RESEARCH ARTICLE

SoGAR: Self-Supervised Spatiotemporal Attention-Based Social Group Activity Recognition

NAGA VS RAVITEJA CHAPPA[1], (Member, IEEE),
PHA NGUYEN[1], (Graduate Student Member, IEEE), **ALEXANDER H. NELSON**[1],
HAN-SEOK SEO[2], **XIN LI**[3], (Fellow, IEEE), **PAGE DANIEL DOBBS**[4],
AND KHOA LUU[1], (Member, IEEE)

[1]Department of EECS, University of Arkansas, Fayetteville, AR 72701, USA
[2]Department of Food Science, University of Arkansas, Fayetteville, AR 72701, USA
[3]Department of Health, Human Performance and Recreation, University of Arkansas, Fayetteville, AR 72701, USA
[4]Department of Computer Science, University at Albany, Albany, NY 12222, USA

Corresponding author: Naga VS Raviteja Chappa (nchappa@uark.edu)

ABSTRACT Social group activity recognition is crucial for various applications including surveillance, human-robot interaction, and behavioral analysis. Current approaches often require extensive manual annotations and rely heavily on pre-trained detectors, limiting their practical applications. Additionally, existing methods struggle to effectively model long-term spatiotemporal relationships in group activities. This paper introduces a novel approach to Social Group Activity Recognition (SoGAR) using Self-supervised Transformers network that can effectively utilize unlabeled video data. To extract spatio-temporal information, we create local and global views with varying frame rates. Our self-supervised objective ensures that features extracted from contrasting views of the same video are consistent across spatio-temporal domains. Our proposed approach efficiently uses transformer-based encoders to alleviate the weakly supervised setting of group activity recognition. By leveraging the benefits of transformer models, our approach can model long-term relationships along spatio-temporal dimensions. Our proposed SoGAR method achieves state-of-the-art results on three group activity recognition benchmarks, namely JRDB-PAR, NBA, and Volleyball datasets, surpassing the current state-of-the-art in terms of F1-score, MCA, and MPCA metrics.

INDEX TERMS Group activity recognition, self-supervised learning, behavioral analysis.
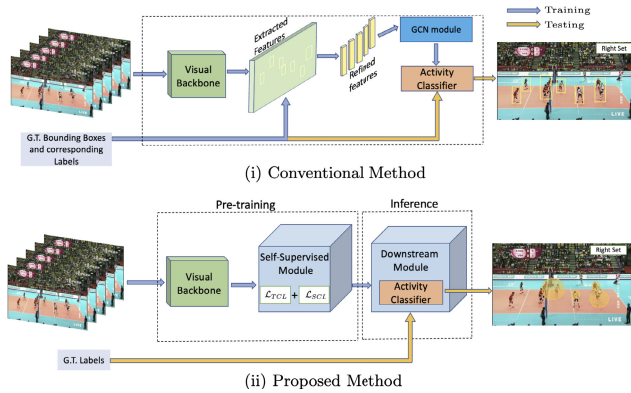
I. INTRODUCTION

Group activity recognition (GAR) has emerged as an important problem in computer vision, with numerous applications in sports video analysis, video monitoring, and social scene understanding. Unlike conventional action recognition methods that focus on identifying individual actions, GAR aims to classify the actions of a group of people in a given video clip as a whole. This requires a deeper

The associate editor coordinating the review of this manuscript and approving it for publication was Abdullah ILIYASU.

understanding of the interactions between multiple actors, including accurate localization of actors and modeling their spatiotemporal relationships [1], [2], [3], [4]. As a result, GAR poses fundamental challenges that need to be addressed in order to develop effective solutions for this problem. In this context, the development of novel techniques for group activity recognition has become an active area of research in computer vision.

Existing methods for GAR require ground-truth bounding boxes and action class labels for training and testing [5], [6], [7], [8], [9], [10], [11], [12], [13]. Bounding box labels

FIGURE 1. Overview of conventional and proposed methods for social activity recognition. The labels in the right image show the predicted labels.

are used to extract actor features and their spatio-temporal relations, which are then aggregated to form a group-level video representation for classification. However, the reliance on bounding boxes and substantial data labeling annotations severely limit their applications.

To address these limitations, some methods simultaneously train person detection and group activity recognition using bounding box labels [14], [15]. Another approach is weakly supervised GAR (WSGAR) learning [16], [17], which does not require individual actor-level labels for training and inference.

Yan et al. [16] proposed WSGAR learning approach that uses a pre-trained detector to generate actor box suggestions and learn to eliminate irrelevant possibilities. However, this method suffers from missing detections when actors are occluded. Kim et al. [17] introduced a detector-free method that captures actor information using partial contexts of token embeddings, but this method can only learn when there is movement in consecutive frames. Moreover, Kim et al. [17] did not consider the consistency of temporal information among different tokens. Hence, there is a need for a GAR approach that can capture temporal information accurately without the limitations of bounding box annotations or detector-based methods.

A. CONTRIBUTIONS OF THIS WORK

In this paper, we propose a new approach to Social Group Activity Recognition called (SoGAR). Our method is unique in that it does not require ground-truth labels during pre-training, and it doesn't rely on an object detector. Instead, our approach uses motion as a supervisory signal from the RGB data modality. Our approach is able to effectively reduce the extensive supervision present in the conventional methods, as demonstrated in Fig. 1. In fact, our method outperforms the DFWSGAR approach introduced by Kim et al. [17]. We also present the comparison of different properties between our approach and other previous methods in Table 1. To handle varying spatial and temporal details within the same deep network, we use a video transformer-based

approach, as described in [18]. This approach allows us to take advantage of varying temporal resolutions within the same architecture. Additionally, the self-attention mechanism in video transformers can capture local and global long-range dependencies in both space and time, providing much larger receptive fields compared to standard convolutional kernels [19].

The proposed SoGAR method differs from the previous methods by leveraging the correspondences from spatio-temporal features which enables the learning of long-range dependencies in both space and time domains. To facilitate this, we introduce a novel self-supervised learning strategy that does temporal collaborative learning and spatiotemporal cooperative learning. This is achieved through the proposed loss functions mentioned in III-B, which match the global features from the whole video sequence to the local features that are sampled in the latent space. Additionally, we utilize the bounding box information to localize the attention of the framework for better learning to improve overall performance. Our proposed method achieves State-of-the-Art (SOTA) performance results on the JRDB-PAR [20], NBA [16] and Volleyball [5] datasets using only the RGB inputs. We conducted extensive experiments and will publish the code for our method.

II. RELATED WORK

### A. GROUP ACTIVITY RECOGNITION (GAR)

In the field of action recognition, group action recognition has become an increasingly popular topic of research due to its wide range of applications in various fields, such as video surveillance, human-robot interaction, and sports analysis. GAR aims to identify the actions performed by a group of individuals and the interactions between them.

Initially, researchers in the field of GAR used probabilistic graphical methods and AND-OR grammar methods to process the extracted features [21], [22], [23], [24], [25], [26], [27], [28]. However, with the advancement of deep learning techniques, methods involving convolutional neural networks (CNN) and recurrent neural networks (RNN) achieved outstanding performance due to their ability to learn high-level information and temporal context [5], [14], [29], [30], [31], [32], [33], [34], [35].

