Dual-AI: Dual-path Actor Interaction Learning for Group Activity Recognition

Anonymous CVPR submission

Paper ID 5612

Abstract

Learning spatial-temporal relation among multiple actors is crucial for group activity recognition. Different group activities often show the diversified interactions between actors in the video. Hence, it is often difficult to model complex group activities from a single view of spatial-temporal actor evolution. To tackle this problem, we propose a distinct Dual-path Actor Interaction (Dual-AI) framework, which flexibly arranges spatial and temporal transformers in two complementary orders, enhancing actor relations by integrating merits from different spatio-temporal paths. Moreover, we introduce a novel Multi-scale Actor Contrastive Loss (MAC-Loss) between two interactive paths of Dual-AI. Via self-supervised actor consistency in both frame and video levels, MAC-Loss can effectively distinguish individual actor representations to reduce action confusion among different actors. Consequently, our Dual-AI can boost group activity recognition by fusing such discriminative features of different actors. To evaluate the proposed approach, we conduct extensive experiments on the widely used benchmarks, including Volleyball [21], Collective Activity [12], and NBA datasets [47]. The proposed Dual-AI achieves state-of-the-art performance on all these datasets. It is worth noting the proposed Dual-AI with 50% training data outperforms a number of recent approaches with 100% training data. This confirms the generalization power of Dual-AI for group activity recognition, even under the challenging scenarios of limited supervision.

1. Introduction

Group Activity Recognition (GAR) is an important problem in video understanding. In this task, we should not only recognize individual action of each actor but also understand collective activity of multiple involved actors. Hence, it is vital to learn spatio-temporal actor relations for GAR [44, 47, 49].

Several attempts have been proposed to model actor relations by building visual attention among actors [6, 17, 19, 24, 44, 47, 49]. However, it is often difficult for joint spatial-temporal optimization [8, 35]. For this reason, the recent approaches in group activity recognition often decompose spatial-temporal attention separately for modeling actor interaction [17, 24, 47]. But single order of space and time is insufficient to describe complex group activities, due to the fact that different group activities often exhibit diversified spatio-temporal interactions.

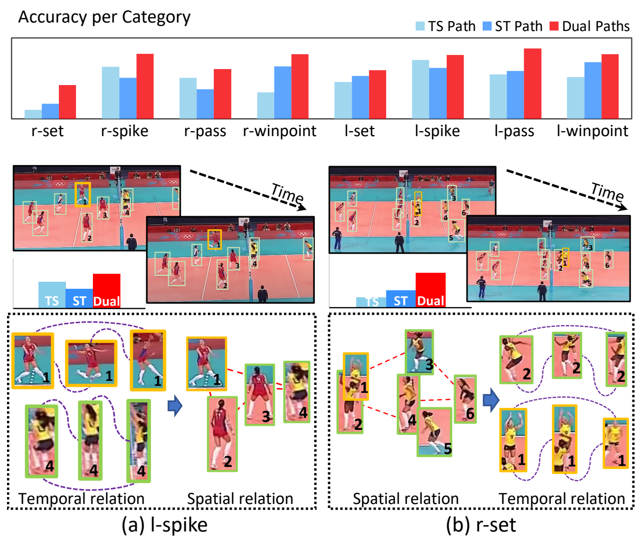

Figure 1. Accuracy per Category and Example of *left spike* and *right set* group activity. Red dashed line and Violet dashed line below show spatial and temporal actor interaction respectively. With spatial and temporal modeling applied in different orders, ST path and TS path learn different spatiotemporal patterns and thereby are skilled at different classes, supported by the accuracy plot.

For example, Fig. 1 (a) refers to the *l-spike* activity in the volleyball, where the hitting player (actor 1) and the defending player (actor 4) move fast to hit and block the ball, while other accompanying players (*e.g.*, actor 2 and actor 3) stand without much movement. Hence, for this group activity, it is better to first understand temporal dynamics of each actor, and then reason spatial interaction among actors in the scene. On the contrary, Fig. 1 (b) refers to the *r-set* activity in the volleyball, where most players in the right-side team are moving cooperatively to tackle the ball falling on different positions, *e.g.*, actor 1 jumps and sets the ball, while actor 2 jumps together to make a fake spiking action. Hence, for this group activity, it is better to reason spatial actor interaction first to understand the action scene, and then model temporal evolutions of each actor. In fact, as shown in the accuracy plot of Fig. 1, the order of space and time interaction varies for different activity categories.

Based on these observations, we propose a distinct Dual-path Actor Interaction (Dual-AI) framework for GAR, which can effectively integrate two complementary spatiotemporal views to learn complex actor relations in videos. Specifically, Dual-AI consists of Spatial-Temporal (ST) and Temporal-Spatial (TS) Interaction Paths, with assistance of spatial and temporal transformers. ST path first takes spatial transformer to capture spatial relation among actors in each frame, and then utilizes temporal transformer to model temporal evolution of each actor over frames. Alternatively, TS path arranges spatial and temporal transformers in a reverse order to describe complementary pattern of actor interaction. In this case, our Dual-AI can comprehensively leverage both paths to generate robust spatiotemporal contexts for boosting GAR.

Furthermore, we introduce a novel Multi-scale Actor Contrastive Loss (MAC-Loss), which is a concise but effective self-supervised signal to enhance actor consistency between two paths. Via such actor supervision in all the frame-frame, frame-video, video-video levels, we can further reduce action confusion between any two individual actors to improve the discriminative power of actor representations in GAR.

Finally, we conduct extensive experiments on the widely-used benchmarks to evaluate our designs. Our Dual-AI simply achieves state-of-the-art performance on all the fully-annotated datasets, such as Volleyball, Collective Activity. More interestingly, our Dual-AI with 50% training data is competitive to a number of recent approaches with 100% training data in Volleyball as shown in Fig. 2, which clearly demonstrates the generalization power of our Dual-AI. Motivated by this, we further investigate the challenging setting with limited actor supervision [47], where Dual-AI also achieves state-of-the-art results on Weak-Volleyball-M and NBA datasets. All these results show the effectiveness of our Dual AI for learning spatiotemporal actor relations in GAR.

