mathematics

MDPI

Article

STAB-GCN: A Spatio-Temporal Attention-Based Graph Convolutional Network for Group Activity Recognition

Fang Liu [1], Chunhua Tian [1], Jinzhong Wang [2,*], Youwei Jin [2], Luxiang Cui [3] and Ivan Lee [4]

[1] School of Computer Science, Shenyang Aerospace University, Shenyang 110136, China; liufang@sau.edu.cn (F.L.); tianchunhua@stu.sau.edu.cn (C.T.)
[2] Public Basic Course Teaching and Research Department, Shenyang Sport University, Shenyang 110102, China; youwei@syty.edu.cn
[3] Sports Training College, Shenyang Sport University, Shenyang 110102, China; cuiluxiang@syty.edu.cn
[4] STEM, University of South Australia, Mawson Lakes 5095, Australia; ivan.lee@unisa.edu.au
* Correspondence: wjz@syty.edu.cn

Abstract: Group activity recognition is a central theme in many domains, such as sports video analysis, CCTV surveillance, sports tactics, and social scenario understanding. However, there are still challenges in embedding actors' relations in a multi-person scenario due to occlusion, movement, and light. Current studies mainly focus on collective and individual local features from the spatial and temporal perspectives, which results in inefficiency, low robustness, and low portability. To this end, a Spatio-Temporal Attention-Based Graph Convolution Network (STAB-GCN) model is proposed to effectively embed deep complex relations between actors. Specifically, we leverage the attention mechanism to attentively explore spatio-temporal latent relations between actors. This approach captures spatio-temporal contextual information and improves individual and group embedding. Then, we feed actor relation graphs built from group activity videos into our proposed STAB-GCN for further inference, which selectively attends to the relevant features while ignoring those irrelevant to the relation extraction task. We perform experiments on three available group activity datasets, acquiring better performance than state-of-the-art methods. The results verify the validity of our proposed model and highlight the obstructive impacts of spatio-temporal attention-based graph embedding on group activity recognition.

Keywords: group activity recognition; sports video analysis; graph convolutional network; spatio-temporal attention

MSC: 68T45

Citation: Liu, F.; Tian, C.; Wang, J.; Jin, Y.; Cui, L.; Lee, I. STAB-GCN: A Spatio-Temporal Attention-Based Graph Convolutional Network for Group Activity Recognition. *Mathematics* **2023**, *11*, 3074. https://doi.org/10.3390/math11143074

Academic Editor: Cornelio Yáñez Márquez

Received: 5 June 2023
Revised: 29 June 2023
Accepted: 7 July 2023
Published: 12 July 2023

1. Introduction

Group activity recognition (GAR) focuses on classifying the activity of a crowd of people working together in a video clip. It is a central topic in many domains, such as sports video analysis [1,2], CCTV surveillance [3–5], sports tactics [6], and social scenario understanding [7–9]. To comprehend the multi-player scenario, the model needs to recognize the individual actions and the group activity. Differing from general individual action recognition, GAR requires the deep and precise learning of spatio-temporal interactions between actors, which entails challenges such as the dynamics of actors and the complexity of their underlying correlations. It is expected to embed latent relations between actors from the spatial and temporal perspectives. Thanks to the graph convolutional network (GCN), this paper designs an effective Spatio-Temporal Attention-Based Graph Convolution Network (STAB-GCN) model for GAR.

Extensive efforts have been made to model actor relations for GAR in videos. In particular, deep learning methods have achieved remarkable improvements in embedding the relations between actors [10–13]. The existing models, as mentioned earlier, generally

have the following disadvantages: high computational costs, low flexibility, and over-fitting. In addition, GCN has been introduced to model the underlying relations between actors based on appearance and location features [2]. However, this paper only considers the local spatial relation between two actors whose distance is below a certain threshold, which may lead to the elimination of some latent important information for GAR.

To solve the aforementioned problem, our proposed STAB-GCN introduces an attention mechanism for actors involved in a group activity with the evolution of the spatio-temporal dimension. Specifically, we build multiple actor relation graphs to model the relations between the pairwise actors, referring to [2]. To the best of our knowledge, only some of the interactions in a basketball game positively impact GAR. In Figure 1, the frame marked with a star is the key frame for the group activity on the left. The red round box denotes the actor performing the key action on the right. Each node represents an actor. The solid line represents the relation between a pair of actors, and the thickness of the straight line represents the strength of the interaction. Actor 3 and actor 9 engage in shooting and defense, respectively. They run quickly to shoot and block at the same time, while a teammate (actor 2) moves up to set the screen on his defender (actor 6). These interactions among "jump shot", "block", and "screen" are considerably stronger than other relations, contributing more to GAR. Therefore, a crowd of actors having the closest underlying relations usually determine the type of group activity. STAB-GCN attends to the key frame and the key actor in a video clip, and performs reasoning about these important semantic activities ("jump shot" or "rebound") for GAR according to the graph structure. In these graphs, the node represents the actor, and their relations are denoted by the line between the two actors. Then, we take the actor relation graphs as input to STAB-GCN, aiming to localize and embed useful contexts of individual action and group activity. STAB-GCN designs a "soft pruning" scheme that converts the original dependency graph into weighted graphs with the full connection. The weight of the edge denotes the strength of the relationship between nodes, which can be learned in an end-to-end fashion by using a spatio-temporal attention mechanism [14]. Finally, our STAB-GCN model obtains a better graph embedding for group activity understanding.