Recent methods for identifying group actions typically utilize attention-based models and require explicit character representations to model spatial-temporal relations in group activities [6], [7], [8], [9], [10], [13], [16], [36], [37], [38], [39], [40], [41]. For example, graph convolution networks are used to learn spatial and temporal information of actors by constructing relational graphs, and spatial and temporal relation graphs are used to infer actor links. Clustered attention is used to capture contextual spatial-temporal information, and transformer encoder-based techniques with different backbone networks are used to extract features for learning actor interactions from multimodal inputs [8]. Additionally, MAC-Loss [42], a combination of spatial and temporal transformers in two complimentary orders, has

TABLE 1. Comparisons in the properties between our proposed approach and other methods. Actor Relation Learning (ARL), Convolutional Neural Networks (CNN), Graph Neural Networks (GNN), Graph Convolutional Networks (GCN), Transformer (TF), TimeSformer (TSformer), Vision Transformer (ViT), Space & Time (ST), Group Activity (G.A.), Individual Actions (I.A.), Bounding Boxes (B.B.)

Methods	Architecture	Source Label	Learning Mechanism	ARL Module
ARG [6]	CNN + GCN	G.A., I.A., B.B.	Fully Supervised	Graph Relational Reasoning
HiGCIN [11]	CNN + GNN	G.A., I.A., B.B.	Fully Supervised	Graph Relational Reasoning
AT [8]	CNN + TF	G.A., I.A., B.B.	Fully Supervised	Joint ST Attention
DIN [36]	CNN + GNN	G.A., I.A., B.B.	Fully Supervised	Graph Relational Reasoning
GroupFormer [13]	CNN + TF	G.A., I.A., B.B.	Fully Supervised	Clustering
Dual-AI [42]	CNN + TF	G.A., I.A., B.B.	Fully Supervised	Joint ST Attention
SAM [16]	CNN + GCN	G.A., B.B.	Weakly Supervised	Graph Relational Reasoning
DFWSGAR [17]	CNN + TF	G.A. (Training & Testing)	Weakly Supervised	Joint ST Attention
Ours	**ViT + TSformer**	**G.A.(Testing)**	**Self-Supervised**	**Divided ST Attention**
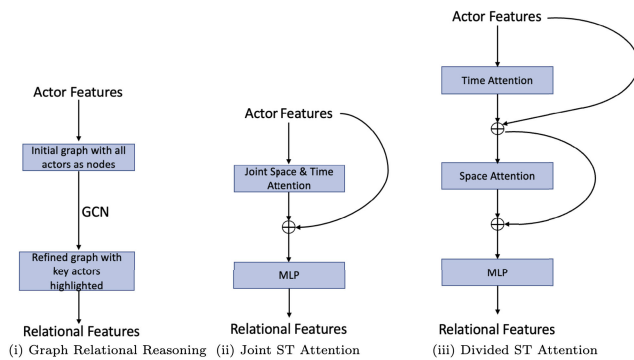
FIGURE 2. Comparison of Actor Relational Learning (ARL) Modules.

been proposed to enhance the learning effectiveness of actor interactions and preserve actor consistency at the frame and video levels. Tamura et al. [43] introduces a framework without using heuristic features for recognizing social group activities and identifying group members. This information is embedded into the features, allowing for easy identification.

Overall, these recent advancements in GAR have made significant progress toward recognizing complex actions performed by a group of individuals in various settings.

1) WEAKLY SUPERVISED GROUP ACTIVITY RECOGNITION (WSGAR)

Various techniques have been developed to address the problem of WSGAR with limited supervision, like training detectors within the framework using bounding boxes. WSGAR is one approach that does not rely on bounding box annotations during training or inference and includes an off-the-shelf item detector in the model. Traditional GAR approaches require accurate annotations of individual actors and their actions, which can be challenging and time-consuming to obtain. Weakly supervised methods aim to relax these requirements by learning from more readily available data such as activity labels, bounding boxes, or even video-level labels. Zhang et al. [44] proposed a technique that employs activity-specific characteristics to enhance WSGAR. It is not particularly designed for GAR. Kim et al. [17] proposed a detector-free approach that uses

transformer encoders to extract motion features. We propose a self-supervised training method specialized for WSGAR and does not necessitate actor-level annotations, object detectors, or labels. As shown in Fig. 2, different approaches handle actor relationships in distinct ways. Graph-based methods like ARG construct explicit relationship graphs, while attention-based methods like AT use joint space-time attention. Our approach, using divided space-time attention, allows for more flexible modeling of actor relationships across both spatial and temporal dimensions, enabling better capture of group dynamics.

2) TRANSFORMERS IN VISION

The transformer architecture was first introduced by Vaswani et al. [45] for sequence-to-sequence machine translation, and since then, it has been widely applied to various natural language processing tasks. Dosovitskiy et al. [46] introduced a transformer architecture not based on convolution for image recognition tasks. Several works [47], [48], [49], [50] used transformer architecture as a general backbone for various downstream computer vision tasks, achieving remarkable performance progress. In the video domain, many approaches [18], [51], [52], [53], [54], [55] utilize spatial and temporal self-attention to learn video representations effectively. Bertasius et al. [18] explored different mechanisms of space and time attention to learn spatiotemporal features efficiently. Fan et al. [54] used multiscale feature aggregation to improve the learning performance of features. Patrick et al. [55] introduced a self-attention block that focuses on the trajectory, which tracks the patches of space and time in a video transformer. While these transformer-based approaches have shown promising results in various vision tasks, they face several limitations when applied to group activity recognition. Methods like [46] focus primarily on static image understanding and don't effectively capture the temporal dynamics crucial for group activities. Video-based approaches [18], [54] use joint space-time attention, which can be computationally expensive and may not effectively separate spatial relationships between actors from temporal evolution of their interactions. Additionally, existing methods often struggle with long-range
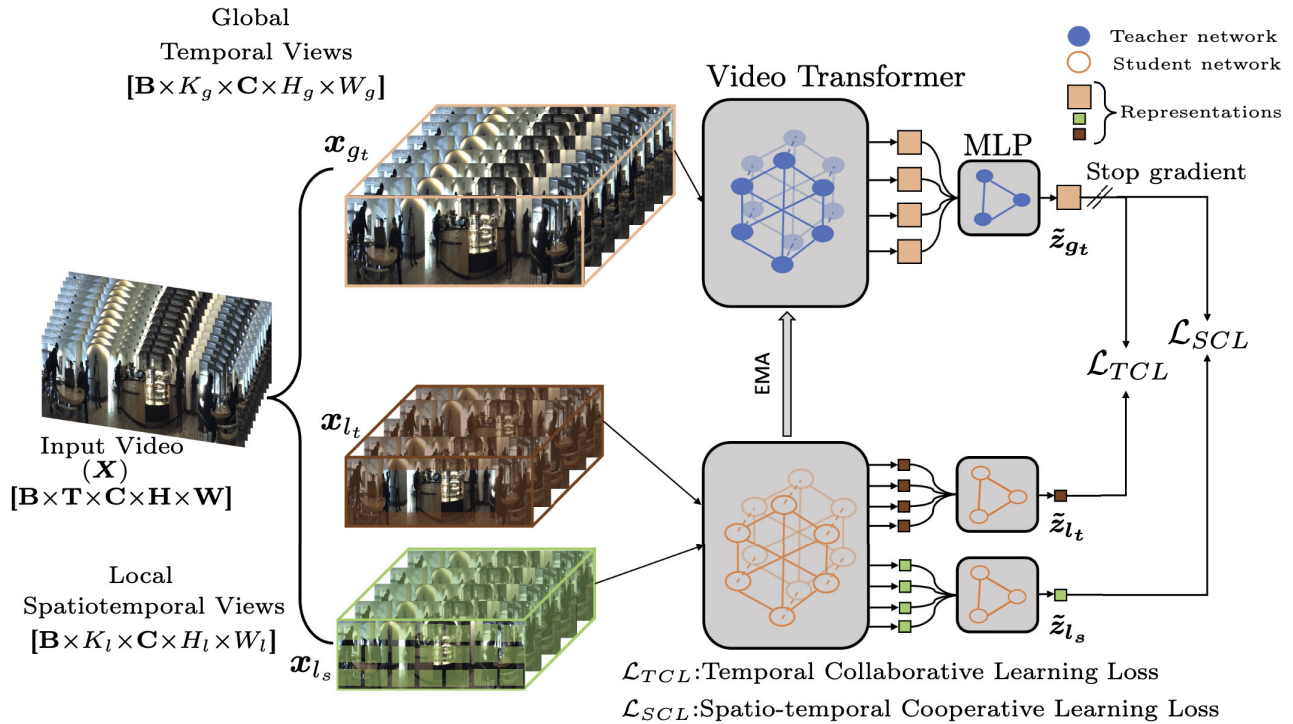
FIGURE 3. The proposed SoGAR framework adopts a sampling strategy that divides the input video into global and local views in temporal and spatial domains. Since the video clips are sampled at different rates, the global and local views have distinct spatial characteristics and limited fields of view and are subject to spatial augmentations. The teacher network takes in global views (x_{gt}) to generate a target, while the student network processes local views (x_{lt} & x_{ls}), where $Kl \leq K_g$. We update the network weights by matching the student local views to the target teacher global views, which involves both *Temporal Collaborative Learning* and *Spatio-temporal Cooperative Learning*. To accomplish this, we employ a standard ViT-Base backbone with separate space-time attention [18] and an MLP that predicts target features from student features.
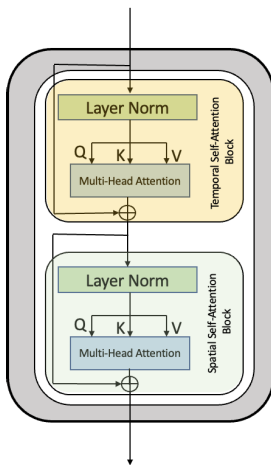
FIGURE 4. Video Transformer Block.

