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 2015. 6

[54] Tianmin Shu, Dan Xie, Brandon Rothrock, Sinisa Todorovic, and Song Chun Zhu. Joint inference of groups, events and human roles in aerial videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4576–4584, 2015. 2

[55] Xiangbo Shu, Jinhui Tang, Guojun Qi, Wei Liu, and Jian Yang. Hierarchical long short-term concurrent memory for human interaction recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2

[56] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. 2021. 6

[57] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 2, 4

[59] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016. 1, 5

[60] Minsi Wang, Bingbing 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. 2

[61] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021. 3

[62] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 1

[63] Zhenhua Wang, Qinfeng Shi, Chunhua Shen, and Anton Van Den Hengel. Bilinear programming for human activity recognition with unknown mrf graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1690–1697, 2013. 2

[64] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9964–9974, 2019. 1, 2, 6

[65] Fanyi Xiao, Joseph Tighe, and Davide Modolo. Modist: Motion distillation for self-supervised video representation learning. 2021. 4

[66] Rui Yan, Jinhui Tang, Xiangbo Shu, Zechao Li, and Qi Tian. Participation-contributed temporal dynamic model for group activity recognition. In *Proceedings of the 26th ACM inter-*

national conference on Multimedia, pages 1292–1300, 2018. 2, 6

[67] 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. 1, 6

[68] 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. 1, 2, 3, 5, 6, 8

[69] Hangjie Yuan and Dong Ni. Learning visual context for group activity recognition. In *AAAI*, volume 35, pages 3261–3269, 2021. 1, 6

[70] Hangjie Yuan, Dong Ni, and Mang Wang. Spatio-temporal dynamic inference network for group activity recognition. In *Proceedings of the IEEE international conference on computer vision*, 2021. 2, 6

[71] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021. 3

[72] Peizhen Zhang, Yongyi Tang, Jian-Fang Hu, and Wei-Shi Zheng. Fast collective activity recognition under weak supervision. *IEEE Transactions on Image Processing*, 29:29–43, 2019. 1

[73] Yanyi Zhang, Xinyu Li, and Ivan Marsic. Multi-label activity recognition using activity-specific features and activity correlations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14625–14635, 2021. 2