Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

REACT: Recognize Every Action Everywhere All At Once

Naga VS Raviteja Chappa

nchappa@uark.edu

University of Arkansas at Fayetteville

Pha Nguyen

University of Arkansas at Fayetteville

Page Daniel Dobbs

University of Arkansas at Fayetteville

Khoa Luu

University of Arkansas at Fayetteville

Research Article

Keywords: Group Activity Recognition (GAR), Action Retrieval, Vision-Language Modeling

Posted Date: March 19th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4109494/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Machine Vision and Applications on July 1st, 2024. See the published version at https://doi.org/10.1007/s00138-024-01561-z.

REACT: Recognize Every Action Everywhere All At Once

Naga VS Raviteja Chappa[1*], Pha Nguyen[1], Page Daniel Dobbs[2], Khoa Luu[1]

[1*]Department of EECS, University of Arkansas, Fayetteville, 72701, Arkansas, USA.
[2]Center for Public Health and Technology, University of Arkansas, Fayetteville, 72701, Arkansas, USA.

*Corresponding author(s). E-mail(s): nchappa@uark.edu;
Contributing authors: panguyen@uark.edu; pdobbs@uark.edu; khoaluu@uark.edu;

Abstract

Group Activity Recognition (GAR) is a fundamental problem in computer vision, with diverse applications in sports video analysis, video surveillance, and social scene understanding. Unlike conventional action recognition, GAR aims to classify the actions of a group of individuals as a whole, requiring a deep understanding of their interactions and spatiotemporal relationships. To address the challenges in GAR, we present REACT (**R**ecognize **E**very **Act**ion Everywhere All At Once), a novel architecture inspired by the transformer encoder-decoder model explicitly designed to model complex contextual relationships within videos, including multi-modality and spatio-temporal features. Our architecture features a cutting-edge Vision-Language Encoder block for integrated temporal, spatial, and multi-modal interaction modeling. This component efficiently encodes spatiotemporal interactions, even with sparsely sampled frames, and recovers essential local information. Our Action Decoder Block refines the joint understanding of text and video data, allowing us to precisely retrieve bounding boxes, enhancing the link between semantics and visual reality. At the core, our Actor Fusion Block orchestrates a fusion of actor-specific data and textual features, striking a balance between specificity and context. Our method outperforms state-of-the-art GAR approaches in extensive experiments, demonstrating superior accuracy in recognizing and understanding group activities. Our architecture's potential extends to diverse real-world applications, offering empirical evidence of its performance gains. This work significantly advances the field of group activity recognition, providing a robust framework for nuanced scene comprehension.

Keywords: Group Activity Recognition (GAR), Action Retrieval, Vision-Language Modeling

1 Introduction

Group activity recognition (GAR) has emerged as a crucial 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. It requires a deeper understanding of the interactions between multiple actors, including accurate localization of actors and modeling their spatiotemporal relationships [1–8]. As a result, GAR poses fundamental challenges that must be addressed to develop practical solutions for this problem. In this context, developing novel techniques for group activity recognition has become an active area of research in computer vision.

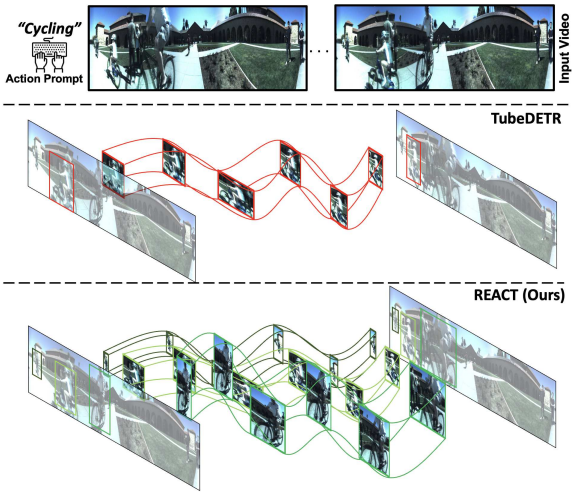


Fig. 1: An example of the response of *REACT model* to the user's *input query*. The user provides a video sequence and an action prompt. Then, the model outputs all the requested actions in the scene by localizing the corresponding actors and provides the overall group activity. **Best viewed in color and zoom in.**

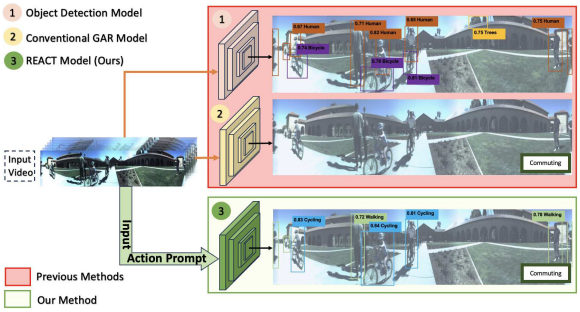


Fig. 2: Comparison between prior methods and our approach. Prior methods do a single classification/detection task while fully supervised, whereas our approach performs group activity classification and query-based action detection simultaneously. **Best viewed in color and zoom in.**

Existing GAR methods require ground-truth bounding boxes and action class labels for training and testing [9–17]. We compare our method with the previous methods in Figure. 2. Bounding box labels 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 [18, 19]. Another approach is weakly supervised (WSGAR) learning [20, 21], which does not require individual actor-level labels for training and inference.

Yan *et al.* [20] proposed the 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.* [21] introduced a detector-free method that captures actor information using partial contexts of token embeddings, but this method can only be learned when there is movement in consecutive frames. Moreover, Kim *et al.* [21] 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.