dependencies in group activities, where interactions between actors may span across extended temporal sequences. Our approach addresses these limitations by introducing divided space-time attention mechanisms specifically designed for group activity recognition, allowing more efficient and effective modeling of both spatial and temporal relationships.

III. THE PROPOSED METHOD

The framework presented in this paper aims to recognize social group activities in a video without depending

on a detector or person-bounding boxes. The proposed method follows a self-supervised training approach within the teacher-student framework for social group activity recognition, as depicted in Fig. 3.

Our method for video representation learning for social group activity recognition differs from other contrastive learning approaches by processing two clips from the same video while altering their spatial-temporal characteristics without requiring memory banks. This approach allows us to capture the intricate and ever-changing nature of group activities where multiple individuals may be moving in different directions and performing different actions simultaneously. Fig. 4 illustrates our Video Transformer Block architecture, which processes spatiotemporal information through separate spatial and temporal attention mechanisms. This design allows our model to efficiently capture both spatial relationships between actors and temporal evolution of group activities, while maintaining computational efficiency through divided attention.

To train our model, we propose a novel loss formulation that matches the features of two distinct clips, thereby enforcing consistency in spatial and temporal changes within the same video. Our loss function encourages the model to learn robust representations that can handle variations in spatial and temporal contexts.

The proposed SoGAR framework is described in detail in the following sections. We demonstrate the effectiveness of

our method on the newly proposed JRDB-PAR dataset [20] along with NBA [16], and Volleyball [30] datasets.

A. SELF-SUPERVISED TRAINING

Videos of social group activities capture rich temporal and spatial information, which is essential for accurate recognition. However, this high temporal dimensionality also makes it challenging to capture the various motion and spatial characteristics of group activities, such as 2p.-fail. (from NBA dataset [16]) or l-winpoint (from Volleyball dataset [5]). To address this challenge, we propose a novel approach that involves predicting different video clips with varying temporal characteristics from each other in the feature space. This approach allows us to learn contextual information that defines the underlying distribution of videos, making the network invariant to motion, scale, and viewpoint variations.

Our self-supervised training framework for video representation learning is formulated as a motion prediction problem consisting of three key components. First, we generate multiple temporal views with different numbers of clips with varying motion characteristics from the same video. Second, we vary the spatial characteristics of these views by generating local and global spatial fields of the sampled clips. Finally, we introduce a loss function that matches the varying views across spatial and temporal dimensions in the latent space.

The proposed approach for social group activity recognition involves predicting multiple video clips with varying temporal and spatial characteristics from a single video. This is achieved through a self-supervised motion prediction problem with three key components: generating multiple temporal views with different numbers of clips and varying motion characteristics, varying the spatial characteristics of these views by generating local and global spatial fields of the sampled clips, and introducing a loss function that matches the varying views across spatial and temporal dimensions in the latent space. By learning contextual information and making accurate predictions even in the presence of various motion, scale, and viewpoint variations, the network becomes invariant to these variations and can capture the complex and dynamic nature of social group activities.

1) PREDICTION OF MOTION VIA SELF-SUPERVISED LEARNING

The temporal dimension of a video is a crucial factor that can significantly affect the motion context and perception of actions captured in the content. For example, the frame rate can capture subtle nuances of body movements and affect the perception of actions, such as walking slowly versus walking quickly. Traditionally, video clips are sampled at a fixed frame rate, which may not be suitable for capturing different motion characteristics of the same action.

Our proposed approach introduces the concept of "temporal views," which refers to a collection of clips sampled at a specific video frame rate. By generating different views with varying resolutions, we can capture different motion

characteristics of the same action and learn contextual information about motion from a low frame rate input. To create motion differences among these views, we randomly sample them and process them using our ViT models. The number of temporal tokens (T) input to ViT varies in different views, allowing us to handle variability in temporal resolutions with a single ViT model.

In addition to varying temporal resolution, we vary the resolution of clips across the spatial dimension within these views. This 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. Using vanilla positional encoding [45], our approach can handle such variability in temporal resolutions with a single ViT model, unlike similar sampling strategies used under multi-network settings [56], [57].

2) ESTABLISHING CORRESPONDENCES ACROSS DIFFERENT VIEWS

Our proposed training strategy seeks to establish the interrelation between a given video's temporal and spatial dimensions. To achieve this, we introduce novel cross-view correspondences by manipulating the field of view during the sampling process. In particular, we generate global and local temporal views from a given video clip to facilitate learning these correspondences.

The global temporal views (x_{g_t}) are generated by randomly sampling K_g frames from a video clip with a fixed spatial size of W_{global} and H_{global}. These views are then fed into the teacher network, which produces an output represented by $\tilde{z}_{g_t}$.

On the other hand, the local spatiotemporal views (x_{l_t} and x_{l_s}) cover a limited portion of the video clip along both spatial and temporal dimensions. We generate these local temporal views by randomly selecting several frames (K_l), which is less than or equal to the number of frames in the global temporal views (K_g), with a spatial size fixed to W_{local} and H_{local}. These views are then fed into the student network, which produces two outputs denoted by $\tilde{z}_{l_t}$ and $\tilde{z}_{l_s}$, respectively.