2. Related Work

Group activity recognition has attracted a large body of work recently due to its wide applications. Early approaches are based on hand-crafted features and typically use probabilistic graphical models [1–3, 22, 23, 43] and AND-OR grammar methods [4, 31]. Recently, methods incorporating convolutional neural networks [7, 21] and recurrent neural networks [7, 13, 20, 21, 25, 29, 32, 39, 45] have achieve remarkable performance, due to the learning

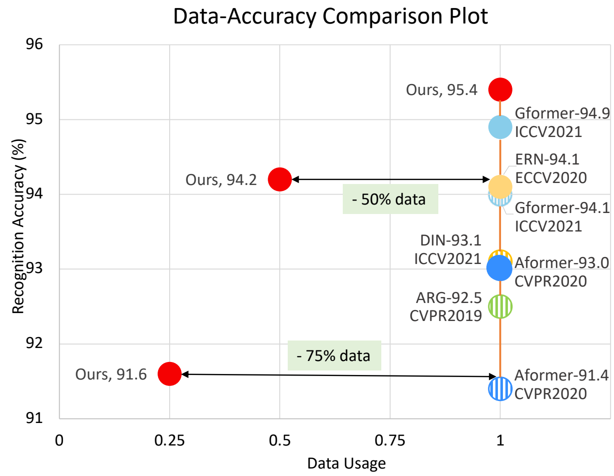

Figure 2. **Accuracy comparison with data in different percentage on Volleyball dataset**. Our method achieves SOTA performance, and achieves 94.2% with 50% data, which is competitive to a number of recent approaches [17, 28, 44] trained with 100% data. Solid point means result with additional optical flow input.

of temporal context and high-level information.

More recent group activity recognition methods [15, 17, 19, 24, 28, 44, 47, 49] often require the explicit representation of spatiotemporal relations, dedicated to apply attention-based methods to model the individual relations for inferring group activity. [44, 49] build relational graphs of the actors and explore the spatial and temporal actor interactions in the same time with graph convolution networks. These methods simulate spatiotemporal interaction of actors in a joint manner. Differently, [47] builds separate spatial and temporal relation graphs subsequently to model the actor relations. [17] encodes temporal information with I3D [10] and constructs spatial relation of the actors with a vanilla transformer. [24] introduces a cluster attention mechanism for better group informative features with transformers. Different from previous approaches, we propose to learn the actor interactions in complementary Spatial-Temporal and Temporal-Spatial views and further promote actor interaction learning with a designed self-supervised loss for effective representation learning.

Vision Transformer has gradually become popular for computer vision tasks. In image domain, ViT [14] firstly introduces a pure transformer architecture without convolution for image recognition. Following works [26, 41, 50] make remarkable progress on enabling transformer architecture to become a general backbone on various kinds of downstream computer vision tasks. In video domain, inspired by ViT, many works [5, 8, 16, 27] explore spatial and temporal self-attention to learn efficient video representation. TimeSformer [8] investigates the different space and time attention mechanisms to learn spatial-temporal representation efficiently. MViT [16] utilizes the multi-

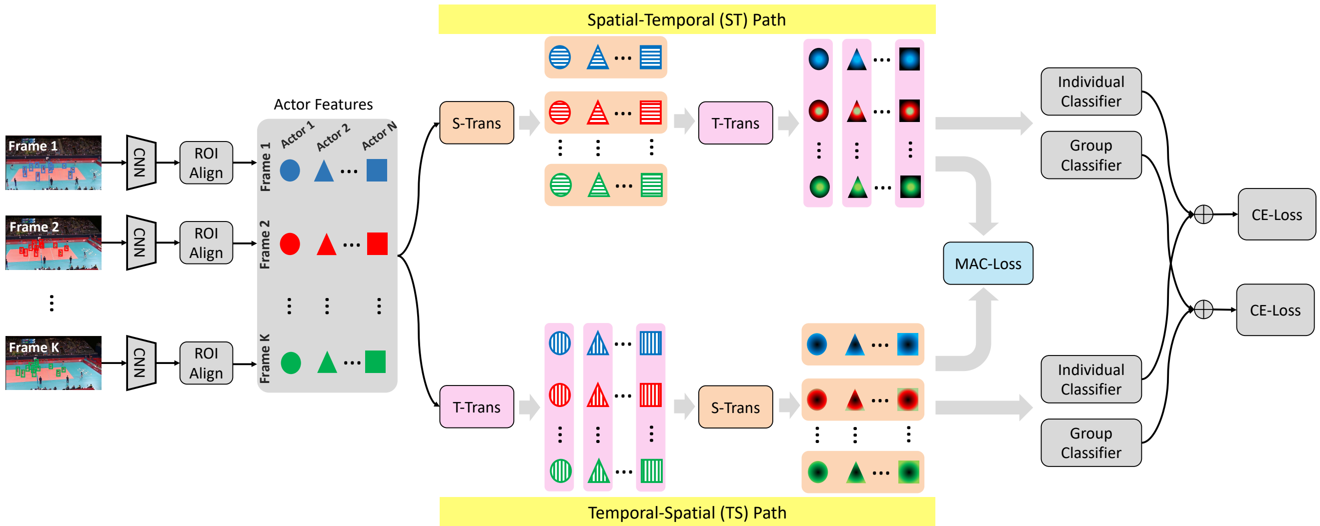

Figure 3. Our Dual-path Actor Interaction (Dual-AI) learning framework, where S-Trans and T-Trans denote Spatial-Transformer and Temporal-transformer respectively. It effectively explores actor evolution in two complementary spatiotemporal views, *i.e.*, ST path and TS path, detailed in Sec. 3.2. Moreover, a Multi-scale Actor Contrastive loss is designed to enable interaction and cooperation of the two paths as in Sec. 3.3.

scale features aggregation to enhance the spatial-temporal representation. Motionformer [27] presents a trajectory-focused self-attention block, which essentially tracks space-time patches for video transformer. The above transformer architectures are designed for general video classification task. It has not been fully explored to tackle the challenging GAR problem with transformers. We propose to construct dual spatiotemporal paths with transformers to flexibly learn actor interactions for group activity recognition.