1.1 Our Contributions in this Work

Inspired by attention mechanisms in video contexts, we introduce a transformer-based architecture, as depicted in Fig. 3, to effectively model complex interactions within videos, incorporating temporal, spatial, and multi-modal features. The **Vision-Language (VL) Encoder block** (Section 3.1) stands as a pivotal element, proficient in encoding sparse spatial and multi-modal interactions, with a dedicated fast branch for recovering temporal details. Our architecture integrates a **Action Decoder Block** (Section 3.2) that refines the understanding of text and video data. It balances context and actor-specific details by merging Encoder and Actor Fusion Block outputs. Utilizing temporal spatial attention for discerning temporal and spatial dependencies and temporal cross attention, aligned with ground-truth bounding boxes, ensures precise retrieval of bounding boxes, bridging semantic and visual contexts. The **Actor Fusion Block** (Section 3.3), positioned at the core, orchestrates a harmonious blend of actor-specific information and textual features. Through concatenation and averaging operations, it achieves a holistic representation enriched by convolution operations extracting local patterns. This refined representation contributes to contextually relevant

output generation in the Decoder Block. In practical terms, our model's efficacy is demonstrated in Fig. 1 through an example response to a user's action query. The results showcase superior performance in accurately recognizing and understanding group activities compared to state-of-the-art methods. Our experiments provide empirical evidence of the significant performance gains achieved by our framework, affirming its potential for diverse real-world applications.

2 Related Work

2.1 Group Activity Recognition (GAR)

In the field of action recognition, group action recognition has become an increasingly popular research topic 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 [22–29]. 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 [9, 18, 30–36].

Recent methods for identifying group actions mainly utilize attention-based models and require explicit character representations to model spatial-temporal relations in group activities [10–14, 17, 20, 37]. 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 extract features for learning actor interactions from multimodal inputs [12]. Additionally, MAC-Loss [38], a combination of spatial and temporal transformers in two complimentary orders, has been proposed to enhance the learning effectiveness of actor interactions and preserve actor consistency at the frame and video levels. Tamura *et al.* [39] 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. These recent advancements in GAR have made significant progress toward recognizing complex actions performed by a group of individuals in various settings.

2.2 Weakly Supervised Group Activity Recognition (WSGAR)

Various techniques have been developed to address the WSGAR problem with limited supervision, such as using bounding boxes to train built-in detectors or activity maps. 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. Weakly supervised methods aim to relax these requirements by learning from more readily available data such as activity labels, bounding boxes, or video-level labels. Zhang et al. [40] proposed a technique that employs activity-specific characteristics to enhance WSGAR. It is not mainly designed for GAR. Kim et al. [21] 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.

2.3 Transformers in Vision

The transformer architecture was first introduced by Vaswani *et al.* [41] for sequence-to-sequence machine translation, and since then, it has been widely applied to various natural language processing tasks. Dosovitskiy *et al.* [42] introduced a transformer architecture not based on convolution for image recognition tasks. Several works [43–46] used transformer architecture as a general backbone for various downstream computer vision tasks, achieving remarkable performance progress. In the video domain, many approaches [47–53] utilize spatial and temporal self-attention to learn video representations effectively. Bertasius *et al.* [50] explored different mechanisms of space and time

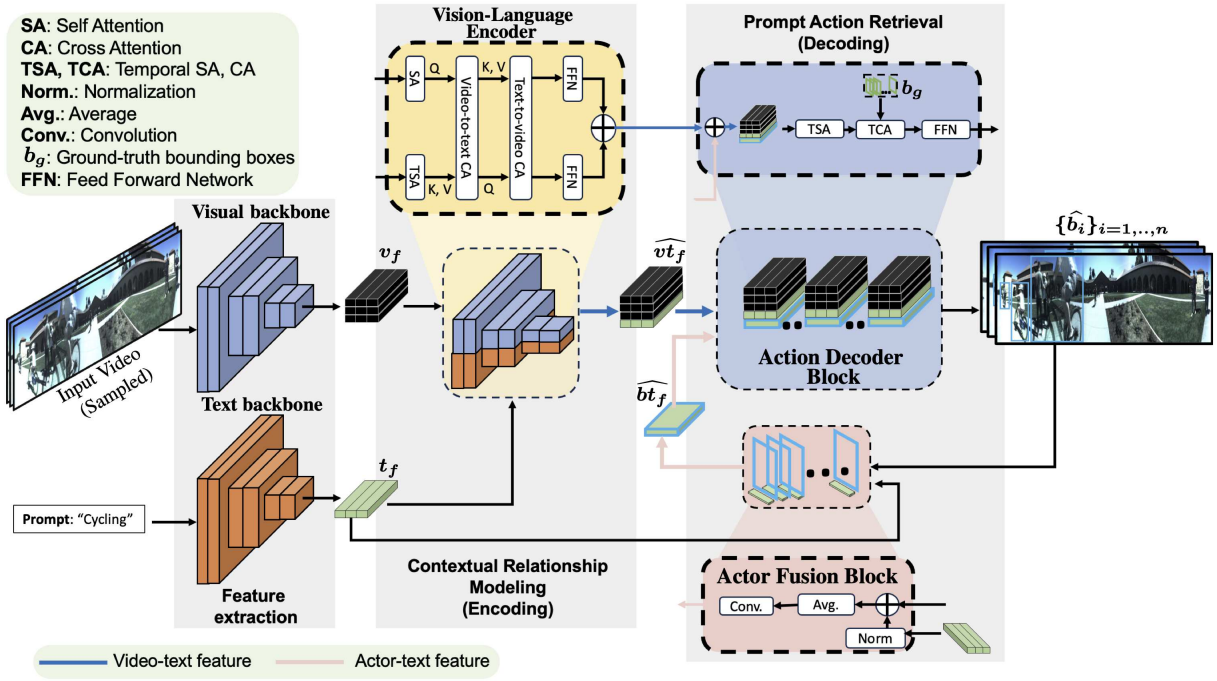


Fig. 3: Overall architecture of the proposed REACT network. The visual and textual representation learning components from our approach incorporate multi-level feature representations. The extracted features are passed through the contextual relationship modeling block to obtain the concatenated multi-modality features. Then, it is passed through the prompt action retrieval block to obtain the detected bounding boxes based on the prompt.

attention to learn spatiotemporal features efficiently. Fan *et al.* [51] used multiscale feature aggregation to improve the learning performance of features. Patrick *et al.* [52] introduced a self-attention block that focuses on the trajectory, which tracks the patches of space and time in a video transformer.

