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

Naga VS Raviteja Chappa¹, Pha Nguyen¹, Alexander H Nelson¹, Han-Seok Seo¹, Xin Li⁵
Page Daniel Dobbs¹, Khoa Luu¹
¹University of Arkansas ⁵West 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 metrics¹.

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 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-

¹The 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 Spatio-temporal Transformers (SPAR-TAN) 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 (gt) to generate a target, while the student model processes local views (lt & ls) where Kl ≤ Kg. 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.

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 resolution.

Local Spatiotemporal Views: We randomly sample $K_g$ clips from a video clip to achieve this. These views are fed into the teacher network which yields an output denoted by $\tilde{f}_g$.

Local Spatiotemporal Views (local): 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 $K_f$ 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 $\tilde{f}_s$ and $\tilde{f}_t$, 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 $X = \{x_t\}_{t=1}^{T}$, where $T$ represents the number of frames, let $g_t$, $l_t$, and $l_s$ represent global temporal views, local temporal and spatial views such that $g_t = \{x_t\}_{t=1}^{K_g}$ and $l_t = l_s = \{x_t\}_{t=1}^{K_t}$, where $g_t$, $l_t$, and $l_s$ are subsets of video $X$ and $K_g \leq K_t$ where $K_g$ and $K_t$ are the number of frames for teacher and student (global and local) inputs. We randomly sample $K_g$ global and $K_t$ 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 $f_g$ and $f_t$. These class tokens are normalized as follows.

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

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

Inter Teacher-Inter Student Loss: Our $g_t$ 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 $g_t$ always passes through the teacher model that serves as the target label. We map the student’s $l_t$ with the teacher’s $g_t$ to create a global-to-local temporal loss as in Eqn. (2).

$$\mathcal{L}_{g_t-l_t} = -\tilde{f}_g \cdot log(\tilde{f}_t),$$ (2)
where $\tilde{f}_g$ and $\tilde{f}_l$ 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 $l_t$ are passed through the student model and mapped to $g_t$ from the teacher model to create the loss function as in Eq. (3).

$$L_{g_t - l_t} = \sum_{n=1}^{q} - \tilde{f}_g * \log(\tilde{f}_l^{(n)}),$$

where $\tilde{f}_l$ are the tokens of the class for $l_t$ 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).

$$\mathcal{L} = L_{g_t - l_t} + L_{g_t - l_s}$$

### 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 $224 \times 224$. 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 $W_g = 224$, $H_g = 224$.
Table 1. Comparisons with the State-of-the-Art GAR models and video backbones on the NBA dataset [68].

<table>
<thead>
<tr>
<th>Method</th>
<th>MCA</th>
<th>MPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSM [39]</td>
<td>66.6</td>
<td>60.3</td>
</tr>
<tr>
<td>VideosWin [41]</td>
<td>64.3</td>
<td>60.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GAR model</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG [64]</td>
<td>59.0</td>
<td>56.8</td>
</tr>
<tr>
<td>AT [20]</td>
<td>47.1</td>
<td>41.5</td>
</tr>
<tr>
<td>SACRF [45]</td>
<td>56.3</td>
<td>52.8</td>
</tr>
<tr>
<td>DIN [70]</td>
<td>61.6</td>
<td>56.0</td>
</tr>
<tr>
<td>SAM [68]</td>
<td>54.3</td>
<td>51.5</td>
</tr>
<tr>
<td>DFWSGAR [30]</td>
<td>75.8</td>
<td>71.2</td>
</tr>
<tr>
<td>Ours</td>
<td>82.1</td>
<td>72.8</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>MCA</th>
<th>Merged MCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully supervised</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSU [7]</td>
<td>Inception-v3</td>
<td>89.9</td>
<td>-</td>
</tr>
<tr>
<td>PCTDM [66]</td>
<td>ResNet-18</td>
<td>90.3</td>
<td>94.3</td>
</tr>
<tr>
<td>StagNet [46]</td>
<td>VGG-16</td>
<td>89.3</td>
<td>-</td>
</tr>
<tr>
<td>ARG [64]</td>
<td>ResNet-18</td>
<td>91.1</td>
<td>95.1</td>
</tr>
<tr>
<td>CRM [6]</td>
<td>I3D</td>
<td>92.1</td>
<td>-</td>
</tr>
<tr>
<td>HiGCIN [67]</td>
<td>ResNet-18</td>
<td>91.4</td>
<td>-</td>
</tr>
<tr>
<td>AT [20]</td>
<td>ResNet-18</td>
<td>90.0</td>
<td>94.0</td>
</tr>
<tr>
<td>SACRF [45]</td>
<td>ResNet-18</td>
<td>90.7</td>
<td>92.7</td>
</tr>
<tr>
<td>DIN [70]</td>
<td>ResNet-18</td>
<td>93.1</td>
<td><strong>95.6</strong></td>
</tr>
<tr>
<td>TCE+STBiP [69]</td>
<td>VGG-16</td>
<td><strong>94.1</strong></td>
<td>-</td>
</tr>
<tr>
<td>GroupFormer [37]</td>
<td>Inception-v3</td>
<td><strong>94.1</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