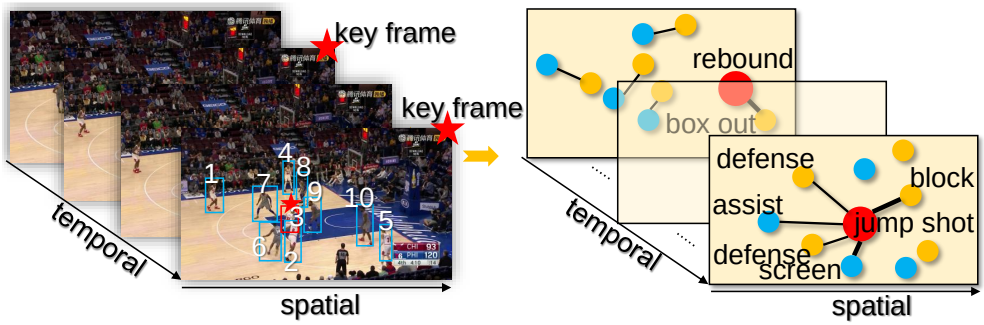

Figure 1. Exploring the latent relations between actors in multi-person scenarios (basketball game) with a spatio-temporal attention mechanism. The numbers are the codes for actors. The yellow and blue circles represent actors from two opposing teams. The red circle represents the key actor for GAR.

It should be emphasized that we consider the spatial relation between actors involved in a dynamic adaptive distance based on a global view, which is different from a localized actor relation graph, as shown in [2]. In addition, a set of K frames are sampled consistently from every video clip to ensure sufficient temporal semantic information. A series of experiments show that STAB-GCN achieves state-of-the-art performance on three datasets: the Volleyball dataset [10], the Collective Activity dataset [3], and the NBA dataset [15]. The results verify that STAB-GCN exceeds the existing models for GAR in terms of accuracy.

In this paper, our contributions are summarized as follows:

- We propose the novel STAB-GCN that uses the GCN with a spatio-temporal attention mechanism, which learns how to selectively attend to the features in videos. Our STAB-GCN model employs a multilayer structure to yield better graph embedding.
- We build elastic and effective actor relation graphs to capture key actors and the latent relations between actors in multi-person scenarios. It leverages an attention mechanism to dynamically embed the relationship strength between actors and yields multiple actor relation graphs of different structures with the evolution of the spatio-temporal dimension, thus effectively recognizing different group activities.
- The proposed model achieves state-of-the-art performance on the available datasets, i.e., the NBA dataset, the Volleyball dataset, and the Collective Activity dataset. Experimental results demonstrate that our STAB-GCN is efficient and effective in embedding spatio-temporal actor relations in GAR.

The remainder of the paper is organized as follows. Section 2 surveys the related works, and Section 3 introduces our proposed STAB-GCN. Section 4 describes the experimental results. The conclusions are presented in Section 5.

2. Related Work

2.1. Group Activity Recognition

In recent years, group activity recognition has attracted plenty of attention and has recently been applied in valuable work. The previous approaches leverage probabilistic graphical models [16,17] and AND-OR grammar methods [18], which mostly take advantage of hand-crafted features. Deep convolutional neural networks (CNN) [10,11], recurrent neural networks (RNNs) [12,13,16], and long short-term memory [19,20] have achieved outstanding performance owing to the spatio-temporal context and multidimensional information. Bagautdinov et al. [10] propose an integrated model for object detection and GAR by introducing a convolutional neural network that embeds the individual features and concatenates these acquired features to form a collective feature. Ibrahim et al. [11] build a two-stage framework to classify individual action and group action. Ibrahim et al. [13] design a hierarchical network model for the embedding of person-level information by leveraging an RNN. The representative works [15,21] propose a criss-cross graph to improve the recognition accuracy for GAR. Our work differs from the above-mentioned approaches in that it dynamically represents the interaction information by embedding the spatio-temporal features of individual and collective information. It also develops an attention-based GCN to acquire different relationships between actors.

2.2. Transformer Models

Transformer-based models have been paid more and more attention regarding the embedding of the semantic relations between collective features and they represent significant improvements in GAR [8,21–25]. Transformer usually lies on top of the actor features to learn spatio-temporal contexts with conditional random fields [21]. Kirill et al. [22] utilize Vanilla Transformer and I3D to represent actors' temporal features and construct actors' spatial relations. Li et al. [23] propose a cluster attention mechanism and leverage spatio-temporal contexts to efficiently explore collective features with Transformer. Yuan et al. [25] use Transformer to encode individual contexts to recognize individual activity. Bertasius et al. [26] propose TimeSformer to embed spatio-temporal relations with different space and time attention mechanisms. Fan et al. [27] aggregate multi-scale features to improve the embedding of spatio-temporal relations. Motionformer [28] proposes a self-attention block to track spatio-temporal patches for GAR. As discussed above, Transformer-based methods have become a widely applied backbone for video analysis occupations. However, several researchers still face great challenges in fully learning the latent relationships between actors in video clips. We propose a flexible and effective model, STAB-GCN, to embed deep relations between relative actors, which introduces a spatio-temporal attention-based GCN for GAR.

3. Methodology

To embed latent actor relations in multi-person scenarios, we propose a Spatio-Temporal Attention-Based Graph Convolution Network (STAB-GCN) model for GAR. We also give a detailed description in the following subsections. First, we offer an overview of the STAB-GCN model. Second, we describe our proposed spatio-temporal attention mechanism. Then, we present the effective and efficient embedding in the GCN. Finally, we introduce a new fusion strategy to optimize our STAB-GCN model.

3.1. The Model of STAB-GCN

As shown in Figure 2, our proposed STAB-GCN model is divided into three phases: spatio-temporal feature extraction, inferring and embedding on GCN, and feature aggregation. We build feature vectors of actors from sampled frames and construct multiple actor relation graphs. Then, we propose a spatio-temporal attention-based graph convolutional network to perform deep reasoning on the graphs. In the second stage, we feed the node embedding into the n attention guided layer to generate n adjacency matrices by utilizing the spatio-temporal attention mechanism, as shown in the figure; these are transformed into n different fully connected weighted graphs and fed into the densely connected layer to generate a new embedding for GAR, as shown at the lower right. Afterward, a combination layer is utilized to concatenate the outputs of the densely connected layer into a latent embedding. Eventually, we combine the initial feature and the latent feature into feature classifiers for GAR.