3 The Proposed Method

As shown in Fig. 3, given the input video $\mathbf{X}$ and the textual input $\mathbf{t}$, the REACT model aims to localize all the positions $\mathbf{b}$ in the video $\mathbf{X}$ that consists of the group actions of interest based on the input prompt $\mathbf{t}$. Formally, the REACT model can be formulated as follows:

$$\hat{\mathbf{b}} = AD(\mathbb{F}(VE(\mathbf{X}), TE(\mathbf{t}))) \quad (1)$$

where $\hat{\mathbf{b}} \in [0,1]^{T\times 4}$ is the normalized predicted bounding box coordinates, VE and TE are the visual and textual encoders to extract the feature representations of the video input $\mathbf{X}$, i.e., $\boldsymbol{v_f} = VE(\mathbf{X})$, and the textual prompt $\mathbf{t}$, i.e., $\boldsymbol{t_f} = TE(\mathbf{t})$, and $\mathbb{F}$ is the correlation model to exploit the contextual relationship between visual features $\boldsymbol{v_f}$ and the textual features $\boldsymbol{t_f}$. Then, the Action decoder(AD) will decode the contextual features produced by the correlation model $\mathbb{F}$ to localize all the group action regions of interest. Fig. 3 illustrates the proposed framework of REACT.

Contextual Relationship Modeling in GAR involves modeling the multi-modality and spatio-temporal features. Thus, to efficiently cover these correlations, we model the correlation function $\mathbb{F}$ via the attention mechanism. In particular, the correlation model $\mathbb{F}$ is designed as a Transformer network. Our model $\mathbb{F}$ is at the forefront of our proposed approach which comprises vital components for the integrated modeling of temporal, spatial, and multi-modal interactions. It adeptly encodes spatial and multi-modal interactions, efficiently processing sparsely sampled frames.$\mathbb{F}$ is a

transformer block that mixes text and video representations using cross attention. Formally, $\mathbb{F}$ can be formed as follows:

$$\widehat{\boldsymbol{vt_f}} = \mathbb{F}(\boldsymbol{v_f}, \boldsymbol{t_f}) = FFN(T2V(\boldsymbol{t_f}, \boldsymbol{v_f})((V2T(\boldsymbol{v_f}, \boldsymbol{t_f})))) \tag{2}$$

where $\widehat{\boldsymbol{vt_f}}$ is the shared representation of video and text, FFN is a feed-forward neural network, $T2V$ and $V2T$ are the cross-attention models for the visual and textual features. Transforming textual content into a video format is known as text-to-video. This can involve producing dynamic video content, slideshows, or animations based on the text provided as visual representations of written content. T2V can be used for a number of things, such as text-to-video (T2V) automated video generation, marketing content, and instructional films. The method of obtaining textual information from video material is known as "video-to-text." Said words, text in the video, or other pertinent information from the visual content must all be converted into a text format. V2T is frequently used for activities including content indexing, video transcription, and textual representation of video material to increase accessibility. Thanks to the attention mechanism, it has empowered our correlation model $\mathbb{F}$ to produce a shared representation that captures the contextual relationships between text and video in both spatial and temporal dimensions and encapsulates a joint understanding of both text and video content, enabling the model to bridge the semantic gap between the two modalities effectively.

From the shared contextual feature representations produced by the correlation model $\mathbb{F}$, the action decoder aims to localize all the group action positions of interest $\mathbf{b}$. In practice, the group action positions $\mathbf{b}$ could consist of a list of positions of interest $\mathbf{b} = \{\mathbf{b_1}, \mathbf{b_2}, ..., \mathbf{b_N}\}$ where N is the number of group action positions. For example, in a military mission, the group action of enemies could include the positions of many army groups and adversarial armed devices. Additionally, predicting a particular position of group action requires a broader understanding of the contextual feature representation and the surrounding predicted group action positions. Therefore, to efficiently model the group action decoder, in addition the the shared feature representation, the positions of the current predicted group actions are also taken into account. Formally, the Group Action Decoder can be modeled as follows:

$$\widehat{\mathbf{b_i}} = AD(\widehat{\boldsymbol{vt_f}}, AF(\boldsymbol{t_f}, \{\mathbf{b_1}, \mathbf{b_2}, ..., \mathbf{b_N}\})) = AD(\widehat{\boldsymbol{vt_f}}, \widehat{\boldsymbol{bt_f}}) \tag{3}$$

where AF (ActorFusion) is the model that exploits the correlation between the current predicted positions of group actions and the textual features. Intuitively, the list of group action positions $\{\mathbf{b_1}, .., \mathbf{b_{i-1}}\}$ provide the spatial context of subjects in the video. Meanwhile, the textual features $\boldsymbol{t_f}$ represent linguistic aspects of context and semantics related to the group actions and offer a complementary view to the visual information. The AF model distinguishes between capturing detailed spatial information from group action positions and emphasizing the broader action-related information from the textual features. Finally, the ActionDecoder (AD) model will predict the position of group action $\mathbf{b_i}$ based on the contextual feature representation $\widehat{\boldsymbol{vt_f}}$ and the action fusion feature $\widehat{\boldsymbol{bt_f}}$ produced by the AF model. In essence, the AD model uses a wealth of information from the correlation model $\mathbb{F}$ and fusion features from the AF model to associate textual queries $\boldsymbol{t_f}$ with specific regions or subjects within the video frames. To efficiently exploit the correlation of these features in AD, we design AD via the attention mechanism. It enables the AD model to precisely identify and highlight the regions within the video that correspond to the described actions.