Weakly supervised

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>MCA</th>
<th>Merged MCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCTDM [66]</td>
<td>ResNet-18</td>
<td>80.5</td>
<td>90.0</td>
</tr>
<tr>
<td>ARG [64]</td>
<td>ResNet-18</td>
<td>87.4</td>
<td>92.9</td>
</tr>
<tr>
<td>AT [20]</td>
<td>ResNet-18</td>
<td>84.3</td>
<td>89.6</td>
</tr>
<tr>
<td>SACRF [45]</td>
<td>ResNet-18</td>
<td>83.3</td>
<td>86.1</td>
</tr>
<tr>
<td>DIN [70]</td>
<td>ResNet-18</td>
<td>86.5</td>
<td>93.1</td>
</tr>
<tr>
<td>SAM [68]</td>
<td>ResNet-18</td>
<td>86.3</td>
<td>93.1</td>
</tr>
<tr>
<td>DFWSGAR [30]</td>
<td>ResNet-18</td>
<td>90.5</td>
<td>94.4</td>
</tr>
<tr>
<td>Ours</td>
<td>ViT-Base</td>
<td><strong>92.9</strong></td>
<td><strong>95.6</strong></td>
</tr>
</tbody>
</table>

for the teacher input and $W_i = 96$ $H_i = 96$ for the student input, respectively. For the Volleyball dataset, we use $K_g = 5$ ($K_i \in 3, 5$), while for the NBA dataset, we use $K_g = 18$ ($K_i \in 2, 4, 8, 16, 18$). We randomly initialize weights relevant to temporal attention, while spatial 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 $5 \times 10^{-4}$, 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 supervised 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
### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>NBA</th>
<th>Volleyball</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours + TIS [47]</td>
<td>78.45</td>
<td>88.11</td>
</tr>
<tr>
<td>Ours + MC</td>
<td><strong>81.20</strong></td>
<td><strong>90.80</strong></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Patch size</th>
<th>NBA</th>
<th>Volleyball</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>78.71</td>
<td>87.10</td>
</tr>
<tr>
<td>16</td>
<td><strong>81.20</strong></td>
<td><strong>90.80</strong></td>
</tr>
<tr>
<td>32</td>
<td>72.56</td>
<td>79.21</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Multi-view</th>
<th>NBA</th>
<th>Volleyball</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>76.17</td>
<td>88.35</td>
</tr>
<tr>
<td>✓</td>
<td><strong>81.20</strong></td>
<td><strong>90.80</strong></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>NBA</th>
<th>Volleyball</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>69.38</td>
<td>78.59</td>
</tr>
<tr>
<td>✓ ✓ ✓ ✓ ✓</td>
<td>72.90</td>
<td>81.45</td>
</tr>
<tr>
<td>✓ ✓ ✓ ✓ ✓</td>
<td><strong>81.20</strong></td>
<td><strong>90.80</strong></td>
</tr>
</tbody>
</table>

### 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, 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 Table 4. Specifically, we compare the performance of our approach with no variation along the spatial dimension (where all frames have a fixed spatial resolution of \( 224 \times 224 \) 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.
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 correspondence 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.

**Acknowledgment** This work is supported by Arkansas Biosciences Institute (ABI) Grant, NSF WVAR-CRESH and NSF Data Science, Data Analytics that are Robust and Trusted (DART).
References


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


[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