We apply various data augmentation techniques to the spatial dimension by applying color jittering and gray scaling with probability 0.8 and 0.2, respectively, to all temporal views. Moreover, we apply Gaussian blur and solarization with probability 0.1 and 0.2, respectively, to global temporal views.

Our approach is based on the idea that training the model 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. More specifically, our method encourages the model to consider both the spatial and temporal context of the video, where the spatial context denotes the possibilities surrounding a given spatial crop, and the temporal context denotes possible previous or future clips from a given temporal crop. It is essential to note that spatial correspondences also involve a temporal component, as our

approach seeks 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.

B. THE PROPOSED OBJECTIVE FUNCTION

Our model aims to predict different views of the same video, capturing various spatial-temporal variations. To achieve this, we train our model with an objective function that leverages global and local temporal and spatial views.

Let $X = x_t^T$ be a video consisting of T frames, where x_{g_t}, x_{l_t}, and x_{l_s} represent global temporal views, local temporal views, and local spatial views, respectively. Specifically, x_{g_t} contains Kg frames, while x_{l_t} and x_{l_s} both contain K_l frames, where $K_l \leq K_g$ and K_g and K_l are the numbers of frames for teacher and student (global and local) inputs. We randomly sample K_g global and K_l local temporal views as described in III-A2. The student and teacher models process the temporal views to obtain class tokens or features z_g and z_l. We then normalize these class tokens to facilitate training with the objective function.

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

where τ is a temperature parameter used to control the sharpness of the exponential function [58] and $z^{(i)}$ is each element in $\tilde{z}^{(i)} \in \mathbb{R}^n$.
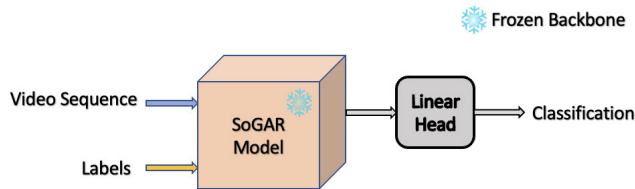
FIGURE 5. Inference. We input the video sequence along with their corresponding labels. The output from the model is fed to the downstream task classifier.

1) TEMPORAL COLLABORATIVE LEARNING LOSS (TCL)

Our x_{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 x_{g_t} always passes through the teacher model that serves as the target label. We map the student's x_{l_t} with the teacher's x_{g_t} to create a global-to-local temporal loss as in (2)

$$\mathcal{L}_{TCL} = -sg(\tilde{z}_{g_t}) * log(\tilde{z}_{l_t}), \qquad (2)$$

where $\tilde{z_{g_t}}$ and $\tilde{z_{l_t}}$ are the tokens of the class for x_{g_t} and x_{l_t} produced by the teacher and student, and sg is the stochastic gradient respectively.

2) SPATIO-TEMPORAL COOPERATIVE LEARNING LOSS (SCL)

The local temporal views x_{l_t} in our approach have a smaller field of vision compared to the global temporal views x_{g_t}, both along the spatial and temporal dimensions. Despite this,

the number of local views is four times higher than that of global views. The student model processes all the local views x_{l_s}, while the teacher model processes only the global views x_{g_t}, which serve as the target. To create the loss function, the local views are mapped to the global views using the teacher model, as described in (3).

$$\mathcal{L}_{SCL} = \sum_{n=1}^{q} -sg(\tilde{z}_{g_t}) * log(\tilde{z}_{l_s}^{(n)}), \qquad (3)$$

where $\tilde{z_{l_s}}$ are the tokens of the class for x_{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 (2) and (3), given as in (4),

$$\mathcal{L} = \mathcal{L}_{TCL} + \mathcal{L}_{SCL} \qquad (4)$$

C. INFERENCE

Our inference framework is depicted in Fig. 5. In this stage, we perform fine-tuning of the self-supervised model that was trained earlier. Specifically, we utilize the pre-trained SoGAR model and fine-tune it with the available labels. This is followed by a linear classifier, and the resulting model is applied to downstream tasks to enhance the overall performance.

IV. EXPERIMENTS
A. DATASETS
1) VOLLEYBALL DATASET

[5] is composed of 55 videos, containing a total of 4,830 labeled clips, including 3,493 for training and 1,337 for testing. The dataset provides annotations for both individual actions and group activities with corresponding bounding boxes. However, in our WSGAR experiments, we only focus on the group activity labels and exclude the individual action annotations. To evaluate our model, we use Multi-class Classification Accuracy (MCA) and Merged MCA metrics. The Merged MCA metric merges the right set and right pass classes into the right pass-set and the left set and left pass classes into the left pass-set, as in previous works like SAM [16] and DFWSGAR [17], to ensure a fair comparison with existing methods.

2) NBA DATASET

Reference [16] used in our experiments contains a total of 9,172 labeled clips from 181 NBA videos, where 7,624 clips are for training and 1,548 for testing. The dataset only provides annotations for group activities and lacks information about individual actions or bounding boxes. For evaluating the model, we use the Multi-class Classification Accuracy (MCA) and Mean Per Class Accuracy (MPCA) metrics. The MPCA metric is used to address the issue of class imbalance in the dataset.

3) JRDB-PAR DATASET

Reference [20] containing 27 categories of individual actions such as walking, talking, etc., 11 categories of social

group activities, and 7 categories of global activities. The dataset consists of 27 videos, which are split into 20 for training and 7 for testing, following the training/validation splitting in JRDB dataset [59]. In total, the dataset contains 27,920 frames with over 628k human bounding boxes. For annotation and evaluation, uniformly sampled keyframes (one keyframe in every 15 frames) are selected, which is consistent with other group activity datasets like CAD [60] and Volleyball [30]. The dataset uses multi-class labels for activity annotation, with each individual/group/frame having multiple activity labels. Following [20], we use the precision, recall, and F1-score (denoted as $\mathcal{P}_g$, $\mathcal{R}_g$, $\mathcal{F}_g$) for evaluation, since social group activity recognition can be considered as a multi-label classification problem.

B. DEEP NETWORK ARCHITECTURE

Our video processing technique employs a Vision Transformer (ViT) [18] to apply attention to both the spatial and temporal dimensions of video clips. The ViT comprises 12 encoder blocks and can handle video clips of size ($B \times T \times C \times W \times H$), where B and C denote the batch size and the number of color channels, respectively. The maximum spatial and temporal sizes are $W = H = 480$ and $T = 18$, respectively, indicating that we extract 18 frames from each video and resize them to 480×480. Our network architecture (see Fig. 3) is designed to accommodate varying input resolution during training, including 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, with each token having an embedding dimension of $\mathbb{R}^m$ [46]. In addition to these spatial and temporal input tokens, we include a single classification token within the architecture as a characteristic vector [61]. This classification token captures the standard features learned by the ViT across the spatial and temporal dimensions of a given video. During training, we use varying spatial and temporal resolutions that satisfy $W \leq 480$, $H \leq 480$, and $T \leq 18$, resulting in different spatial and temporal tokens. Finally, we apply a projection head to the class token of the last ViT encoder [58], [62].