3.1 Vision-Language (VL) Encoder

The VL Encoder typically receives input features from their respective backbone networks in this method. These backbones are specialized models designed to extract meaningful information from set of L text data $\boldsymbol{t_f} \in \mathbb{R}^{L \times d}$ and video data $\boldsymbol{v_f} \in \mathbb{R}^{T \times HW \times d}$ for all T frames and HW spatial resolution, independently optimizing their representations for their specific modalities.

The VL Encoder receives the features $\boldsymbol{t_f}$ and $\boldsymbol{v_f}$. It also uses visual features derived from the video, including frames, motion patterns, bounding boxes, and object presence.

The VL Encoder utilizes attention mechanisms to comprehend these features effectively. It begins with self-attention on text, allowing it to weigh the importance of words and understand the context

within the text. Simultaneously, it performs temporal self-attention on video features to capture temporal dependencies and dynamics.

The model then progresses to cross-attention, connecting the video and text modalities. The "video-to-text" step focuses on video features and aligns them with relevant text. The "text-to-video" step attends to video elements based on the textual context. This bidirectional interaction ensures a holistic understanding of both modalities.

Finally, the information from self-attention and cross-attention is combined, creating multimodal features. These features merge enhanced text and video information, bridging the gap between the two modalities. The VL Encoder facilitates joint text and video analysis, benefiting tasks like video captioning and action recognition. It empowers models to grasp contextual relationships between text and video, enhancing multi-modal understanding.

3.2 Action Decoder (AD) Block

The *AD* Block is a crucial part of the architecture responsible for producing meaningful bounding box outputs. It combines information from the encoder and the actor fusion blocks.

First, it concatenates outputs from these sources. The Encoder Block provides a broad multimodal understanding of text and video, while the Actor Fusion Block adds actor-specific details like normalized text features and actor bounding boxes. This merging process combines contextual understanding with fine-grained actor-specific information.

Then, the *AD* Block employs temporal spatial attention. This mechanism allows the model to consider temporal (sequential) and spatial (object relationships) aspects in video data. It is particularly valuable for tasks like action recognition and actor localization.

Next, it uses temporal cross-attention, incorporating ground-truth bounding box data. Ground-truth bounding boxes precisely locate objects or actors in video frames. This step refines the model's understanding of object and actor movements over time, aligning textual descriptions with the video content.

Finally, the *AD* Block generates bounding boxes as grounding outputs based on text queries. It leverages the information gathered from previous stages, aligning textual descriptions with specific regions or objects in video frames. This output is vital for tasks like actor localization and action recognition, connecting text with visual content to pinpoint regions corresponding to described actions or objects.

3.3 Actor Fusion (AF) Block

The *AF* block begins by combining two vital sources of information, including the bounding box data $\hat{\mathbf{b}}$ and the normalized text features $\mathbf{t_f}$. The bounding box data provides precise location details for objects or actors in video frames, enhancing the model's spatial understanding.

Simultaneously, normalized text features offer textual descriptions and contextual information related to actions or objects in the video, providing a linguistic perspective.

After combining these sources, the *AF* block calculates an average over the concatenated features. This averaging serves multiple purposes. It creates a balanced representation that includes actor-specific details (from the bounding box) and semantic context (from the text features). Additionally, it emphasizes action-related information, aligning with the action recognition task.

The *AF* block uses convolution operations to refine the feature representation further (presented as $\widehat{\mathbf{bt_f}}$). Convolutions are essential in deep learning for capturing local data patterns. They extract and highlight relevant spatial and temporal features from the combined actor-specific information in this context. These convolutional operations are the key to improving the representation before passing it back to the *AD* Block.

3.4 Training Loss

The input training data is in the form of a set of frames, where each set of frames belongs to a video, which is annotated with a text label representing the group activity and individual actions s and the corresponding set of bounding boxes b.

We train our architecture with a linear combination of four losses

$$\mathcal{L} = \lambda_{\mathcal{L}_1}\mathcal{L}_{\mathcal{L}_1}(\hat{b}, b) + \lambda_{gIoU}\mathcal{L}_{gIoU}(\hat{b}, b) + \lambda_{gIoU}\mathcal{L}_{gIoU}(\hat{n}, n) \tag{4}$$

where $b \in [0, 1]$ denotes the normalized ground truth box coordinates and $\hat{b}$ the predicted bounding boxes. Finally, different $\lambda_\bullet$ are scalar weights of the individual losses. $\mathcal{L}_{\mathcal{L}_1}$ is a $\mathcal{L}_1$ loss on bounding box coordinates. $\mathcal{L}_{gIoU}$ is a generalized "intersection over union" (IoU) loss [54] on the bounding boxes. Both $\mathcal{L}_1$ and $\mathcal{L}_{gIoU}$ are used for spatial and temporal grounding. Losses are computed at each layer of the decoder following [55].

3.5 Inference

A machine learning model uses a labelled dataset to learn from during the training phase. It presents input data, makes predictions, compares those forecasts to the actual labels, and iteratively (epoch-by-epoch) adjusts its internal parameters (weights) to minimise prediction error. The model moves onto the inference stage after training. During this stage, fresh, unobserved data is predicted using the trained model. Based on the patterns it discovered during training, the model uses the input data to generate predictions. During the testing phase, we employ distinct operations for individual action and social group predictions. For individual actions, a softmax operation is applied to the predictions from cross-entropy functions, and a sigmoid operation is performed on predictions from binary cross-entropy functions. Our approach follows a hierarchical strategy, initiating label selection from the first partition and progressing to subsequent ones in the hierarchy only if the "Other" class is predicted. In the context of social group prediction, we leverage the predicted counts of individual actions and social groups. To deduce the activity of each predicted social group, we adopt a strategy where the social activity label is determined by the most frequent action labels among its members.