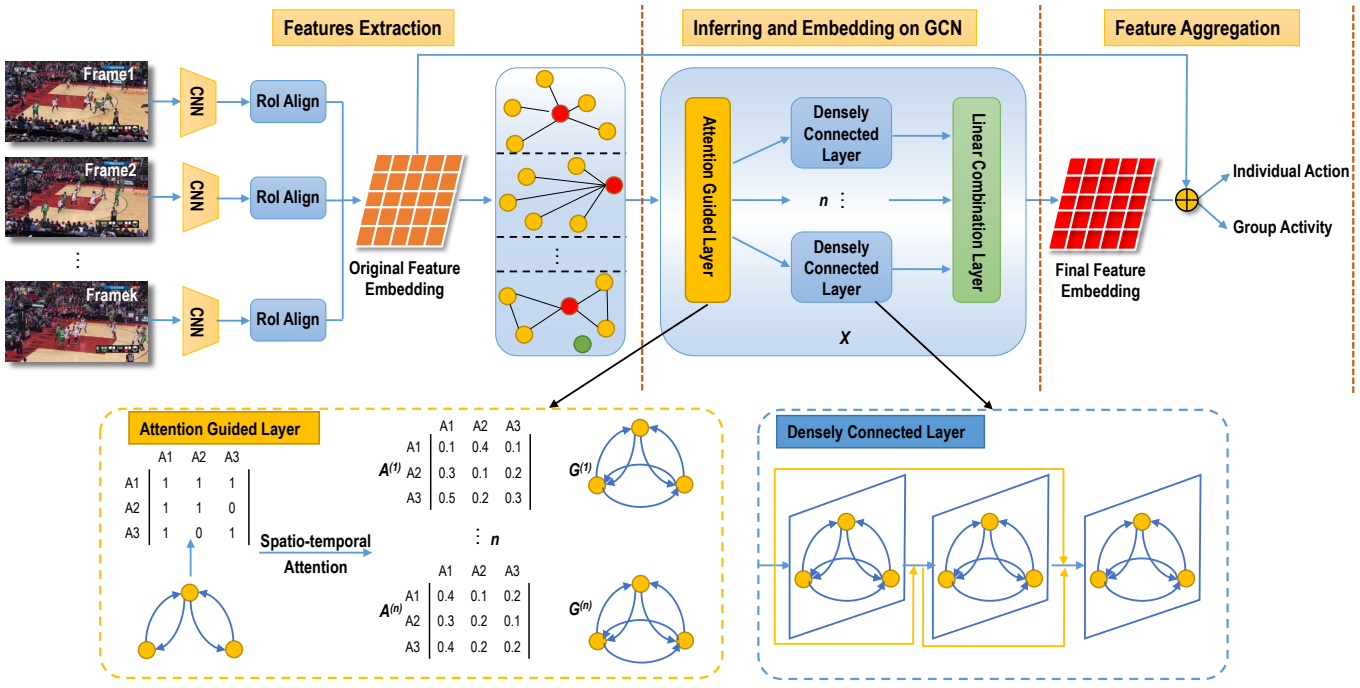

Figure 2. The proposed STAB-GCN model for GAR.

The uniform sampling technique obtains K frames in the first stage from a video clip. Then, we extract frame-level feature vectors with the strategy proposed in [10] based on the K frames. To verify the performance of STAB-GCN, we introduce Inception-v3 as a backbone to embed latent features for each sampled frame. Meanwhile, we also perform extensive experiments with other state-of-the-art backbones to show the superiority of STAB-GCN. Then, RoIAlign [29] is used to acquire the actor-level features on the frame-level features based on bounding boxes of N actors. After this, we leverage an fc layer to aggregate each actor feature into an M-dimensional vector, which is represented as a matrix $W \in \mathbb{R}^{K \times N \times M}$.

Given these actor features, we construct multiple actor relation graphs to embed latent relationships. In the graphs, every node denotes an actor in a multiple-person scenario and every edge denotes a relation between pairwise actors. The weight of every node denotes the relationship strength according to relative actors' appearance features and coordinate positions. Then, inferring and embedding in the GCN consists of X identical blocks, which include an attention-guided layer, a densely connected layer, and a combination embedding layer. The attention-guided layer mainly transforms multiple actor graphs into n fully connected weighted graphs and corresponding adjacency matrices by utilizing spatio-temporal attention, which further ensures no loss of latent valuable spatio-temporal relation information for GAR. Moreover, the attention-guided layer further highlights the spatio-temporal dependencies of different graph convolution features. Then, n adjacency matrices are input into n densely connected layers to produce a novel spatio-temporal embedding that is subsequently fed into a combination embedding layer and combined to generate the final feature embedding.

In the final stage, we leverage two classifiers to recognize individual actions and group activities by fusing the original feature embedding and final feature embedding. Specifically, we introduce a fully connected layer for the recognition of individual actions. Then, the actor-level feature vectors are max-pooled to produce group-level vectors. After this, we utilize another fully connected layer for the recognition of group activities.

3.2. Spatio-Temporal Attention Mechanism

As an important component of the Transformer network, the self-attention mechanism can also be successfully used to reason about actors' relations and interactions [20]. To embed latent relations for complex group activities, we propose a spatio-temporal attention mechanism to capture actor interactions, which is divided into a spatial attention unit and temporal attention unit in this paper. Following the Transformer concept, we describe how to build the spatial and temporal actor relation modules in detail.

First, we build actor relation graphs by introducing the method proposed in [2], where each node denotes an actor, and each edge denotes the relation between pairwise actors based on their appearance features and the 2D coordinates. We can obtain multiple graphs to embed latent relation features. Subsequently, the multiple graphs are fed into the spatio-temporal attention unit to acquire a more comprehensive relation representation.