1) SELF-DISTILLATION

Our approach, depicted in Fig. 3, employs a teacher-student setup for self-distillation based on the methodology proposed in [58] and [62]. The teacher and student models share the same architecture, consisting of a ViT backbone and a predictor MLP. However, only the student model is directly trained, while the teacher model is updated through an exponential moving average (EMA) of the student weights at each training step [58]. This design allows us to use a unified network to process various input clips.

C. IMPLEMENTATION DETAILS

To prepare the JRDB-PAR, NBA and Volleyball datasets for our analysis, we sampled frames at a rate of T (K_g) using segment-based sampling. The choice of Kg was determined by the characteristic temporal span of activities in each dataset. For Volleyball, Kg=5 was selected as volleyball group activities typically complete within 5 frames. For NBA, Kg=18 was chosen to capture the longer sequences of basketball plays involving multiple player interactions. For JRDB-PAR, Kg=8 was selected based on the dataset's frame sampling rate of 15 frames per clip to maintain temporal coherence of social group activities. Next, we resized the frames to $W_g = 480$ & $H_g = 480$ for the teacher input and $W_l = 96$ & $H_l = 96$ for the student input. In the case of the Volleyball dataset, we set K_g to 5 ($K_l \in 3, 5$), while for the NBA dataset, we set K_g to 18 ($K_l \in 2, 4, 8, 16, 18$). For JRD-PAR dataset, we used K_g to 8 ($K_l \in 2, 4, 8, 16, 18$). We initialized temporal attention weights randomly, while spatial attention weights were initialized using a ViT model trained self-supervised over ImageNet-1K [63]. This initialization scheme facilitated faster convergence of space-time ViT, as seen in the supervised setting [18]. We trained using an Adam optimizer [64] with a learning rate of 5×10^{-4}, scaled using a cosine schedule with a linear warm-up over five epochs [65], [66]. Additionally, we applied weight decay scaled from 0.04 to 0.1 during training. For the downstream task, we trained a linear classifier on our pretrained SPARTAN backbone. During training, the backbone was frozen, and we trained the classifier 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. For datasets with available bounding box annotations (JRDB-PAR and Volleyball), we utilized these coordinates during pre-training to generate local views centered around actor locations, without using their corresponding action labels. For NBA dataset, which lacks bounding box annotations, random spatial crops were used for local views.

D. COMPARISON WITH STATE-OF-THE-ART METHODS
1) JRDB-PAR DATASET

We conducted a comparative study to evaluate our proposed approach alongside state-of-the-art methods in GAR and WSGAR using the JRDB-Par dataset. We involved fully supervised and weakly supervised settings to evaluate the dataset. The comparison results are presented in Table 2. In the fully supervised setting, our method outperforms the existing social group activity recognition frameworks significantly in all the metrics. In the weakly supervised setting, our proposed method outperformed existing GAR and WSGAR methods by a considerable margin, achieving 8.7 of $\mathcal{P}_g$, 12.7 of $\mathcal{R}_g$ and 9.9 of $\mathcal{F}_g$. Additionally, we evaluated this dataset using ResNet-18 and ViT-Base backbones, where ViT-Base proved to be better, which is analyzed in the ablation study section. Despite their impressive performance in WSGAR, our approach outperformed them all.

TABLE 2. Comparative results of the social group activity recognition on JRDB-PAR dataset [20].

Method	Group Activity		
	$\mathcal{P}_g$	$\mathcal{R}_g$	$\mathcal{F}_g$
Fully supervised			
ARG [6]	34.6	29.3	30.7
SA-GAT [10]	36.7	29.9	31.4
JRDB-Base [60]	44.6	46.8	45.1
Ours	**49.3**	**47.1**	**48.7**
Weakly supervised			
AT [8]	21.2	19.1	19.8
SACRF [9]	42.9	35.5	37.6
Dynamic [36]	37.5	27.1	30.6
HiGCIN [11]	39.3	30.1	33.1
ARG [6]	26.9	21.5	23.3
SA-GAT [10]	28.6	24.0	25.5
JRDB-Base [60]	38.4	33.1	34.8
Ours	**47.1**	**45.8**	**44.9**

TABLE 3. Comparisons with the State-of-the-Art GAR models and video backbones on the NBA dataset [16].

Method	MCA	MPCA
Video backbone		
TSM [68]	66.6	60.3
VideoSwin [69]	64.3	60.6
GAR model		
ARG [6]	59.0	56.8
AT [8]	47.1	41.5
SACRF [9]	56.3	52.8
DIN [36]	61.6	56.0
SAM [16]	54.3	51.5
DFWSGAR [17]	75.8	71.2
Ours	**83.3**	**73.5**

2) NBA DATASET

Table 3 lists the outcomes of our comparison study on NBA dataset. Our approach outperforms existing GAR and WSGAR methods significantly, achieving 7.5% MCA and 2.3% MPCA. SAM's results [16] from [17] are also listed. RGB frames are exclusively used as input to ensure a fair comparison across approaches and video backbones, including ResNet-18 TSM [67] and VideoSwin-T [68]. Comparing our approach to these strong backbones, our method prevails. Evaluating our proposed approach against current video backbones and state-of-the-art methods in GAR and WSGAR, our comparison study utilizes the NBA dataset. Notably, results of SAM [16] are referenced from [17].

3) VOLLEYBALL DATASET

In the volleyball dataset, we reproduce results using only the RGB input and ResNet-18 backbone, respectively, to ensure a fair comparison. To have consistent comparison, we compare our approach against the latest GAR and WSGAR methods in two supervision levels: fully supervised and weakly supervised. The results show that our ResNet-18 trained model surpasses most fully supervised frameworks, showing a remarkable enhancement in MCA and MPCA metrics. The first and second sections display the outcomes of earlier techniques in fully supervised and weakly supervised contexts, respectively. Employing the ViT-Base backbone, our approach excels in weakly supervised conditions,

outperforming all GAR and WSGAR models. By utilizing the transformer architecture to leverage spatiotemporal features, we achieve a significant lead of 2.4% in MCA and 1.2% in Merged MCA. Notably, these levels differ in their use of actor-level labels like ground-truth bounding boxes and individual action class labels during training and inference. In the weakly supervised setting, the group action classification labels are substituted with ground-truth bounding boxes of actors minus their corresponding actions. Table 4 showcases the results. Additionally, our approach fares better than current GAR methods employing less comprehensive actor-level supervision, such as [8], [9], [14], [32], and [35].

TABLE 4. Comparison with the state-of-the-art methods on the Volleyball dataset. [5].

Method	Backbone	MCA	Merged MCA
Fully supervised			
SSU [14]	Inception-v3	89.9	-
PCTDM [35]	ResNet-18	90.3	94.3
StagNet [32]	VGG-16	89.3	-
ARG [6]	ResNet-18	91.1	95.1
CRM [70]	I3D	92.1	-
HiGCIN [11]	ResNet-18	91.4	-
AT [8]	ResNet-18	90.0	94.0
SACRF [9]	ResNet-18	90.7	92.7
DIN [36]	ResNet-18	93.1	**95.6**
TCE+STBiP [12]	VGG-16	**94.1**	-
GroupFormer [13]	Inception-v3	**94.1**	-
Weakly supervised			
PCTDM [35]	ResNet-18	80.5	90.0
ARG [6]	ResNet-18	87.4	92.9
AT [8]	ResNet-18	84.3	89.6
SACRF [9]	ResNet-18	83.3	86.1
DIN [36]	ResNet-18	86.5	93.1
SAM [16]	ResNet-18	86.3	93.1
DFWSGAR [17]	ResNet-18	90.5	94.4
Ours	ResNet-18	91.8	94.5
	ViT-Base	**93.1**	**95.9**

E. ABLATION STUDY

We conduct a thorough analysis of the various components that contribute to the effectiveness of our approach, which is an extension of analysis from [70]. In particular, we assess the impact of five distinct elements: a) Impact of different backbone networks, b) Impact of knowledge distillation, and c) Impact of ground-truth bounding box information

1) DIFFERENT BACKBONE NETWORKS

We investigated the effect of different backbone networks on our framework. We conducted the experiments presented in Table 5. Our results show that ResNet-18 performs better than the other Convolutional Neural Network (CNN) backbones, but overall performance is optimal with ViT-Base backbone because the spatiotemporal features of the input video with varying views are well leveraged by the transformer architecture for videos [18]. Also, when both networks share the same backbone, they perform better rather than having distinct backbone networks.