3. Method

To learn complex actor relations in the group activities, we propose a distinct Dual-path Actor Interaction (Dual-AI) framework for GAR. In this section, we introduce our Dual-AI in detail. First, we describe an overview of Dual-AI framework. Then, we explain how to build the interaction paths, with assistance of spatial and temporal transformers. Next, we introduce a Multi-scale Actor Contrastive Loss (MAC-Loss) to further improve actor consistency between paths. Finally, we describe the training objectives to optimize our Dual-AI framework.

3.1. Framework Overview

As shown in Fig. 3, our Dual-AI framework consists of three important steps. First, we need to extract actor features from backbone. Specifically, we sample K frames from the input video. To make a fair comparison with the previous works in GAR [7, 24, 44, 48, 49], we choose ImageNet-pretrained Inception-v3 [33] as backbone to extract feature of each sampled frame. Then, we apply RoIAlign [18] on the frame feature, which can generate actor features in this frame from bounding boxes of N actors. After that, we adopt a fully-connected layer to further encode each actor feature into a C dimensional vector. For convenience, we denote all the actor vectors as $\mathbf{X} \in \mathbb{R}^{K \times N \times C}$. More details can be found in Sec. 4.2.

After extracting actor feature vectors, we next learn spatiotemporal interactions among these actors in the video. Different from the previous approaches [17, 44, 46, 47, 49], we disentangle spatiotemporal modeling into consecutive spatial and temporal interactions in different orders. Specifically, we design spatial and temporal transformers as basic actor relation modules. By flexibly arranging these transformers in two reverse orders, we can enhance actor relations with complementary integration of both spatial-temporal (ST) and temporal-spatial (TS) interaction paths. Finally, we design training losses to optimize our Dual-AI framework. In particular, we introduce a novel Multi-scale Actor Contrastive Loss (MAC-Loss) between two paths, which can effectively improve discriminative power of individual actor representations, by actor consistency in all the frame-frame, frame-video, video-video levels. Subsequently, we integrate actor representations of two paths to recognize individual actions and group activities.

3.2. Dual-path Actor Interaction

To capture complex relations for diversified group activities, we propose a novel dual path structure to describe actor interactions. To start with, we build basic spatial and temporal actor relation units, with assistance of transformers. Then, we explain how to construct dual paths for spatiotemporal actor interactions.

3.2.1 Spatial/Temporal Actor Relation Units

To understand spatiotemporal actor evolution in videos, we first construct basic units to describe spatial and temporal actor relations. Since there is no prior knowledge about actor relation, we propose to use transformer to model such relation by the powerful self-attention mechanism.

Spatial Actor Transformer. In order to model the spatial relation of the actors in single frame, we design a concise spatial actor transformer (S–Trans). Specifically, we denote $\mathbf{X}^k \in \mathbb{R}^{N\times C}$ as the feature vectors of N actors in the k-th frame. The spatial relation among these actors are modeled by $\hat{\mathbf{X}}^k = \text{S–Trans}(\mathbf{X}^k)$, which consists of three modules as follows,

$$\mathbf{X}' = \text{SPE}(\mathbf{X}^k) + \mathbf{X}^k, \tag{1}$$
$$\mathbf{X}'' = \text{LN}(\mathbf{X}' + \text{MHSA}(\mathbf{X}')), \tag{2}$$
$$\hat{\mathbf{X}}^k = \text{LN}\big(\mathbf{X}'' + \text{FFN}(\mathbf{X}'')\big). \tag{3}$$

First, we use spatial position encoding (SPE) to add spatial structure information of the actors in the scene, as in Eq. (1). We represent spatial position of each actor with center point of its bounding box and encode the spatial positions with PE function in [9, 17]. Second, we use multi-head self-attention (MHSA) [37] module to reason the spatial interaction of the actors in the scene, as in Eq. (2). Finally, we use feed-forward network (FFN) [37] to further improve learning capacity of the spatial actor relation unit, as in Eq. (3).

Temporal Actor Transformer. In order to model the temporal evolution of single actor across frames, we design a temporal actor transformer (T–Trans) following the way in Eqs. (1) to (3). Differently, we use the input as the feature vectors of the n-th actor across K frames, *i.e.*, $\mathbf{X}^n \in \mathbb{R}^{K\times C}$. In this case, the MHSA module can reason the evolution of actor n in different time steps. Moreover, to add temporal sequence information of actor n, temporal position encoding (TPE) is used instead of SPE, which encodes frame index $\{1, ..., K\}$ with PE function in [37]. Finally, we can get actor features enhanced by temporal interactions, as $\hat{\mathbf{X}}^n = \text{T–Trans}(\mathbf{X}^n)$.

3.2.2 Dual Spatiotemporal Paths of Actor Interaction

Once the spatial and temporal relations of actors are built, we can further integrate them to construct spatiotemporal representation of the actor evolution. As discussed in Sec. 1, the single order of space and time is insufficient to understand the complex actor interactions, leading to the failure of inferring group activities. Thus, we propose a dual spatiotemporal paths framework for GAR to capture the complex interaction of the actors.

It consists of two complementary spatiotemporal modeling patterns for actor evolution, *i.e.*, Spatial-Temporal (ST) and Temporal-Spatial (TS), by switching the order of space and time as:

$$\mathbf{X}_{\text{ST}} = \text{T–Trans}(\mathbf{X} + \text{MLP}(\text{S–Trans}(\mathbf{X}))) \tag{4}$$
$$\mathbf{X}_{\text{TS}} = \text{S–Trans}(\mathbf{X} + \text{MLP}(\text{T–Trans}(\mathbf{X}))), \tag{5}$$

where we adopt a residual structure to enhance the actor representation. MLP with parameters in shape $C \times C$ is used to add non-linearity. By reshaping the frame and actor dimension as batch dimension, S–Trans and T–Trans reason about spatial and temporal actor interaction respectively.