3.2.1. Spatial Attention Unit

First, we design a spatio-temporal actor Transformer. From the spatial dimension, $Y^j \in \mathbb{R}^{N \times M}$ denotes the initial feature vectors of N actors in the j-th frame. Furthermore, we utilize the Transformer model $\hat{Y}^j = S - Trans(Y^j)$ to attend to the spatial relationships among these actors, which mainly includes three important modules as follows:

$$Y' = SPE(Y^j) + Y^j, \tag{1}$$

$$Y'' = LN(Y' + MHSA(Y')), \tag{2}$$

$$\hat{Y}^j = LN(Y'' + FFN(Y'')). \tag{3}$$

In this paper, spatial position encoding (SPE) is applied to capture the spatial features of actors and generate the spatial feature vectors Y' in the scenario, as shown in Equation (1). Meanwhile, we leverage the central point coordinates of each actor's bounding box to represent the spatial positions, which are subsequently encoded with the position encoding (PE) function from [22]. Then, we introduce a multi-head self-attention (MHSA) module to embed the spatial interaction of the actors [14] and acquire the corrective feature vectors Y'', as shown in Equation (2). Lastly, we utilize a feed-forward network (FFN) to further boost the performance of the spatial relation inference [14], as shown in Equation (3). $\hat{Y}^j$ denotes the final spatial feature vectors of N actors in the j-th frame.

3.2.2. Temporal Attention Unit

To capture the temporal features of each actor across frames, we design a temporal attention unit according to the Transformer model. We take the feature vectors of the i-th actor as an input across K frames, as shown in $Z^i \in \mathbb{R}^{K \times M}$. We also introduce temporal position encoding (TPE) to encode the temporal features of K frames, as shown in Equation (4). Furthermore, we use the MHSA module to attend to the temporal evolution of actor i across different time steps, as shown in Equation (5). We similarly use FFN to improve the learning accuracy of the temporal relation unit, as shown in Equation (6). Finally, the actor features improved by temporal interactions are obtained by $\hat{Z}^i = T - Trans(Z^i)$.

$$Z' = TPE(Z^i) + Z^i, \tag{4}$$

$$Z'' = LN(Z' + MHSA(Z')), \tag{5}$$

$$\hat{Z}^i = LN(Z'' + FFN(Z'')). \tag{6}$$

Z^i denotes the initial temporal feature vectors of the i-th actor. Z' and Z'' denote the modified temporal feature vectors. $\hat{Z}^i$ denotes the generative temporal feature vectors of the i-th actor.

3.2.3. Spatio-Temporal Actor Relation Embedding

Based on actors' spatial and temporal relations, we construct a spatio-temporal feature embedding matrix E through the tensor product, as shown in Equation (7). The actor embedding is reweighted and combined in terms of the diversified spatio-temporal context by observing and integrating spatial and temporal relative features. The spatial attention unit mainly focuses on the relationships between different actors in the scenario. In contrast, the temporal attention unit is able to capture the evolution of actor interactions across different frames. Thus, we propose a spatio-temporal attention mechanism to recognize multiple activities with different spatio-temporal patterns.

$$E = \left[\hat{Y}^1, \quad \hat{Y}^2, \quad \ldots, \quad \hat{Y}^K\right] \otimes \left[\hat{Z}^1, \quad \hat{Z}^2, \quad \ldots, \quad \hat{Z}^N\right]'. \tag{7}$$

By leveraging the embedding of actors due to a spatio-temporal attention mechanism, we can predict individual actions and group activities more efficiently and effectively.

3.3. Embedding on Graph Convolutional Network

Unlike state-of-the-art strategies, which obtain a smaller graph than the initial one, we leverage a spatio-temporal attention mechanism to build a larger, fully connected graph. Then, we introduce a densely connected layer into the STAB-GCN model to mine deeper information on large graphs, which contributes to the capture of rich local and global information to learn a better actor relation.

In the densely connected layer, we define $g_j^{(l)}$ as the integration of the initial node embedding x_j and the node embedding induced in different layers ($h_j^{(1)} \cdots h_j^{(l-1)}$). The node embedding in the l layer is shown as follows:

$$g_j^{(l)} = \left[x_j; h_j^{(1)}; \cdots; h_j^{(l-1)}\right]. \tag{8}$$

Each densely connected layer generally consists of L sub-layers whose dimension is determined by their number and the entered feature dimension d. In this paper, we set $L = 4$ and $d = 1024$, respectively. Then, we obtain the new embedding by concatenating the output of each sub-layer. To our knowledge, the dimension of the sub-layers in the original GCN is not smaller than the input dimension. However, we reduce the dimension to improve the learning efficiency further.

Considering n attention-guided adjacency matrices, we utilize n densely connected layers correspondingly. The inference of each layer is listed as follows:

$$h_{t_i}^{(l)} = \sigma\left(\sum_{j=1}^{N} \tilde{A}_{ij}^{(t)} W_t^{(l)} g_j^{(l)} + b_t^{(l)}\right), \tag{9}$$

where $t = 1, \cdots, n$, $W_t^{(l)} \in \mathbb{R}^{\frac{d}{L} \times d^{(l)}}$ denotes the weight matrix, and $d^{(l)} = d + \frac{l-1}{L} \times d$. $h_{t_i}^{(l)}$ denotes the induced representation of node i at the l-th layer in the t-th densely connected layer. σ is an activation function. $b_t^{(l)}$ denotes the bias vector. d denotes the input feature dimension.

The STAB-GCN model introduces the combination embedding layer to generate an embedding from n separate densely connected layers. Specifically, the final output H'' is shown as

$$H'' = W'H' + b', \tag{10}$$

where H' denotes the output obtained by integrating the outputs obtained from the densely connected layers, and $H' = \left[h^{(1)}; \cdots; h^{(n)}\right] \in \mathbb{R}^{d \times n}$. W' and b' denote a weight matrix and a bias vector. In the end, STAB-GCN fuses the output relational features with the original features to generate the scenario embedding, which is taken as the input of two classifiers to make individual action and group activity predictions.