TABLE 5. Different backbones. The most optimal backbone for our framework is ViT-Base outperforming the other backbones.

Backbone	JRDB-PAR ($\mathcal{F}_g$)	NBA (MCA)	Volleyball (MCA)
Inception-v3	31.8	69.3	78.6
VGG-16	35.1	72.9	81.5
I3D	36.3	76.7	85.8
ResNet-18	39.6	78.1	89.2
ViT-S	41.3	80.2	91.1
ViT-B	**44.9**	**83.3**	**93.1**

2) IMPACT OF KNOWLEDGE DISTILLATION (KD)

To evaluate the effect of knowledge distillation, we conducted experiments as presented in Table 6. To be specific, we compared the performance of our approach in the absence of KD, i.e., the student and teacher networks learn independently, and there is no transfer of information from the student to teacher network. This shows very poor performance. Hence, KD is determined to be one of the key factors in the optimal performance of the proposed framework. This also proves that exponential moving average (EMA) aids feature learning across the networks to improve performance.

TABLE 6. Impact of Knowledge Distillation (KD): The framework is proved to work better when there is knowledge distillation with EMA which infers student-teacher network learns the spatiotemporal features for different views on all the datasets.

KD	JRDB-PAR	NBA	Volleyball
✗	34.2	75.2	86.4
✓	**44.9**	**83.3**	**93.1**

3) IMPACT OF GROUND-TRUTH BOUNDING BOX (G.T. BB'S) INFORMATION

To evaluate the impact of actor localization information, we compared two approaches for generating local views: (1) using ground-truth bounding box coordinates to create local views centered around actors, and (2) using random spatial crops without any prior location information. The G.T. BB's were only used during pre-training to guide the attention mechanism, without utilizing their corresponding action labels. During the pre-training step, the social group activity recognition is highly leveraged by the actor localization information. So, we perform experiments as shown in Table 7 to evaluate the performance of our method on this information. Specifically, we used random crops in the initial experiment in all the input views, which yields poor performance for JRDB-PAR and Volleyball datasets but the NBA dataset performs well as there is no bounding box information from the dataset. In contrast, we used the G.T. BB's exclusively without their corresponding labels for the other experiment to prove the optimal performance of our method.

F. QUALITATIVE RESULTS

We conducted an analysis to understand how our method aggregates feature for various social group activities.

TABLE 7. Impact of ground-truth bounding box information (G.T. BB's)): When we provided the bounding box information during the pre-training, it is proved that the performance is optimal rather than using random crops.

|	JRDB-PAR	NBA	Volleyball
Random Crops	40.7	**83.3**	88.5
G.T. BB's	**44.9**	-	**93.1**

We visualized the attention locations of the transformer encoder in Fig. 6 and Fig. 7 for JRDB-PAR and Volleyball datasets, showing locations with the top five and top four attention weights in the last layer of the encoder. The yellow circles represent the attention locations. The size of the yellow circles denotes whether the locations are in the high or low-resolution feature maps, giving a rough indication of the image areas affecting the generated features. Our findings reveal that features are generally aggregated from low-resolution feature maps when group members are situated in broader areas, and the opposite is true. These results indicate that the proposed framework can effectively aggregate features based on the distribution of group members, thereby contributing to improving the performance of social group activity recognition. While we present qualitative results for JRDB-PAR and Volleyball datasets here, future work will extend this analysis to the NBA dataset to provide a comprehensive understanding of our model's attention mechanisms across different sports scenarios.
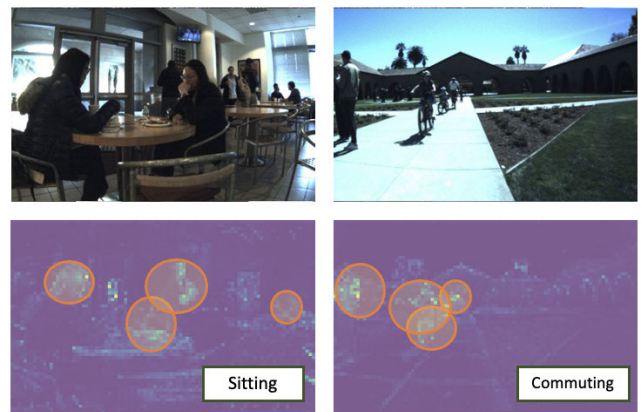
FIGURE 6. Visualization of the attention locations on the JRDB-PAR dataset. We show the locations of the top five attention weights from the transformer heads.

G. LIMITATIONS

While our method achieves state-of-the-art performance, it has several limitations:

- The self-supervised training phase requires a large amount of unlabeled video data for optimal performance.
- The current approach may struggle with very crowded scenes where multiple group activities occur simultaneously.
- Performance can be affected when there are extreme lighting changes or severe occlusions.
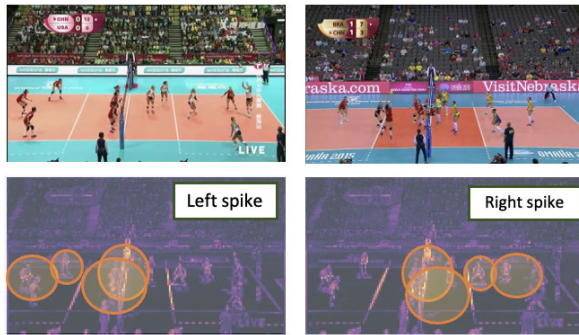
FIGURE 7. Visualization of the attention locations on the Volleyball dataset. We show the locations of the top four attention weights from the transformer heads.

- The method assumes relatively stable camera views and may need adaptations for highly dynamic camera movements.

V. CONCLUSION AND FUTURE WORK

Our paper presents a new self-supervised video model named SoGAR, which is based on a video transformer architecture. The method entails generating multiple views of a video, which differ in terms of their spatial and temporal characteristics. To capture the motion characteristics and cross-view relationships between the clips, we define two sets of correspondence learning tasks. The self-supervised objective is to reconstruct one view from another in the latent space of both the teacher and student networks. Furthermore, our SoGAR model can capture long-term spatio-temporal dependencies and perform dynamic inference within a single framework. We evaluate SoGAR on three benchmark datasets for social group activity recognition and demonstrate its superior performance over existing state-of-the-art models.

Future works including extensive qualitative analysis across all datasets, including visualization of attention patterns for basketball plays in the NBA dataset, to provide deeper insights into how the model handles different types of group activities and sports scenarios. Also, providing comprehensive computational analysis comparing the runtime efficiency of our approach with existing methods, providing insights into the practical deployment considerations of different group activity recognition approaches will benefit the research community.