By stacking spatial and temporal transformers in different orders, the actor representation is reweighted and aggregated according to different spatiotemporal context. ST path first reasons about the interaction of different actors in the scene of each frame. Then, the temporal evolution is modeled to reweight the built actor interaction across different frames. As such, ST path is skilled at recognizing activities with distinct spatial arrangement, such as *set* in volleyball games. This activity requires the player to move to a new position and set the ball, usually accompanied by other players moving or jumping for fake spiking. Complementarily, TS path reasons about the actor evolution, in the opposite order of ST path. It considers temporal dynamics of each actor in the first place, and then reasons about spatial actor interaction to understand the scene. Hence, it is skilled at recognizing activities with distinct actor evolution patterns, such as *spike* in volleyball games, which requires hitter to jump and quickly hit the ball.

Subsequently, to fully take advantage of such complementary characteristic, we feed the representation of actors from ST and TS paths to generate individual actions and group activity predictions, and fuse them as final predictions of dual spatiotemporal paths.

3.3. Multi-scale Actor Contrastive Learning

The actor representation is reweighted and aggregated by dual spatiotemporal paths, however, the modeling process is independent. To promote cooperation of these two complementary paths, we design a self-supervised Multi-scale Actor Contrastive loss (MAC-loss). As dual spatiotemporal paths model evolution of each actor in different patterns, we define a pretext task of actor consistency. Specifically, we design such constraints in multiple scales of frame and video levels.

Frame-Frame Actor Contrastive Loss. The frame representation of the actor in one path should be similar with its corresponding frame representation in the other path, while different from other frame representation of this actor in the path. As shown in Fig. 4 (a), taking actor n in ST path as an example, we attract frame representation in k-th frame ($\mathbf{X}_{\text{ST}}^{n,k}$) to its corresponding representation from TS path ($\mathbf{X}_{\text{TS}}^{n,k}$). Meanwhile, we repel the representation of actor

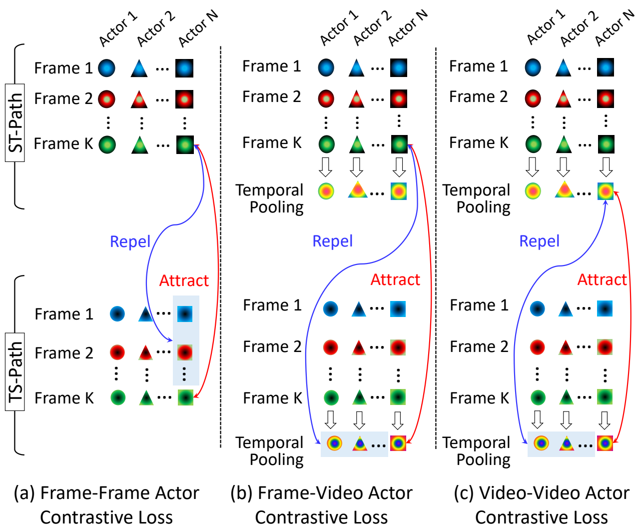

Figure 4. Illustration of MAC-loss for Actor N. It consists of three levels, *i.e*., frame-frame, frame-video and video-video. The blue block means the source of negative pairs. For simplicity, we only show the constraints from ST path to TS path. It is similar for the constraints from TS path to ST path.

n in other frames from TS path ($\mathbf{X}_{\text{TS}}^{n,t}$, where $t \neq k$),

$$\mathcal{L}_{ff}(\mathbf{X}_{\text{ST}}^{n,k}, \mathbf{X}_{\text{TS}}^{n,k}) = -\log \frac{h(\mathbf{X}_{\text{ST}}^{n,k}, \mathbf{X}_{\text{TS}}^{n,k})}{\sum_{t=1}^{K} h(\mathbf{X}_{\text{ST}}^{n,k}, \mathbf{X}_{\text{TS}}^{n,t})}, \quad (6)$$

where $h(\mathbf{u}, \mathbf{v}) = \exp(\frac{\mathbf{u}^\top \mathbf{v}}{||\mathbf{u}||_2 ||\mathbf{v}||_2})$ is the exponential of cosine similarity measure. Vice versa, the loss for actor n in TS path can be obtained by $\mathcal{L}_{ff}(\mathbf{X}_{\text{TS}}^{n,k}, \mathbf{X}_{\text{ST}}^{n,k})$.

Frame-Video Actor Contrastive Loss. The frame representation of the actor in one path should be consistent with its video representation in the other path, while different from video representation of other actors in the path. As shown in Fig. 4 (b), taking actor n in ST path as an example, we attract its frame representation $\mathbf{X}_{\text{ST}}^{n,k}$ to its video representation $\tilde{\mathbf{X}}_{\text{TS}}^{n}$ from TS path, which is obtained by pooling frame representation $\mathbf{X}_{\text{TS}}^{n,1:K}$. Meanwhile, we repel the video representation of other actors in the minibatch from TS path ($\tilde{\mathbf{X}}_{\text{TS}}^{i}$, where $i \neq n$),

$$\mathcal{L}_{fv}(\mathbf{X}_{\text{ST}}^{n,k}, \tilde{\mathbf{X}}_{\text{TS}}^{n}) = -\log \frac{h(\mathbf{X}_{\text{ST}}^{n,k}, \tilde{\mathbf{X}}_{\text{TS}}^{n})}{\sum_{i=1}^{B\times N} h(\mathbf{X}_{\text{ST}}^{n,k}, \tilde{\mathbf{X}}_{\text{TS}}^{i})}, \quad (7)$$

where B denotes the minibatch size. Vice versa, the loss for actor n in TS path can be obtained by $\mathcal{L}_{fv}(\mathbf{X}_{\text{TS}}^{n,k}, \tilde{\mathbf{X}}_{\text{ST}}^{n})$.