3.4. Training Objective

Our model is trained with the standard cross-entropy loss in an end-to-end manner. In STAB-GCN, $\overline{S}_g$ denotes the score of group activity recognition, which is obtained from our proposed framework. $\overline{S}_i$ denotes the score of individual action prediction based on actor feature embedding. In this paper, we introduce the cross-entropy loss to optimize the training process, as shown in Equation (11).

$$\Psi = \Psi_1(S_g, \overline{S}_g) + \lambda \Psi_2(S_i, \overline{S}_i), \tag{11}$$

where Ψ_1 and Ψ_2 are the functions of cross-entropy loss. S_i denotes the ground truth for the individual actions. S_g denotes the ground truth for group activities. λ represents the hyper-parameter to equalize the two terms.

4. Experiments

4.1. Datasets

In our experiments, two public datasets (the Volleyball dataset and the Collective Activity dataset) are leveraged for GAR. In addition, we introduce an available NBA dataset to verify the performance of our proposed model.

The Volleyball dataset [11] is made up of 4830 clips originating from 55 volleyball games; it includes 4 group activity categories, i.e., pass, spike, set, and winpoint. Meanwhile, 9 individual action labels (spiking, jumping, waiting, failing, setting, standing, digging, moving, standing) and the player's bounding boxes are used in the middle frame of each clip. Following the settings in [11], we leverage 3220 clips as a training set and 1610 clips as a testing set. Thanks to [10], we solve the problem of the lack of benchmark bounding boxes for unlabelled frames.

The Collective Activity dataset [3] consists of 44 video clips, which are divided into 5 group activities, namely crossing, waiting, queuing, walking, and talking, and 6 individual actions, namely N/A, crossing, queuing, waiting, talking, and walking. Group activity annotation depends on the most individual action labels in a clip. According to the settings in [30], we select two thirds of the video clips as a training set and the rest as the testing set.

The NBA dataset contains 181 NBA games from 2019, downloaded from the web, with 9172 video clips and 9 group activity categories: two points success, two points failure, layup success, layup failure, three points success, and three points failure for offensive

rebounds. We choose 7624 clips for training and 1548 clips for testing, in accordance with the experimental settings of [15].

4.2. Implementation Details

In the process of implementation, Inception-v3 is used as our backbone network and our detailed settings are consistent with [25,31]. To verify the performance of STAB-GCN, we also utilize state-of-the-art models as the backbone networks for other datasets and make a further comparison with previous approaches. We leverage linear embedding to obtain actor feature vectors with 1024 dimensions. STAB-GCN has a 256-dimensional two-tier architecture. PyTorch is used as the implementation platform in this paper. We utilize 4 Tesla T4 GPUs to analyze the video clips.

4.3. Comparison with Up-to-Date Methods

We made a series of experiments to compare STAB-GCN with state-of-the-art methods on the Volleyball dataset. As observed in Table 1, the accuracy of STAB-GCN is 94.8% with Inception-v3 and 84.3% with ResNet-18 for group activity and individual action recognition, respectively, showing better performance. STAB-GCN surpasses other up-to-date approaches for group activity recognition. Although STAB-GCN is almost equal to the suboptimal method (Dual [1]) for individual action recognition, it is 0.4 percentage points better than Dual for GAR in terms of accuracy. STAB-GCN is able to embed the most relevant feature information between actors. Furthermore, STAB-GCN outperforms ARG [2] by a good margin, which is attributed to the proposed attention mechanism to capture the key actors and the latent relationship strength.

Table 1. Performance comparison between STAB-GCN and other methods (Volleyball).

Method	Backbone	Individual Action	Group Activity
HDTM [11]	AlexNet	-	81.9
CERN [32]	VGG-16	-	83.3
StagNet [30]	VGG-16	-	89.3
HRN [13]	VGG-19	-	89.5
AFormer [22]	I3D	-	91.4
DIN [31]	ResNet-18	-	93.1
SSU [10]	Inception-v3	81.8	90.6
ARG [2]	Inception-v3	83.0	92.5
TCE + STBiP [25]	Inception-v3	-	93.3
GFormer [23]	Inception-v3	83.7	94.1
Dual [1]	Inception-v3	84.4	94.4
STAB-GCN	Inception-v3	82.9	94.8
STAB-GCN	ResNet-18	84.3	92.2

Referring to Table 2, we obtain the experimental results on the Collective Activity dataset. STAB-GCN leverages the spatio-temporal attention-based GCN to achieve 96.5% accuracy for GAR, which verifies the generalization and effectiveness of our method in embedding the latent relations between actors. Dual acquires 95.2% accuracy by leveraging a dual-path actor interaction framework and is 1.3 percentage points worse than STAB-GCN.

As shown in Table 3, STAB-GCN outperforms the mentioned approaches by a large margin. The accuracy of STAB-GCN is 0.6 percentage points higher than that of state-of-the-art methods on the NBA dataset. The results demonstrate that STAB-GCN can boost the embedding ability and acquire collective activity representations. According to the above-mentioned analysis, STAB-GCN is fully competent for GAR.

Table 2. Performance comparison between STAB-GCN and other methods (Collective Activity).

Method	Backbone	Group Activity
SIM [12]	AlexNet	81.2
HDTM [11]	AlexNet	87.5
PCTDM [33]	AlexNet	92.2
CERN [32]	VGG-16	87.2
StagNet [30]	VGG-16	89.1
PRL [34]	VGG-16	93.8
SPA + KD [35]	VGG-16	92.5
CRM [36]	I3D	94.2
ARG [2]	ResNet-18	91.0
DIN [31]	ResNet-18	95.3
HiGCIN [37]	ResNet-18	93.0
TCE + STBiP [25]	Inception-v3	95.1
SBGAR [16]	Inception-v3	86.1
Dual [1]	ResNet-18	95.2
STAB-GCN	Inception-v3	92.1
STAB-GCN	ResNet-18	96.5

Table 3. Performance comparison between STAB-GCN and other methods (NBA).