REFERENCES

[1] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland. Springer, Jan. 2016, pp. 20–36.

[2] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[3] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[4] K. Ranasinghe, M. Naseer, S. Khan, F. S. Khan, and M. S. Ryoo, "Self-supervised video transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2874–2884.

[5] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1971–1980.

[6] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu, "Learning actor relation graphs for group activity recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9964–9974.

[7] G. Hu, B. Cui, Y. He, and S. Yu, "Progressive relation learning for group activity recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 980–989.

[8] K. Gavrilyuk, R. Sanford, M. Javan, and C. G. M. Snoek, "Actor-transformers for group activity recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 839–848.

[9] R. R. A. Pramono, Y. T. Chen, and W. H. Fang, "Empowering relational network by self-attention augmented conditional random fields for group activity recognition," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, Aug. 2020, pp. 71–90.

[10] M. Ehsanpour, A. Abedin, F. S. Saleh, Q. Shi, I. Reid, and H. Rezatofighi, "Joint learning of social groups, individuals action and sub-group activities in videos," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2020, pp. 177–195.

[11] R. Yan, L. Xie, J. Tang, X. Shu, and Q. Tian, "HiGCIN: Hierarchical graph-based cross inference network for group activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6955–6968, Jun. 2020.

[12] H. Yuan and D. Ni, "Learning visual context for group activity recognition," in *Proc. AAAI*, May 2021, vol. 35, no. 4, pp. 3261–3269.

[13] S. Li, Q. Cao, L. Liu, K. Yang, S. Liu, J. Hou, and S. Yi, "GroupFormer: Group activity recognition with clustered spatial–temporal transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13648–13657.

[14] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese, "Social scene understanding: End-to-end multi-person action localization and collective activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4315–4324.

[15] P. Zhang, Y. Tang, J.-F. Hu, and W.-S. Zheng, "Fast collective activity recognition under weak supervision," *IEEE Trans. Image Process.*, vol. 29, pp. 29–43, 2019.

[16] R. Yan, L. Xie, J. Tang, X. Shu, and Q. Tian, "Social adaptive module for weakly-supervised group activity recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2020, pp. 208–224.

[17] D. Kim, J. Lee, M. Cho, and S. Kwak, "Detector-free weakly supervised group activity recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20083–20093.

[18] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proc. ICML*, 2021, vol. 2, no. 3, p. 4.

[19] M. Naseer, K. Ranasinghe, S. Khan, M. Hayat, F. Shahbaz Khan, and M.-H. Yang, "Intriguing properties of vision transformers," 2021, *arXiv:2105.10497*.

[20] R. Han, H. Yan, J. Li, S. Wang, W. Feng, and S. Wang, "Panoramic human activity recognition," in *Proc. Eur. Conf. Comput. Vis.*, Tel Aviv, Israel. Cham, Switzerland: Springer, Oct. 2022, pp. 244–261.

[21] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S. Zhu, "Cost-sensitive top-down/bottom-up inference for multiscale activity recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2012, pp. 187–200.

[22] M. R. Amer, S. Todorovic, A. Fern, and S.-C. Zhu, "Monte Carlo tree search for scheduling activity recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1353–1360.

[23] M. R. Amer, P. Lei, and S. Todorovic, "HiRF: Hierarchical random field for collective activity recognition in videos," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2014, pp. 572–585.

[24] M. R. Amer and S. Todorovic, "Sum product networks for activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 800–813, Apr. 2015.

[25] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1549–1562, Aug. 2012.

[26] T. Lan, L. Sigal, and G. Mori, "Social roles in hierarchical models for human activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1354–1361.

[27] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S.-C. Zhu, "Joint inference of groups, events and human roles in aerial videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4576–4584.

[28] Z. Wang, Q. Shi, C. Shen, and A. van den Hengel, "Bilinear programming for human activity recognition with unknown MRF graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1690–1697.

[29] Z. Deng, A. Vahdat, H. Hu, and G. Mori, "Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4772–4781.

[30] M. S. Ibrahim and G. Mori, "Hierarchical relational networks for group activity recognition and retrieval," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 721–736.

[31] X. Li and M. C. Chuah, "SBGAR: Semantics based group activity recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2876–2904.

[32] M. Qi, J. Qin, A. Li, Y. Wang, J. Luo, and L. V. Gool, "StagNet: An attentive semantic RNN for group activity recognition," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2018, pp. 101–117.

[33] X. Shu, J. Tang, G.-J. Qi, W. Liu, and J. Yang, "Hierarchical long short-term concurrent memory for human interaction recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1110–1118, Mar. 2021.

[34] M. Wang, B. Ni, and X. Yang, "Recurrent modeling of interaction context for collective activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3048–3056.

[35] R. Yan, J. Tang, X. Shu, Z. Li, and Q. Tian, "Participation-contributed temporal dynamic model for group activity recognition," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1292–1300.

[36] H. Yuan, D. Ni, and M. Wang, "Spatio-temporal dynamic inference network for group activity recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7456–7465.

[37] N. V. S. R. Chappa, P. Nguyen, T. H. N. Le, P. D. Dobbs, and K. Luu, "HAtt-flow: Hierarchical attention-flow mechanism for group-activity scene graph generation in videos," *Sensors*, vol. 24, no. 11, p. 3372, May 2024.

[38] N. Venkata Sai Raviteja Chappa and K. Luu, "LiGAR: LiDAR-guided hierarchical transformer for multi-modal group activity recognition," 2024, *arXiv:2410.21108*.

[39] N. V. S. R. Chappa, P. Nguyen, P. D. Dobbs, and K. Luu, "React: Recognize every action everywhere all at once," *Mach. Vis. Appl.*, vol. 35, no. 4, p. 102, Jul. 2024.

[40] N. V. R. Chappa, P. D. Dobbs, B. Raj, and K. Luu, "FLAASH: Flow-attention adaptive semantic hierarchical fusion for multi-modal tobacco content analysis," 2024, *arXiv:2410.19896*.

[41] T.-D. Truong, R. T. N. Chappa, X.-B. Nguyen, N. Le, A. P. G. Dowling, and K. Luu, "OTAdapt: Optimal transport-based approach for unsupervised domain adaptation," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 2850–2856.

[42] M. Han, D. J. Zhang, Y. Wang, R. Yan, L. Yao, X. Chang, and Y. Qiao, "Dual-AI: Dual-path actor interaction learning for group activity recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2990–2999.

[43] M. Tamura, R. Vishwakarma, and R. Vennelakanti, "Hunting group clues with transformers for social group activity recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2022, pp. 19–35.

[44] Y. Zhang, X. Li, and I. Marsic, "Multi-label activity recognition using activity-specific features and activity correlations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14625–14635.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, vol. 30, Jun. 2017, pp. 5998–6008.

[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[47] M. Li, W. Cai, R. Liu, Y. Weng, X. Zhao, C. Wang, X. Chen, Z. Liu, C. Pan, M. Li, Y. Zheng, Y. Liu, F. D. Salim, K. Verspoor, X. Liang, and X. Chang, "FFA-IR: Towards an explainable and reliable medical report generation benchmark," in *Proc. 35th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track (Round 2)*, Aug. 2021, pp. 1–13.