Video-Video Actor Contrastive Loss. Furthermore, we constrain the consistency of video representation of each actor across dual paths, as shown in Fig. 4 (c). We achieve this by minimizing cosine similarity measure $\mathcal{L}_{vv}$ of corresponding video representation ($\tilde{\mathbf{X}}_{\text{TS}}^{n}, \tilde{\mathbf{X}}_{\text{ST}}^{n}$). Our proposed MAC-loss is then formed as

$$\mathcal{L}_{MAC} = \lambda_{ff}\mathcal{L}_{ff} + \lambda_{fv}\mathcal{L}_{fv} + \lambda_{vv}\mathcal{L}_{vv}, \quad (8)$$

where $\lambda_{\{\cdot\}}$ denote weights for the different components.

3.4. Training objectives

Our network can be trained in an end-to-end manner to simultaneously predict individual actions of each actor and group activity. Combining with standard cross-entropy loss, the final loss for recognition is formed as

$$\mathcal{L}_{cls} = \mathcal{L}_{CE}(\frac{\hat{y}_{\text{ts}}^G + \hat{y}_{\text{st}}^G + \hat{y}_{scene}^G}{3}, y^G) + \lambda \mathcal{L}_{CE}(\frac{\hat{y}_{\text{ts}}^I + \hat{y}_{\text{st}}^I}{2}, y^I), \quad (9)$$

where $\hat{y}_{\{ts,st\}}^I$ and $\hat{y}_{\{ts,st\}}^G$ denote individual action and group activity predictions from TS and ST paths. y^I and y^G represent the ground truth labels for the target individual actions and group activity. $\hat{y}_{scene}^G$ denotes the scene prediction produced by separate group activity classifier, using features directly from backbone. λ is the hyper-parameter to balance the two items. Finally, we combine all the losses to train our Dual-AI framework,

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{MAC}. \quad (10)$$

During inference, we infer the individual actions and group activity by averaging the predictions from the dual spatiotemporal paths.

4. Experiments

4.1. Dataset

Volleyball Dataset. This dataset [21] consists of 4,830 labeled clips (3493/1337 for training/testing) from 55 volleyball games. Each clip is annotated with one of 8 group activity classes. Middle frame of each clip is annotated with 9 individual action labels and their bounding boxes.

Collective Activity Dataset. This dataset [12] contains 44 short videos with every ten frames annotated with individual action labels and their bounding boxes. The group activity class of each clip is determined by the largest number of the individual action classes. We follow [45, 46, 49] to merge the *crossing* and *walking* into *moving*.

Weak-Volleyball-M Dataset. This dataset [47] is adapted from Volleyball dataset while merging *pass* and *set* categories to have total 6 group activity classes, and discarding all individual annotations (including individual action labels and bounding boxes) for weakly supervised GAR.

NBA Dataset. This dataset [47] contains 9,172 annotated clips (7624/1548 for training and testing) from 181 NBA game videos, each of which belongs to one of the 9 group activities. No individual annotations, such as individual action labels and bounding boxes, are provided.

4.2. Implementation Details

We select the Inception-v3 model as our CNN backbone, following widely used settings [7, 24, 44, 48, 49] in GAR. We

Method	Backbone	Data Ratio	Optical Flow	Individual Action	Group Activity
HDTM [21]	AlexNet	100%		-	81.9
CERN [30]	VGG16	100%		-	83.3
StageNet [29]	VGG16	100%		-	89.3
HRN [20]	VGG19	100%		-	89.5
SSU [7]	Inception-v3	100%		81.8	90.6
AFormer [17]	I3D	100%		-	91.4
ARG [44]	Inception-v3	100%		83.0	92.5
TCE+STBiP [48]	Inception-v3	100%		-	93.3
DIN [49]	ResNet-18	100%		-	93.1
GFormer [24]	Inception-v3	100%		83.7	94.1
Ours	Inception-v3	25%		82.1	89.7
	Inception-v3	50%		83.0	92.7
	Inception-v3	100%		**84.4**	**94.4**
SBGAR [25]	Inception-v3	100%	✓	-	66.9
CRM [6]	I3D	100%	✓	-	93.0
Aformer [17]	I3D	100%	✓	83.7	93.0
JLSG [15]	I3D	100%	✓	83.3	93.1
ERN [28]	R50-FPN+I3D	100%	✓	81.9	94.1
GFormer [24]	I3D	100%	✓	84.0	94.9
Ours	Inception-v3	25%	✓	83.0	91.6
	Inception-v3	50%	✓	84.0	94.2
	Inception-v3	100%	✓	**85.3**	**95.4**

Table 1. Comparison with state-of-the-art methods on **Volleyball dataset** in term of Acc.%.

Method	Backbone	MPCA
HDTM [21]	AlexNet	89.7
PCTDM [45]	AlexNet	92.2
CERN-2 [30]	VGG-16	88.3
Recurrent [40]	VGG-16	89.4
stagNet [29]	VGG-16	89.1
SPA+KD [34]	VGG-16	92.5
PRL [19]	VGG-16	93.8
CRM [6]	I3D	94.2
ARG [44]	ResNet-18	92.3
HiGCIN [46]	ResNet-18	93.0
DIN [49]	ResNet-18	95.3
TCE+STBiP [48]	Inception-v3	95.1
Ours	ResNet-18	**96.0**
	Inception-v3	**96.5**

Table 2. Comparisons with previous state-of-the-art methods on **Collective Activity datatset**.

also use ResNet-18 model as backbone for Collective Activity Dataset, following widely used settings [46, 49]. We apply the ROI-Align with crop size 5×5 and a linear embedding to get actor features with dimension $C = 1024$. Each Spatial or Temporal transformer has one attention layer with 256 embedding dimension. The $\lambda_{ff}, \lambda_{fv}, \lambda_{vv}$ in MAC-Loss are all set 1.