Method	Backbone	Group Activity
SACRF [24]	Inception-v3	56.3
TSM [38]	Inception-v1	66.6
AFormer [22]	ResNet-18	47.1
DIN [31]	ResNet-18	61.6
SAM [15]	ResNet-18	54.3
ARG [2]	Inception-v3	59.0
DFW [39]	ResNet-18	75.8
Dual [1]	Inception-v3	50.2
STAB-GCN	Inception-v3	78.1
STAB-GCN	ResNet-18	78.4

4.4. Ablation Study

To analyze the effectiveness of the different components of our proposed STAB-GCN model, we implement a series of ablation studies on the NBA dataset by utilizing the recognition accuracy metric.

4.4.1. Multiple Sub-Layers

To obtain latent relation information, we verify the effectiveness of the number of sub-layers in the densely connected layer on the NBA dataset. We conduct some relative experiments on a different number of sub-layers. Referring to Table 4, we find that using multiple sub-layers is helpful for group activity accuracy, compared with only one sub-layer, and can improve the accuracy from 75.3% to 78.4%. When the number of sub-layers is 4, the accuracy in group activity recognition is the best.

Table 4. Effectiveness of number of sub-layers L.

Sub-Layer	1	4	8	16	32
Accuracy	75.3	78.4	78.1	77.5	76.8

4.4.2. Effectiveness of Spatio-Temporal Attention-Based GCN

We perform ablation studies with the following settings to verify the effectiveness of different attention mechanisms on the spatio-temporal semantic information captured by STAB-GCN. (1) None: we do not use an attention mechanism. (2) SA: we utilize a spatial attention mechanism to capture the most important features of sampled frames without considering temporal evolution. (3) STA: we utilize a spatio-temporal attention mechanism to embed the latent features of individual and group activities. (4) STAB-GCN: we jointly embed spatial and temporal context information and integrate it with the GCN. With the exception of the different components of the model, all settings are the same. As shown in Table 5, our proposed STAB-GCN improves the performance in individual and group activity recognition from 74.2% to 78.4% on the NBA dataset. Furthermore, STAB-GCN obtains a performance boost by 4.2%, which is mainly due to the spatio-temporal attention-based GCN. The results also demonstrate that STAB-GCN is effective for GAR.

Table 5. Effectiveness of different components of STAB-GCN.

Method	NBA	Volleyball
None	74.2	90.6
SA	75.9	92.4
STA	77.6	93.2
STAB-GCN	78.4	94.8

4.4.3. Scene Information

In this paper, we adopt different fusion methods to verify the effectiveness of scene information, i.e., early stage, middle stage, and late stage. Referring to Table 6, late scene fusion is better than the other two methods. Specifically, the accuracy is improved by around 0.4 percent. The scene information includes the global context semantics, which helps to infer the relations between actors and obtain a more effective and efficient feature embedding for GAR.

Table 6. Effectiveness of scene information.

Scene Fusion	NBA	Volleyball
w/o	78.0	94.0
Early	77.6	93.6
Middle	78.1	94.1
Late	78.4	94.8

4.5. Visualization Analysis

As shown in Figure 3, we considered the *t*-SNE [40] visualization of group activity feature embedding on the NBA dataset. (1) FC: we only use FC layers instead of STAB-GCN. (2) SA: we use only a spatial attention mechanism. (3) STA: we use a spatio-temporal attention mechanism. (4) STAB-GCN: we use the spatio-temporal-based GCN. Specifically, we transform the group activity representation into a two-dimensional map based on *t*-SNE. The feature embedding from STAB-GCN can be clustered more effectively compared with FC, SA, and STA. The results show that our STAB-GCN is suitable for the recognition of group activity.

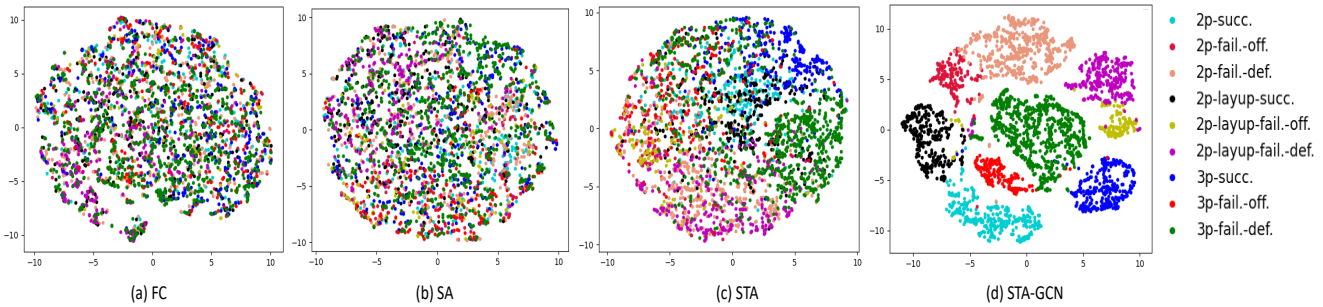

Figure 3. *t*-SNE [40] visualization of effectiveness of different components (NBA).

As shown in Figure 4, we considered the visualization of the STAB-GCN attention maps on the NBA dataset. The results imply that STAB-GCN can capture the related actors and the most important group activities by utilizing the spatial attention unit, the temporal attention unit, and the graph convolutional network, which greatly improves the accuracy in the recognition of individual actions and group activities in the video clips.