[48] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-Token ViT: Training vision transformers from scratch on ImageNet," 2021, *arXiv:2101.11986*.

[49] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[50] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," 2021, *arXiv:2102.12122*.

[51] M. Han, Y. Wang, X. Chang, and Y. Qiao, "Mining inter-video proposal relations for video object detection," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland. Springer, Jan. 2020, pp. 431–446.

[52] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," 2021, *arXiv:2103.15691*.

[53] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "UniFormer: Unifying convolution and self-attention for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12581–12600, Oct. 2022.

[54] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," 2021, *arXiv:2104.11227*.

[55] M. Patrick, D. Campbell, Y. M. Asano, I. Misra, F. Metze, C. Feichtenhofer, A. Vedaldi, and J. F. Henriques, "Keeping your eye on the ball: Trajectory attention in video transformers," in *Proc. NeurIPS*, Dec. 2021, pp. 12493–12506.

[56] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6202–6211.

[57] K. Kahatapitiya and M. S. Ryoo, "Coarse-fine networks for temporal activity detection in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8381–8390.

[58] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9630–9640.

[59] M. Ehsanpour, F. Saleh, S. Savarese, I. Reid, and H. Rezatofighi, "JRDB-act: A large-scale dataset for spatio-temporal action, social group and activity detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20983–20992.

[60] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2011, pp. 3273–3280.

[61] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[62] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1–14.

[63] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, Dec. 2014, pp. 1–15.

[65] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your ViT? Data, augmentation, and regularization in vision transformers," 2021, *arXiv:2106.10270*.

[66] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," 2021, *arXiv:2104.02057*.

[67] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7083–7093.

[68] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video Swin transformer," 2021, *arXiv:2106.13230*.

[69] S. M. Azar, M. G. Atigh, A. Nickabadi, and A. Alahi, "Convolutional relational machine for group activity recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7892–7901.

[70] N. V. S. R. Chappa, P. Nguyen, A. H. Nelson, H.-S. Seo, X. Li, P. D. Dobbs, and K. Luu, "SPARTAN: Self-supervised spatiotemporal transformers approach to group activity recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 5157–5167.

NAGA VS RAVITEJA CHAPPA (Member, IEEE) received the B.Tech. degree in electronics and communication engineering from the MVGR College of Engineering (affiliated to JNTU-K), India, in 2018, and the master's degree in computer engineering major from the Electrical and Computer Engineering, Purdue University, Indianapolis. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Computer Engineering, University of Arkansas. His research interests include vision-language models, group behavior analysis and their activity analysis, video understanding, and self-supervised learning-based methods. He also serves as a Reviewer for CVPR, NeurIPS, ICLR, ICCV, ECCV, ICML, WACV, IEEE Access, MTAP journal, and *Pattern Recognition*.
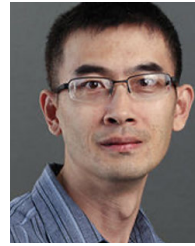
XIN LI (Fellow, IEEE) received the B.S. degree (Hons.) in electronic engineering and information science from the University of Science and Technology of China, Hefei, in 1996, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 2000. He was a member of the Technical Staff with the Sharp Laboratories of America, Camas, WA, USA, from August 2000 to December 2002. From January 2003 to August 2023, he was a Professor of lane with the Department of Computer Science and Electrical Engineering. He is currently a Professor with the Department of Computer Science, University at Albany. His contributions to image interpolation, restoration, and compression.

PHA NGUYEN (Graduate Student Member, IEEE) received the B.Sc. degree from Vietnam National University Ho Chi Minh City-University of Science in Computer Science, in 2020. He is currently pursuing the Ph.D. degree in computer science. He is a Research Assistant with the Computer Vision and Image Understanding (CVIU) Laboratory, University of Arkansas, USA. His particular research interests are focused on deep learning, machine learning, and their applications in computer vision, primarily multiple object tracking. He also serves as a reviewer for CVPR, ICCV, ECCV, and IEEE Access.

PAGE DANIEL DOBBS received the M.S. and Ph.D. degrees in community health promotion from the College of Education and Health Professions, University of Arkansas. She joined the University of Arkansas, in 2020. Her research focuses on social factors that influence perceptions and behaviors among vulnerable populations. She has published quantitative, qualitative, and mixed-methods research that examined the use of cigarettes and electronic cigarettes (e-cigarettes) among youth, young adults, minority populations, and pregnant women. She received the Career Award (Loopholes, Enforcement Challenges, and Tobacco Industry Interference with Tobacco Control Policies) from the National Institute of Health, entitled, "Loopholes, Enforcement Challenges, and Tobacco Industry Interference with Tobacco Control Policies."

ALEXANDER H. NELSON received the Ph.D. degree from the University of Maryland, Baltimore County (UMBC), in 2017. He is currently an Assistant Professor with the Department of Computer Science and Computer Engineering, University of Arkansas. His primary research interests include wearable and ubiquitous computing. At the University of Arkansas, Alexander leads the ÆSIR (Applied Embedded Systems and IoT Research) Laboratory.

HAN-SEOK SEO is currently a Professor and the Director of the University of Arkansas Sensory Science Center, Department of Food Science, University of Arkansas, Fayetteville, AR, USA. He is a Creative Sensory Scientist who combines multidisciplinary backgrounds and skills in order to contribute to improved quality of life through healthy and happy eating behavior. His research interests include identifying mechanisms of multisensory interaction and integration with a focus on chemosensory cues, developing methods to improve eating quality, creating novel methodology of sensory evaluation, and investigating impacts of sensory disorders on eating quality. He has published more than 120 articles in peer-reviewed journals, and he serves as an Editorial Board Member of multiple journals, including *Journal of Sensory Studies*, *Food Quality and Preference*, *Foods*, *Journal of Culinary Science and Technology*, *Journal of Food Science*, and *Korean Journal of Food and Cookery Science*. He also serves as an Associate Editor and a Section Editor for *Food Research International* and *Current Opinion in Food Science*, respectively.

KHOA LUU (Member, IEEE) is currently an Assistant Professor and the Director of the Computer Vision and Image Understanding (CVIU) Laboratory, Department of Computer Science and Computer Engineering, University of Arkansas, Fayetteville. He was the Research Project Director of the Cylab Biometrics Center, Carnegie Mellon University (CMU), USA. He has received four patents and two best paper awards and co-authored more than 100 papers in conferences and journals. He was the Vice-Chair of Montreal Chapter IEEE SMCS in Canada, from September 2009 to March 2011. He is teaching computer vision, image processing, and introduction to artificial intelligence courses with the CSCE Department, University of Arkansas. His research interests include various topics, including biometrics, image processing, computer vision, machine learning, deep learning, multifactor analysis, and compressed sensing. He is a PC member of AAAI and ICPRAI, in 2020 and 2021. He is a Co-Organizer and the Chair of the CVPR Precognition Workshop, in 2019, 2020, and 2021; the MICCAI Workshop, in 2019 and 2020; and the ICCV Workshop, in 2021. He is a reviewer for several top-tier conferences and journals, such as CVPR, ICCV, ECCV, NeurIPS, ICLR, FG, BTAS, IEEE-TPAMI, IEEE-TIP, *Journal of Pattern Recognition*, *Journal of Image and Vision Computing*, *Journal of Signal Processing*, *Journal of Intelligence Review*, IEEE Access, and Transactions.

∙ ∙ ∙