For Vollyball and Weak-Volleyball-M, we randomly select $K = 3$ frames with 720×1280 resolution for training and 9 frames for testing, corresponding to 4 frames before the middle frame and 4 frames after. For Collective Activity dataset, we utilize $K = 10$ frames (480×720) of each video clip for training and testing. For NBA dataset, we select $K = 3$ frames (720×1280) around middle frame of each video for training and take 20 frames for testing. For Volleyball and Collective Activity dataset, we use annotated bounding boxes provided by the datasets for training and testing to make fair comparison, *i.e.*, $N = 12$ and $N = 13$ respectively. For NBA and Weak-Volleyball-M datasets, we detect person bounding boxes with MMDetection Toolbox [11] following [47], and set maximum actor number $N = 16$ and $N = 20$ respectively. More details can be found in supplementary materials.

Method	Backbone	Modality	NBA Acc./Mean Acc.	Weak Vlb.-M Acc.
TSN* [38]	Incep-v1	RGB	– / 37.8	–
I3D* [10]	I3D	RGB	– / 32.7	–
Nlocal* [42]	I3D-NLN	RGB	– / 32.3	–
ARG* [44]	Incep-v3	RGB	– / –	90.7
SAM [47]	Res-18	RGB	– / –	93.1
SAM [47]	Incep-v3	RGB	49.1 / 47.5	94.0
Ours	Incep-v3	RGB	51.5 / 44.8	95.8
	Incep-v3	Flow	56.8 / 49.1	96.1
	Incep-v3	Fusion	**58.1 / 50.2**	**96.5**

Table 3. Comparision with state-of-the-art methods on **NBA and Weak-Volleyball-M dataset** following metrics adopted in [47]. * means the results are from [47].

Method	5%	10%	25%	50%	100%
PCTDM [45]	53.6	67.4	81.5	88.5	90.3
AFormer [17]	54.8	67.7	84.2	88.0	90.0
HiGCIN [46]	35.5	55.5	71.2	79.7	91.4
ERN [28]	41.2	52.5	73.1	75.4	90.7
ARG [44]	69.4	80.2	87.9	90.1	92.3
DIN [49]	58.3	71.7	84.1	89.9	93.1
Ours	**76.2**	**85.5**	**89.7**	**92.7**	**94.4**

Table 4. Comparison with state-of-the-art methods trained with Volleyball dataset of different data ratios in term of group activity recognition Acc.%.

4.3. SOTA Comparison

Full Setting. This setting allows us to train our model with all data fully annotated with group activities and individual annotations. We compare our method with the state-of-the-art approaches on Volleyball and Collective Activity dataset. As shown in Tab. 1, our approach (94.4%) with only RGB frames and Inception backbone has already outperformed other SOTA methods with computationally high backbones (I3D, FPN) and additional optical flow input. Furthermore, equipped with RGB and optical flow late fusion, our method can improve the SOTA result by a large margin to 95.4%. Remarkably, even with only 50% data, our method still surpasses the vast majority of the SOTA methods with 100% data, *e.g.*, Ours (50%) vs. SARF (100%): 94.2 vs. 93.1. As shown in Tab. 2, our approach also achieves state-of-the-art performance on Collective Activity dataset. These results demonstrate the effectiveness of our method.

Weakly Supervised Setting. Under this setting we use all raw data and group activity annotations, without any individual annotations. We follow the [47] to report re-

Dual-Path	Weak Volleyball-M	Limited Volleyball	Full Volleyball
S-S	88.9	88.4	91.2
T-T	91.6	87.9	90.9
S-T	93.0	89.3	92.2
T-S	92.6	89.5	92.1
ST-ST Cross	92.1	88.7	91.7
ST-TS Fusion	**94.2**	**90.8**	**93.3**

Table 5. Effectiveness of our Dual Path Actor Interaction.

Components of MAC-loss			Data Ratio	
F-F	F-V	V-V	50%	100 %
			90.8	93.3
✓			91.2	93.5
	✓		91.0	93.3
		✓	91.6	93.6
✓	✓	✓	**92.1**	**94.0**

Table 6. Effectiveness of our MAC-loss. Different components are ablated on Volleyball dataset in term of Acc.%.

sults on Weak-Volleyball-M dataset and NBA dataset. As shown in Tab. 3, our method surpasses all the existing methods by a good margin, establishing new state-of-the-art results. Specifically, our approach improves the previous SOTA [47] by 2.5% on Weak-Volleyball-M and by 9% on NBA dataset in term of Acc.%. It indicates that our Dual-AI framework can enhance the learning ability of the model to obtain robust representation and achieve promising performance in the case individual annotations missing.

Limited Data Setting. In this setting, we train our method with random sampled data in different ratios to show the generalization power of our method. To compare the results under this setting, we implement a number of previous SOTA methods that have the officially-published codes available. As shown in Tab. 4, our method surpasses previous SOTA methods in all data ratios. Moreover, with the available training data decreasing, the performance of our method remains promising and the gain against other methods gets enlarged, which demonstrates the robustness of our method.

4.4. Ablation Study

Dual Spatial Temporal Paths. To validate the effectiveness of our Dual Spatiotemporal Paths, we investigate six settings. Particularly, we experiment with 50% data for limited Volleyball. In addition to T-S and S-T introduced in Section Sec. 3.2, other two paths, *i.e.*, S-S and T-T are introduced to validate in a broader range. S-S/T-T means that features go through two successive Spatial/Temporal-Transformer, respectively. ST-ST Cross denotes the way where features from Spatial-Transformer and Temporal-Transformer are fused in the middle and then fed into a second Spatial/Temporal-Transformer, to achieve a cross-enhanced spatiotemporal actor interaction. As shown in Tab. 5, our Dual Paths is better than ST-ST Cross and achieves the best result under different setting. The reason is that, dual-path TS and ST are good at inferring different group activities and the learned representation from ST and TS can complement each other, leading to a better performance. This demonstrates that our dual path ST-TS is a preferable way to comprehensively leverage both paths to generate robust spatiotemporal contexts for boosting group activity recognition.