3.6 Discussion

Need for concatenated attention? The method described earlier, which we will call "concatenation-based fusion," involves the combination of features. It performs self-attention on features from individual sources and then computes cross-attention across different sources. It is in contrast to "cross-attention-based fusion." With flexibility and fewer assumptions about the data's spatial pattern, the Transformer model offers significant modeling freedom. Compared to cross-attention-based fusion, concatenation-based fusion is more efficient due to operation sharing and reduces model parameters through weight sharing. Weight sharing is a critical design aspect for symmetric metric learning between two data branches. In concatenation-based fusion, we implement this property in feature extraction and feature fusion. In a nutshell, concatenation-based fusion enhances both efficiency and performance.

What if we use a query-based decoder? Many transformer-based models in vision tasks draw inspiration from the vanilla Transformer decoder. They incorporate a learnable query to extract specific target features from the encoder. For instance, in [55], they use object queries; in [56], they refer to target queries. However, our experimental findings reveal that a query-based decoder faces challenges, including slower convergence and subpar performance. The vanilla Transformer decoder, which is fundamentally a generative model, may not be the ideal choice for classification tasks. Furthermore, employing a single universal target query for all types of objects could potentially create a performance bottleneck. It is worth emphasizing that REACT primarily functions as an "encoder" model within the conventional Transformer encoder-decoder framework.

4 Experiment Results

4.1 Experiment Settings

Upstream task We consistently used the RoBERTa [57] model throughout all the experiments to extract the textual features, whereas the visual backbone varied for different experiments to have fair comparisons. We use hyper-parameters T = 15 (JRDB-PAR) & 8 (Volleyball), N= 6 (JRDB-PAR) & 8 (Volleyball), $\lambda_{\mathcal{L}_1} = 5$, $\lambda_{gIoU} = 2$ and d=256. We initialized temporal attention weights randomly, while spatial attention weights were initialized using a ViT model trained self-supervised over ImageNet-1K [58]. This initialization scheme facilitated faster convergence of space-time ViT, as seen in the supervised setting [59]. We trained using an Adam optimizer [60] with a learning rate of 5×10^{-4}, scaled using a cosine schedule with a

linear warm-up over five epochs [61, 62]. Additionally, we applied weight decay scaled from 0.04 to 0.1 during training.

Downstream tasks We trained a linear classifier on our pretrained 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.

4.2 Dataset Details

Volleyball Dataset [9]: With 55 videos and 4,830 labeled clips (3,493 training, 1,337 testing), this dataset includes annotations for individual actions and group activities with bounding boxes. In WSGAR experiments, we focus solely on group activity labels, excluding individual action annotations. Evaluation metrics include Multi-class Classification Accuracy (MCA) and Merged MCA, ensuring fair comparisons with existing methods.

JRDB-PAR Dataset [63]: Featuring 27 action, 11 social groups, and seven global activity categories, this dataset comprises 27 videos (20 training, seven testing) with 27,920 frames and 628k human bounding boxes. Uniformly sampled keyframes (every 15 frames) are used for annotation and evaluation. Precision ($\mathcal{P}_g$), recall ($\mathcal{R}_g$), and F1-score ($\mathcal{F}_g$) are the evaluation metrics for social group activity recognition, treated as a multi-label classification problem [63].

4.3 Classification Task Evaluation

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. Our method significantly outperforms the existing social group activity recognition frameworks in the fully supervised setting in all metrics. In the weakly supervised setting, our proposed method outperformed existing GAR and WSGAR methods by a considerable margin, achieving *1.2* of $\mathcal{P}_g$, *1.7* of $\mathcal{R}_g$ and *2.3* of $\mathcal{F}_g$. Additionally, we evaluated this dataset using ResNet-18, ViT-B/16, and ViT-B/32 backbones, where ViT-B/16 proved to be better (whose results are presented in Table. 2), which is further analyzed in Appendix 3. Despite their impressive performance in WSGAR, our approach outperformed them all.

Table 1: Comparison with the state-of-the-art methods on the Volleyball dataset [9]. MCA and MPCA represent Mean Class Accuracy and Mean Per Class Accuracy. The *green* line indicates the best result.

Method	Backbone	MCA	Merged MCA
Fully supervised			
SSU [18]	Inception-v3	89.9	-
PCTDM [36]	ResNet-18	90.3	94.3
StagNet [33]	VGG-16	89.3	-
ARG [10]	ResNet-18	91.1	95.1
CRM [64]	I3D	92.1	-
HiGCIN [15]	ResNet-18	91.4	-
AT [12]	ResNet-18	90.0	94.0
SACRF [13]	ResNet-18	90.7	92.7
DIN [37]	ResNet-18	93.1	95.6
TCE+STBiP [16]	VGG-16	94.1	-
GroupFormer [17]	Inception-v3	94.1	-
Weakly supervised			
PCTDM [36]	ResNet-18	80.5	90.0
ARG [10]	ResNet-18	87.4	92.9
AT [12]	ResNet-18	84.3	89.6
SACRF [13]	ResNet-18	83.3	86.1
DIN [37]	ResNet-18	86.5	93.1
SAM [20]	ResNet-18	86.3	93.1
DFWSGAR [21]	ResNet-18	90.5	94.4
SoGAR [65]	ResNet-18	91.8	94.5
SPARTAN [66]	ViT-B/16	92.9	95.6
SoGAR [65]	ViT-B/16	93.1	95.9
Ours	ResNet-18	92.3	94.6
	ViT-B/32	**93.4**	**96.3**
	ViT-B/16	**94.2**	**96.7**

Volleyball dataset. We evaluate our approach against the latest GAR and WSGAR methods in two supervision levels: fully supervised and weakly supervised. These differ in using actor-level labels such as ground-truth bounding boxes and individual action class labels in training and inference. To ensure a fair comparison, we report the results of previous methods and reproduce results using only the RGB input and ResNet-18 backbone, respectively. In the weakly supervised setting, we replace the group action classification labels with ground-truth bounding boxes of the actors without their corresponding actions so that the actors localization is learned during the pre-training stage. Table. 1 presents the results, with the first and second sections showing the results of earlier techniques in fully supervised and weakly supervised

Table 2: Comparative results of the group activity recognition on JRDB-PAR dataset [63]. $\mathcal{P}_g$, $\mathcal{R}_g$, and $\mathcal{F}_g$ represent Precision, Recall and F1-score. In these experiments, we use ViT-B/16 as the backbone.