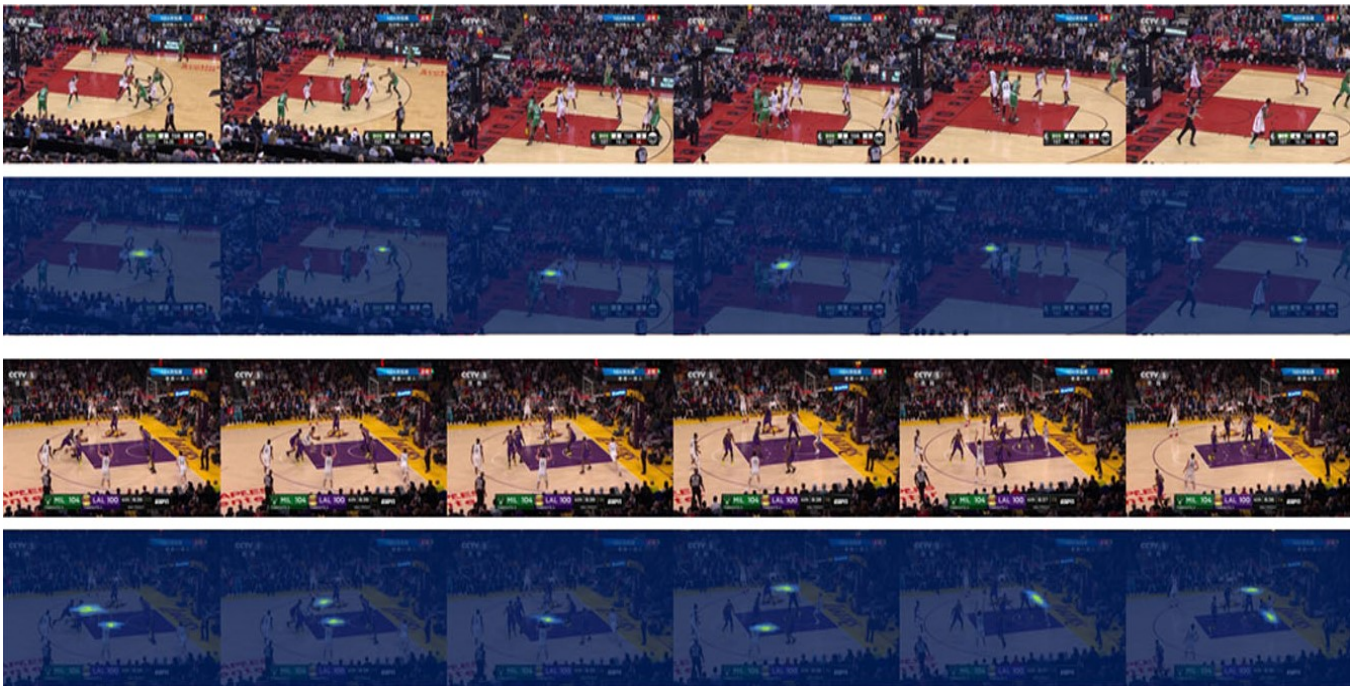

Figure 4. Visualization of our proposed STAB-GCN attention maps on the NBA dataset.

5. Conclusions

This paper proposes the Spatio-Temporal Attention-Based Graph Convolutional Network to embed the latent relations between actors. We present a series of experiments on three public datasets to verify the performance of our proposed STAB-GCN. The proposed model leverages the spatial and temporal attention mechanism to further infer the latent actor features on graph convolutional networks, and it achieves a good outcome compared with up-to-date methods. In the end, the experimental results demonstrate that STAB-GCN can embed actor interactions in a multi-person scenario.

Author Contributions: Conceptualization, F.L. and J.W.; methodology, J.W.; software, C.T. and J.W.; validation, F.L., C.T. and J.W.; formal analysis, J.W.; investigation, J.W.; resources, J.W.; data curation, J.W. and I.L.; writing—original draft preparation, J.W.; writing—review and editing, F.L. and I.L.; visualization, C.T. and J.W.; supervision, J.W.; project administration, Y.J.; funding acquisition, L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Doctoral Research Startup Fund Program of Liaoning Province under Grant No. 2020-BS-272, and the Scientific Research Fund Project of the Educational Department of Liaoning Province under Grant No. LJKZ1052 and No. LQN2020ST02.

Data Availability Statement: The data that support the findings of this study are openly available at https://github.com/mostafa-saad/deep-activity-rec#dataset (accessed on 3 September 2022), http://www.eecs.umich.edu/vision/activity-dataset.html (accessed on 18 October 2022), https://ruiyan1995.github.io/SAM.html (accessed on 28 October 2022), reference number [3,10,15].

Conflicts of Interest: The corresponding authors declare on behalf of all authors that there is no conflict of interest.