Scene Fusion	Data Ratio	
	50%	100%
w/o	92.1	94.0
Early	92.0	93.9
Middle	92.2	94.0
Late	**92.7**	**94.4**

Table 7. Effectiveness of scene information.

SPE	TPE	Individual Action	Group Activity
		83.4	93.3
	✓	83.8	93.8
✓		84.0	93.7
✓	✓	**84.4**	**94.4**

Table 8. Impact of spatial and temporal transformer structure. Different combinations of PEs are ablated in term of Acc.%.

Multi-scale Actor Contrastive Loss. We explore the performance of our network with different components of MAC loss. As shown in Tab. 6, with different component of consistent loss (frame-frame, frame-video, video-video), our network consistently outperforms w/o consistent loss. By utilizing all components of MAC-loss, our network can achieve the best results. Note that, given less available training data, the loss can help network get a larger accuracy improvement. It demonstrates that the MAC-loss can enable cooperation of the dual complementary modeling process, thereby enhancing the learned representation from ST and TS paths, especially with limited available data.

Scene Information. We investigate the effectiveness of scene information, by exploring the way to fuse scene context in a early, middle and late fusion manner. As shown in Tab. 7, late scene context fusion is the best choice. Regardless of the available data ratio, the scene information can improve the performance by around 0.6 in term of Acc.%. This is because that scene information can provide global-level context, which can supplement the actor-level relation modeling and is crucial to GAR.

Spatial and Temporal Position Encoding. In the last ablation stage, we measure the importance of Spatial and Temporal Position Encoding. As shown in Tab. 8, either equipped with SPE or TPE, the performance of our method can be improved. These results demonstrate that SPE and TPE can provide useful spatial and temporal structure prior, which is beneficial to spatiotemporal action interaction learning.

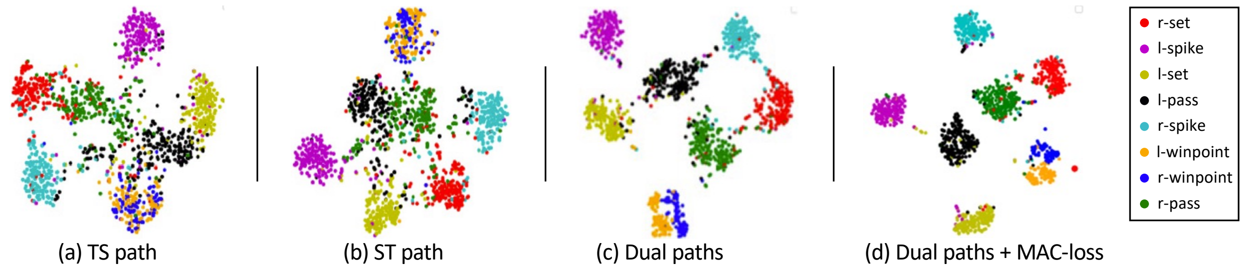

Figure 5. t-SNE [36] visualization of video representation on the Volleyball dataset learned by different variants of our Dual-AI model: ST path only, TS path only, Dual spatiotemporal paths, and final Dual-AI model.

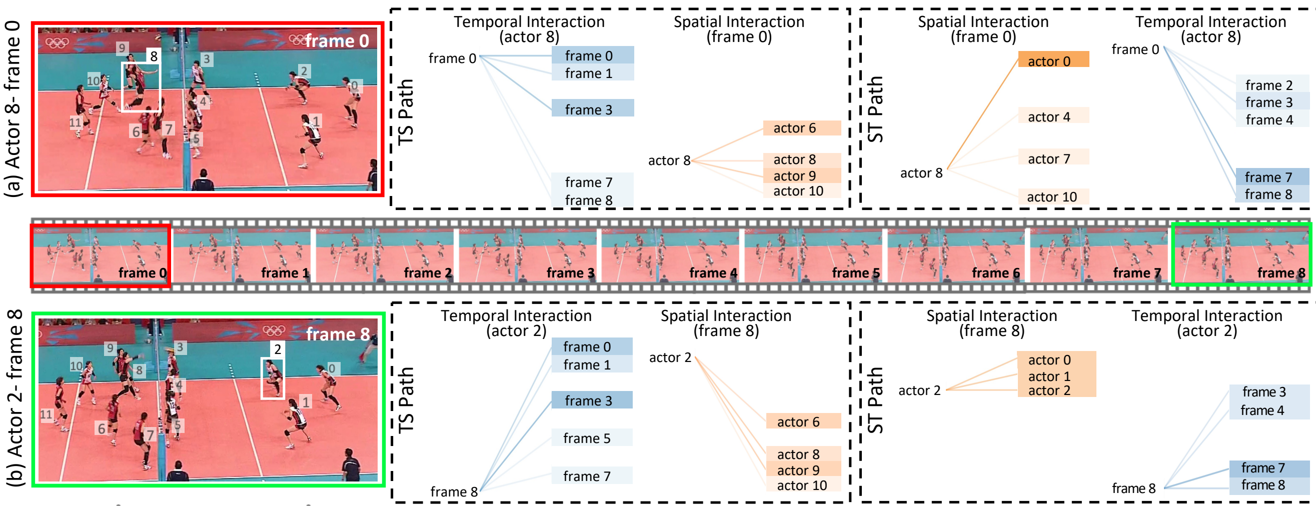

Figure 6. Actor interaction visualization for *l-spike* activity with connected lines. Brighter color indicates stronger relation. (a) For actor 8 in frame 0, we visualize the temporal interaction with same actors in different frames for ST and TS paths; similarly, we visualize the spatial interaction with different actors in frame 0. (b) We visualize the actor interaction for actor 2 in frame 8 in the same way.