Method	Classification		
	$\mathcal{P}_g$	$\mathcal{R}_g$	$\mathcal{F}_g$
Fully supervised			
PCTDM [36]	32.1	27.2	28.9
ARG [10]	34.6	29.3	30.7
HiGCIN [15]	35.1	34.4	31.8
AT[12]	34.8	37.6	30.9
SA-GAT [14]	36.7	29.9	31.4
JRDB-Base [67]	44.6	46.8	45.1
SoGAR [65]	49.3	47.1	48.7
Ours	**50.1**	**48.9**	**49.2**
Weakly supervised			
AT[12]	21.2	19.1	19.8
SACRF[13]	42.9	35.5	37.6
Dynamic[37]	37.5	27.1	30.6
HiGCIN[15]	39.3	30.1	33.1
ARG[10]	26.9	21.5	23.3
SA-GAT[14]	28.6	24.0	25.5
JRDB-Base[67]	38.4	33.1	34.8
SoGAR [65]	47.1	45.8	44.9
Ours	**48.3**	**47.5**	**47.2**

environments, respectively. Our model trained on the ResNet-18 backbone outperforms most fully supervised frameworks by significantly improving the MCA and MPCA metrics. We show that using ViT-Base backbone, our approach significantly outperforms all GAR and WSGAR models in weakly supervised conditions, beating them by 2.4% of MCA and 1.2% of Merged MCA by leveraging the spatiotemporal features using the transformer architecture. Moreover, our approach is better than the current GAR methods that employ less thorough actor-level supervision, such as [12, 13, 18, 33, 36].

4.4 Human Action Retrieval Evaluation

Our experimental results in Table. 3 show that our proposed method outperforms the existing state-of-the-art approaches in terms of the R@K metric. Specifically, our method achieved a significantly higher R@1 score than the other methods, with a margin of *4.8*. Additionally, our method achieved higher R@5 and R@10 scores than all other methods, demonstrating its superior performance in video object retrieval. Furthermore, our method demonstrated a more consistent performance across two datasets and experimental setups. We attribute this to our method's ability to effectively capture the video frames' global and local features and its incorporation of text embeddings for improved semantic understanding of the video content.

4.5 Ablation Study

In this section, we verify the effectiveness of each component in the REACT framework on downstream group activity recognition tasks and verify the significance of the actor fusion block on the proposed method. In Table. 4, irrespective of the visual backbone, results indicate that all the components, including feature extraction, encoding, and decoding, can bring significant gains individually. From Table. 5, results show the importance of actor fusion block in all the experiments. More ablation results can be seen in Appendix, including the detailed ablation on the effectiveness of each component on downstream human action retrieval tasks.

4.6 Visualization

In order to gain insights into the visualization capabilities of our proposed method and demonstrate its significance, we conducted an analysis that involved examining the visualizations based on the input action query. The visualizations are presented in Fig. 4, which showcases the localization of actions based on the input query for the JRDB-PAR and Volleyball datasets. The results of the decoder's generated proposals during the inference stage indicate that our method effectively localizes the specified actions. It demonstrates the ability of our framework to accurately recognize and pinpoint specific actions within a group activity context. Furthermore, the t-SNE plots are visualized in the Appendix. The results illustrated in these plots demonstrate that the proposed framework effectively recognizes all actions within a given video. Moreover, the framework enhances the overall performance in group activity recognition.

5 Conclusion

We introduce REACT, a novel video transformer-based model for group activity recognition using

Table 3: Human Action-retrieval task results on transformer-based methods.

Method	Backbone	Volleyball			JRDB-PAR		
		R@1	R@5	R@10	R@1	R@5	R@10
AT[12]	ResNet18	61.2	79.2	95.2	46.3	54.4	74.4
DFWSGAR [21]		64.7	80.4	85.3	47.6	57.9	77.1
SPARTAN [66]		68.6	80.6	95.7	54.6	60.1	80.2
Ours		**71.3**	**84.5**	**96.3**	**61.4**	**71.3**	**87.4**
AT[12]	ViT-B/32	54.2	77.6	93.6	41.5	49.2	68.7
DFWSGAR [21]		57.6	59.8	84.4	44.2	53.6	70.8
SPARTAN [66]		58.1	81.3	95.4	49.5	59.2	74.3
Ours		**64.4**	**84.2**	**95.4**	**55.6**	**67.2**	**81.3**
AT[12]	ViT-B/16	62.3	84.6	97.3	50.2	58.1	77.2
DFWSGAR [21]		65.7	59.8	84.4	54.3	62.7	79.8
SPARTAN [66]		69.4	84.5	97.4	59.1	65.3	82.1
Ours		**74.2**	**85.6**	**97.7**	**67.2**	**71.4**	**88.2**

Table 4: Ablation study of each Group Activity Recognition task component. **VB** denotes Visual Backbone and **FE** denotes Feature Extraction.