References

1. Han, M.F.; Zhang, D.J.; Wang, Y.L.; Yan, R.; Yao, L.N.; Chang, X.J.; Qiao, Y. Dual-AI: Dual-path Actor Interaction Learning for Group Activity Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 2980–2989.
2. Wu, J.C.; Wang, L.M.; Wang, L.; Guo, J.; Wu, G.S. Learning actor relation graphs for group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 9964–9974.
3. Choi, W.G.; Shahid, K.; Savarese, S. What are they doing? Collective activity classification using spatio-temporal relationship among people. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Kyoto, Japan, 29 September–2 October 2009; pp. 1282–1289.
4. Yu, S.; Xia, F.; Li, S.H.; Hou, M.L.; Sheng, Q.Z. Spatio-Temporal Graph Learning for Epidemic Prediction. *ACM Trans. Intell. Syst. Technol.* **2023**, *14*, 1–25. [CrossRef]
5. Abdel-Basset, M.; Hawash, H.; Chang, V.; Chakrabortty, R.K.; Ryan, M. Deep Learning for Heterogeneous Human Activity Recognition in Complex IoT Applications. *IEEE Internet Things J.* **2022**, *9*, 5653–5665. [CrossRef]
6. Kong, L.; Pei, D.; He, R.; Huang, D.; Wang, Y. Spatio-Temporal Player Relation Modeling for Tactic Recognition in Sports Videos. *IEEE Trans. Circ. Syst. Video Technol.* **2022**, *32*, 6086–6099. [CrossRef]
7. Bourached, A.; Gray, R.; Griffiths, R.R.; Jha, A.; Nachev, P. Hierarchical graph-convolutional variational autoencoding for generative modelling of human motion. *arXiv* **2022**, arXiv:2111.12602.
8. Li, K.C.; Wang, Y.L.; Zhang, J.H.; Gao, P.; Song, G.L.; Liu, Y.; Li, H.S.; Qiao, Y. Uniformer: Unifying convolution and self-attention for visual recognition. *arXiv* **2022**, arXiv:2201.09450.
9. Yu, S.; Xia, F.; Wang, Y.R.; Li, S.H.; Febrinanto, F.G.; Chetty, M.H. PANDORA: Deep Graph Learning Based COVID-19 Infection Risk Level Forecasting. *IEEE Trans. Comput. Soc. Syst.* **2022**, 1–14. [CrossRef]
10. Bagautdinov, T.; Alahi, A.; Fleuret, F.; Fua, P.; Savarese, S. Social scene understanding: End-to-end multiperson action localization and collective activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4315–4324.
11. Ibrahim, M.S.; Muralidharan, S.; Deng, Z.W.; Vahdat, A.; Mori, G. A hierarchical deep temporal model for group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1971–1980.
12. Deng, Z.W.; Vahdat, A.; Hu, H.X.; Mori, G. 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 (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4772–4781.
13. Ibrahim, M.S.; Mori, G. Hierarchical relational networks for group activity recognition and retrieval. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 721–736.
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
15. Yan, R.; Xie, L.X.; Tang, J.H.; Shu, X.B.; Tian, Q. Social Adaptive Module for Weakly-Supervised Group Activity Recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 208–224.
16. Li X.; Chuah, M.C. Sbgar: Semantics based group activity recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22 October 2017; pp. 2876–2885.
17. Kumar A.; Rawat, Y.S. End-to-End Semi-Supervised Learning for Video Action Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 14680–14690.
18. Shu, T.M.; Xie, D.; Rothrock, B.; Todorovic, S.; Zhu, S.C. Joint inference of groups, events and human roles in aerial videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4576–4584.
19. Yan, R.; Shu, X.; Yuan, C.; Tian Q.; Tang, J. Position-Aware Participation-Contributed Temporal Dynamic Model for Group Activity Recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 7574–7588. [CrossRef] [PubMed]
20. Tang, J.; Shu, X.; Yan, R.; Zhang, L. Coherence Constrained Graph LSTM for Group Activity Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 636–647. [CrossRef] [PubMed]

21. 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 Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 5 March 2020; pp. 71–90.
22. Gavrilyuk, K.; Sanford, R.; Javan, M.; Snoek, C.G. Actor-transformers for group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–20 June 2020; pp. 839–848.
23. Li, S.C.; Cao, Q.G.; Liu, L.B.; Yang, K.L.; Liu, S.N.; Hou, J.; Yi, S. Groupformer: Group activity recognition with clustered spatial-temporal transformer. In Proceedings of the IEEE Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 13668–13677.
24. Pramono, R.R.A.; Fang, W.H.; Chen, Y.T. Relational reasoning for group activity recognition via self-attention augmented conditional random field. *IEEE Trans. Image Process.* **2021**, *20*, 4752–4768. [CrossRef] [PubMed]
25. Yuan, H.J.; Ni, D. Learning visual context for group activity recognition. In Proceedings of the Conference on Association for the Advance of Artificial Intelligence (AAAI), Online, 2–9 February 2021; pp. 3261–3269.
26. Bertasius, G.; Wang, H.; Torresani, L. Is space-time attention all you need for video understanding? In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021.
27. Fan, H.Q.; Xiong, B.; Mangalam, K.; Li, Y.H.; Yan, Z.C.; Malik, J.; Feichtenhofer, C. Multiscale vision transformers. In Proceedings of the International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 6804–6815.
28. Patrick, M.; Campbell, D.; Asano, Y.M.; Metze, I.M.F.; Feichtenhofer, C.; Vedaldi, A.; Henriques, J. Keeping your eye on the ball: Trajectory attention in video transformers. In Proceedings of the Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 6–14 December 2021.
29. He, K.M.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2961–2969.
30. Qi, M.; Wang, Y.; Qin, J.; Li, A.; Luo, J.; Gool, L.V. StagNet: An Attentive Semantic RNN for Group Activity and Individual Action Recognition. *IEEE Trans. Circ. Syst. Video Technol.* **2020**, *30*, 549–565. [CrossRef]
31. Yuan, H.J.; Ni, D.; Wang, M. Spatio-temporal dynamic inference network for group activity recognition. In Proceedings of the IEEE conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 7456–7465.
32. Shu, T.M.; Todorovic, S.; Zhu, A.C. CERN: Confidence-energy recurrent network for group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4255–4263.
33. Yan, R.; Tang, J.H.; Shu, X.B.; Li, Z.C.; Tian, Q. Participation-contributed temporal dynamic model for group activity recognition. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 1292–1300.
34. Hu, G.Y.; Cui, B.; He, Y.; Yu, S. Progressive relation learning for group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–20 June 2020; pp. 980–989.
35. Tang, Y.S.; Wang, Z.A.; Li, P.Y.; Lu, J.W.; Yang, M.; Zhou, J. Mining semantics-preserving attention for group activity recognition. In Proceedings of the 26th ACM international conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 1283–1291.
36. Azar, S.M.; Atigh, M.G.; Nickabadi, A.; Alahi, A. Convolutional relational machine for group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 7892–7901.
37. Yan, R.; Xie, L.X.; Tang, J.H.; Shu, X.B.; Tian, Q. Higcin: Hierarchical graph-based cross inference network for group activity recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 6955–6968. [CrossRef] [PubMed]
38. Lin, J.; Gan, C.; Han, S. Tsm: Temporal shift module for efficient video understanding. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–3 November 2019; pp. 7083–7093.
39. Kim, D.K.; Lee, J.S.; Cho, M.S.; Kwak, S. Detector-Free Weakly Supervised Group Activity Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 20083–20093.
40. Maaten, L.V.; Hinton, G. Visualizing data using t-sne. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.