4.5. Visualization

Group Feature Visualization. Fig. 5 shows the t-SNE [36] visualization of the learned representation. We project video representation extracted from Volleyball validation dataset to 2-D dimension using t-SNE. We can see that learned representation from Dual Path transformer (c) can be grouped better than single Temporal-Spatial path (a) and Spatial-Temporal path (b). Furthermore, equipped with MAC-loss, our Dual-AI network (d) is able to differentiate group representations much better. These results demonstrate the effectiveness of our Dual-AI framework.

Spatial/Temporal Actor Attention Visualization. We visualize the actor interaction of *l-spike* activity in Fig. 6. The attention weight between actors is represented by connected lines, and the brightness of the lines represents the scale of the attention weight. Orange and Blue lines correspond to the Spatial and Temporal interaction, respectively. As shown by spatial interaction in Fig. 6 (a), the spiking player (actor 8) is more related with accompanying players in TS path, who are "moving" (actor 6 and 10) and "standing" (actor 9). Differently, in ST path, actor 8 has wider connections with accompanying players (*e.g.*, actor 7 and actor 10) and defending players (*e.g.*, actor 0 and actor 4). Similarly, as shown by spatial interaction in Fig. 6 (b), the actor 2 is related to different accompanying and defending players in TS path and ST path respectively, showing complementary patterns. As for temporal interaction in both (a) and (b), the anchor actor is more related with early frames (frame 0 and frame 3) in TS path, while more related with late frames (frame 7 and frame 8) in ST path, showing highly complementary patterns.

5. Conclusion

In this work, we develop a Dual-AI framework to flexibly learn actor interactions in Spatial-Temporal and Temporal-Spatial views. Furthermore, we design a distinct MAC-loss to enable cooperation of dual paths for effective actor interaction learning. We conduct experiments on three datasets and establish new state-of-the-art results under different data settings. Particularly, our method with 50% data surpasses a number of recent methods trained with 100% data. The comprehensive ablation experiments and visualization results show that our method is able to learn actor interaction in a complementary way.

References

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

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

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

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

[5] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021. 2

[6] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional relational machine for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7892–7901, 2019. 1, 6

[7] Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4315–4324, 2017. 2, 3, 5, 6

[8] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning,*, 2021. 1, 2

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 4

[10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6299–6308, 2017. 2, 6

[11] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[12] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *ICCV Workshops*, pages 1282–1289. IEEE, 2009. 1, 5

[13] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4772–4781, 2016. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[15] Mahsa Ehsanpour, Alireza Abedin, Fatemeh Saleh, Javen Shi, Ian Reid, and Hamid Rezatofighi. Joint learning of social groups, individuals action and sub-group activities in videos. In *European Conference on Computer Vision*, pages 177–195. Springer, 2020. 2, 6

[16] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. *arXiv preprint arXiv:2104.11227*, 2021. 2

[17] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 839–848, 2020. 1, 2, 3, 4, 6

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

[19] Guyue Hu, Bo Cui, Yuan He, and Shan Yu. Progressive relation learning for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 980–989, 2020. 1, 2, 6

[20] Mostafa S Ibrahim and Greg Mori. Hierarchical relational networks for group activity recognition and retrieval. In *European Conference on Computer Vision*, pages 721–736, 2018. 2, 6

[21] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1980, 2016. 1, 2, 5, 6

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

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

[24] Shuaicheng Li, Qianggang Cao, Lingbo Liu, Kunlin Yang, Shinan Liu, Jun Hou, and Shuai Yi. Groupformer: Group activity recognition with clustered spatial-temporal transformer. *Proceedings of the IEEE international conference on computer vision*, 2021. 1, 2, 3, 5, 6

[25] Xin Li and Mooi Choo Chuah. Sbgar: Semantics based group activity recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2876–2885, 2017. 2, 6

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE international conference on computer vision*, 2021. 2

[27] Mandela Patrick, Dylan Campbell, Yuki M Asano, Ishan Misra Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, Jo Henriques, et al. Keeping your eye on the ball: Trajectory attention in video transformers. In *NeurIPS*, 2021. 2, 3

[28] Rizard Renanda Adhi Pramono, Yie Tarng Chen, and Wen Hsien Fang. Empowering relational network by self-attention augmented conditional random fields for group activity recognition. In *European Conference on Computer Vision*, pages 71–90. Springer, 2020. 2, 6

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

[30] Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. Cern: confidence-energy recurrent network for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5523–5531, 2017. 6

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

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

[33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3

[34] Yansong Tang, Zian Wang, Peiyang Li, Jiwen Lu, Ming Yang, and Jie Zhou. Mining semantics-preserving attention for group activity recognition. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1283–1291, 2018. 6

[35] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6450–6459, 2018. 1

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

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

[38] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016. 6

[39] Minsi Wang, Bingbing Ni, and Xiaokang Yang. Recurrent modeling of interaction context for collective activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3048–3056, 2017. 2

[40] Minsi Wang, Bingbing Ni, and Xiaokang Yang. Recurrent modeling of interaction context for collective activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, July 2017. 6

[41] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021. 2

[42] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 6

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

[44] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9964–9974, 2019. 1, 2, 3, 5, 6

[45] Rui Yan, Jinhui Tang, Xiangbo Shu, Zechao Li, and Qi Tian. Participation-contributed temporal dynamic model for group activity recognition. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1292–1300, 2018. 2, 5, 6

[46] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Higcin: hierarchical graph-based cross inference network for group activity recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 3, 5, 6

[47] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Social adaptive module for weakly-supervised group activity recognition. In *European Conference on Computer Vision*, pages 208–224. Springer, 2020. 1, 2, 3, 5, 6, 7

[48] Hangjie Yuan and Dong Ni. Learning visual context for group activity recognition. In *AAAI*, volume 35, pages 3261–3269, 2021. 3, 5, 6

[49] Hangjie Yuan, Dong Ni, and Mang Wang. Spatio-temporal dynamic inference network for group activity recognition. In *Proceedings of the IEEE international conference on computer vision*, 2021. 1, 2, 3, 5, 6

[50] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021. 2