VB	*FE*	*Encoding*	Decoding	JRDB-PAR	Volleyball
ResNet-18	✓			36.5	82.4
	✓	✓		41.3(+4.8)	86.5(+4.1)
	✓	✓	✓	**46.4**(+5.1)	**92.3**(+5.8)
ViT-B/32	✓			31.8	80.2
	✓	✓		39.2(+7.4)	84.3(+4.1)
	✓	✓	✓	**45.7**(+6.5)	**93.4**(+9.1)
ViT-B/16	✓			38.3	79.7
	✓	✓		41.4(+3.1)	85.7(+6.0)
	✓	✓	✓	**47.2**(+6.2)	**96.7**(+11.0)

Table 5: Ablation study results on the Actor Fusion component in the architecture. For all the experiments below, we use ViT-B/32 as the visual backbone for the framework. The metrics we use for the Volleyball dataset are *Merged MCA*, and for the JRDB-PAR dataset $\mathcal{F}_g$.

Actor Fusion	JRDB-PAR	Volleyball
✗	33.7	76.1
✓	**47.2**	**96.7**

contrastive learning. Our approach generates diverse spatio-temporal views from a single video, leveraging different scales and frame rates. Correspondence learning tasks capture motion properties and cross-view relationships between sampled clips and textual information. The core of our approach is a contrastive learning objective that reconstructs video-text modalities in the latent space. The input-output pair in the V2T architecture is reversed from that of the T2V network topology. Using the fundamental cGAN architecture, our network is made up of a Discriminator D and a Generator G. REACT effectively models long-range spatio-temporal dependencies and performs dynamic inference within a single architecture. REACT outperforms state-of-the-art models in group activity recognition, showcasing its superior performance.

Limitations. While our methods excel in group activity classification and human action-retrieval tasks, they may need to be optimized for other visual-language problems due to the lack of textual descriptions in the datasets. Future research directions could focus on incorporating additional data



Fig. 4: Visualization based on input action query. The top two rows are the results from the JRDB-PAR dataset, and the bottom row is from the Volleyball dataset. **Best viewed in color and zoom in.**

involving question-answering, phrases for video grounding, and other related textual information to enhance the versatility of our methods.

References

[1] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV, pp. 20–36 (2016). Springer

[2] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR, pp. 6299–6308 (2017)

[3] Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR, pp. 7794–7803 (2018)

[4] Ranasinghe, K., Naseer, M., Khan, S., Khan, F.S., Ryoo, M.S.: Self-supervised video transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2874–2884 (2022)

[5] Nguyen, T.-T., Nguyen, P., Luu, K.: Hig: Hierarchical interlacement graph approach to

scene graph generation in video understanding. arXiv preprint arXiv:2312.03050 (2023)

[6] Nguyen, P., Quach, K.G., Duong, C.N., Phung, S.L., Le, N., Luu, K.: Multi-camera multi-object tracking on the move via single-stage global association approach. arXiv preprint arXiv:2211.09663 (2022)

[7] Nguyen, P., Quach, K.G., Kitani, K., Luu, K.: Type-to-track: Retrieve any object via prompt-based tracking. Advances in Neural Information Processing Systems **36** (2024)

[8] Quach, K.G., Le, N., Duong, C.N., Jalata, I., Roy, K., Luu, K.: Non-volume preserving-based fusion to group-level emotion recognition on crowd videos. Pattern Recognition **128**, 108646 (2022)

[9] Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G.: A hierarchical deep temporal model for group activity recognition. In: CVPR, pp. 1971–1980 (2016)

[10] Wu, J., Wang, L., Wang, L., Guo, J., Wu, G.: Learning actor relation graphs for group activity recognition. In: CVPR, pp. 9964–9974 (2019)

[11] Hu, G., Cui, B., He, Y., Yu, S.: Progressive relation learning for group activity recognition. In: CVPR, pp. 980–989 (2020)

[12] Gavrilyuk, K., Sanford, R., Javan, M., Snoek, C.G.: Actor-transformers for group activity recognition. In: CVPR, pp. 839–848 (2020)

[13] Pramono, R.R.A., Chen, Y.T., Fang, W.H.: Empowering relational network by self-attention augmented conditional random fields for group activity recognition. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, pp. 71–90 (2020). Springer

[14] Ehsanpour, M., Abedin, A., Saleh, F., Shi, J., Reid, I., Rezatofighi, H.: Joint learning of social groups, individuals action and sub-group activities in videos. In: ECCV, pp. 177–195 (2020). Springer

[15] Yan, R., Xie, L., Tang, J., Shu, X., Tian, Q.: Higcin: hierarchical graph-based cross inference network for group activity recognition (2020)

[16] Yuan, H., Ni, D.: Learning visual context for group activity recognition. In: AAAI, vol. 35, pp. 3261–3269 (2021)

[17] Li, S., Cao, Q., Liu, L., Yang, K., Liu, S., Hou, J., Yi, S.: Groupformer: Group activity recognition with clustered spatial-temporal transformer. ICCV (2021)

[18] Bagautdinov, T., Alahi, A., Fleuret, F., Fua, P., Savarese, S.: Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In: CVPR, pp. 4315–4324 (2017)

[19] Zhang, P., Tang, Y., Hu, J.-F., Zheng, W.-S.: Fast collective activity recognition under weak supervision. IEEE Transactions on Image Processing **29**, 29–43 (2019)

[20] Yan, R., Xie, L., Tang, J., Shu, X., Tian, Q.: Social adaptive module for weakly-supervised group activity recognition. In: ECCV, pp. 208–224 (2020). Springer

[21] Kim, D., Lee, J., Cho, M., Kwak, S.: Detector-free weakly supervised group activity recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20083–20093 (2022)

[22] Amer, M.R., Xie, D., Zhao, M., Todorovic, S., Zhu, S.-C.: Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In: ECCV, pp. 187–200 (2012). Springer

[23] Amer, M.R., Todorovic, S., Fern, A., Zhu, S.-C.: Monte carlo tree search for scheduling activity recognition. In: ICCV, pp. 1353–1360 (2013)

[24] Amer, M.R., Lei, P., Todorovic, S.: Hirf: Hierarchical random field for collective activity recognition in videos. In: ECCV, pp. 572–585 (2014). Springer

[25] Amer, M.R., Todorovic, S.: Sum product networks for activity recognition **38**(4), 800–813 (2015)

[26] Lan, T., Wang, Y., Yang, W., Robinovitch, S.N., Mori, G.: Discriminative latent models for recognizing contextual group activities **34**(8), 1549–1562 (2011)

[27] Lan, T., Sigal, L., Mori, G.: Social roles in hierarchical models for human activity recognition. In: CVPR, pp. 1354–1361 (2012). IEEE

[28] Shu, T., Xie, D., Rothrock, B., Todorovic, S., Chun Zhu, S.: Joint inference of groups, events and human roles in aerial videos. In: CVPR, pp. 4576–4584 (2015)

[29] Wang, Z., Shi, Q., Shen, C., Van Den Hengel, A.: Bilinear programming for human activity recognition with unknown mrf graphs. In: CVPR, pp. 1690–1697 (2013)

[30] Deng, Z., Vahdat, A., Hu, H., Mori, G.: Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In: CVPR, pp. 4772–4781 (2016)

[31] Ibrahim, M.S., Mori, G.: Hierarchical relational networks for group activity recognition and retrieval. In: ECCV, pp. 721–736 (2018)

[32] Li, X., Choo Chuah, M.: Sbgar: Semantics based group activity recognition. In: ICCV, pp. 2876–2885 (2017)

[33] Qi, M., Qin, J., Li, A., Wang, Y., Luo, J., Van Gool, L.: stagnet: An attentive semantic rnn for group activity recognition. In: ECCV, pp. 101–117 (2018)

[34] Shu, X., Tang, J., Qi, G., Liu, W., Yang, J.: Hierarchical long short-term concurrent memory for human interaction recognition (2019)

[35] Wang, M., Ni, B., Yang, X.: Recurrent modeling of interaction context for collective activity recognition. In: CVPR, pp. 3048–3056 (2017)

[36] Yan, R., Tang, J., Shu, X., Li, Z., Tian, Q.: Participation-contributed temporal dynamic model for group activity recognition. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 1292–1300 (2018)

[37] Yuan, H., Ni, D., Wang, M.: Spatio-temporal dynamic inference network for group activity recognition. In: ICCV (2021)

[38] Han, M., Zhang, D.J., Wang, Y., Yan, R., Yao, L., Chang, X., Qiao, Y.: Dual-ai: Dual-path actor interaction learning for group activity recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2990–2999 (2022)

[39] Tamura, M., Vishwakarma, R., Vennelakanti, R.: Hunting group clues with transformers for social group activity recognition. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV, pp. 19–35 (2022). Springer

[40] Zhang, Y., Li, X., Marsic, I.: Multi-label activity recognition using activity-specific features and activity correlations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14625–14635 (2021)

[41] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NIPS, pp. 5998–6008 (2017)

[42] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[43] Li, M., Cai, W., Liu, R., Weng, Y., Zhao, X., Wang, C., Chen, X., Liu, Z., Pan, C., Li, M., *et al.*: Ffa-ir: Towards an explainable and reliable medical report generation benchmark. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)

[44] Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. arXiv preprint arXiv:2101.11986 (2021)

[45] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)

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

[47] Han, M., Wang, Y., Chang, X., Qiao, Y.: Mining inter-video proposal relations for video object detection. In: European Conference on Computer Vision, pp. 431–446 (2020). Springer

[48] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., Schmid, C.: Vivit: A video vision transformer. arXiv preprint arXiv:2103.15691 (2021)

[49] Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y.: UniFormer: Unifying Convolution and Self-attention for Visual Recognition (2022)

[50] Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML, vol. 2, p. 4 (2021)

[51] Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. arXiv preprint arXiv:2104.11227 (2021)

[52] Patrick, M., Campbell, D., Asano, Y.M., Metze, I.M.F., Feichtenhofer, C., Vedaldi, A., Henriques, J., *et al.*: Keeping your eye on the ball: Trajectory attention in video transformers. In: NeurIPS (2021)

[53] Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Tubedetr: Spatio-temporal video grounding with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16442–16453 (2022)

[54] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 658–666 (2019)

[55] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV, pp. 213–229 (2020). Springer

[56] Yan, B., Peng, H., Fu, J., Wang, D., Lu, H.: Learning spatio-temporal transformer for visual tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10448–10457 (2021)

[57] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[58] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. IJCV (2015)

[59] Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML, vol. 2, p. 4 (2021)

[60] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)

[61] Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., Beyer, L.: How to train your vit? data, augmentation, and regularization in vision transformers (2021)

[62] Chen*, X., Xie*, S., He, K.: An empirical study of training self-supervised vision transformers (2021)

[63] Han, R., Yan, H., Li, J., Wang, S., Feng, W.,

Wang, S.: Panoramic human activity recognition. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV, pp. 244–261 (2022). Springer

[64] Azar, S.M., Atigh, M.G., Nickabadi, A., Alahi, A.: Convolutional relational machine for group activity recognition. In: CVPR, pp. 7892–7901 (2019)

[65] Chappa, N.V., Nguyen, P., Nelson, A.H., Seo, H.-S., Li, X., Dobbs, P.D., Luu, K.: Sogar: Self-supervised spatiotemporal attention-based social group activity recognition. arXiv preprint arXiv:2305.06310 (2023)

[66] Chappa, N.V., Nguyen, P., Nelson, A.H., Seo, H.-S., Li, X., Dobbs, P.D., Luu, K.: Spartan: Self-supervised spatiotemporal transformers approach to group activity recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5157–5167 (2023)

[67] Ehsanpour, M., Saleh, F., Savarese, S., Reid, I., Rezatofighi, H.: Jrdb-act: A large-scale dataset for spatio-temporal action, social group and activity detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20983–20992 (2022)