Learning Semantics-Preserving Attention and Contextual Interaction for Group Activity Recognition

Yansong Tang, *Student Member, IEEE*, Jiwen Lu, *Senior Member, IEEE*, Zian Wang, Ming Yang, *Member, IEEE*, and Jie Zhou, *Senior Member, IEEE*

***Abstract*— In this paper, we investigate the problem of group activity recognition by learning semantics-preserving attention and contextual interaction among different people. Conventional methods usually aggregate the features extracted from individual persons by pooling operations, which lack physical meaning and cannot fully explore the contextual information for group activity recognition. To address this, we develop a Semantics-Preserving Teacher-Student (SPTS) networks architecture. Our SPTS networks first learn a Teacher Network in the semantic domain that classifies the *word* of group activity based on the *words* of individual actions. Then, we design a Student Network in the appearance domain that recognizes the group activity according to the input video. We enforce the Student Network to mimic the Teacher Network in the learning procedure. In this way, we allocate semantics-preserving attention to different people, which is more effective to seek the key people and discard the misleading people, while no extra labeled data are required. Moreover, a group of people inherently lie in a graph-based structure, where the people and their relationship can be regarded as the nodes and edges of a graph, respectively. Based on this, we build two graph convolutional modules on both the Teacher Network and the Student Network to reason the dependency among different people. Furthermore, we extend our approach on action segmentation task based on its intermediate features. The experimental results on four datasets for group activity analysis clearly show the superior performance of our method in comparison with the state-of-the-art.**

***Index Terms*— Semantics-preserving, attention, group activity recognition, Teacher-Student networks.**

I. Introduction

GROUP activity recognition (*a.k.a.* collective activity recognition), which refers to discerning what a group of people are doing in a video, has attracted growing attention in the realm of computer vision over the past decade [1]–[7]. There are wide real-world applications for group activity recognition including traffic surveillance, social role understanding and sports video analysis. Compared with conventional action recognition which focuses on a single person, group activity recognition is a more challenging task as it requires further understanding of high-level relationships among different people. Hence, it is desirable to design a model to aggregate the individual dynamics across people and exploit their contextual information for effective group activity recognition.

Over the past few years, great efforts have been devoted to mining the contextual information for group activity recognition. In the early period, a typical series of approaches are developed to design graph-based structure models based on hand-crafted features [7]–[10]. However, these methods require strong prior knowledge and lack discriminative power to model the temporal evolution of group activity. In recent years, with the spectacular progress of deep learning methods, researchers have attempted to build different deep neural networks [2], [3] for group activity recognition. Most of these methods treat all participants with equal importance, and integrate the features of individual actions by simple pooling operators. However, the group activity is usually sensitive to a few key persons, whose actions essentially define the activity, and other people may bring ambiguous information and mislead the recognition process. Let's take Fig. 1 as an example. The bottom of Fig. 1 shows a frame sampled from a video clip in Volleyball dataset [2]. Obviously, the "spiking" person shall provide more discriminative information for recognizing the "right spike" activity, and those "standing" people may bring some confounding information. To address these, several attention-based methods [5], [11] have been proposed to assign different weights to different people. Specifically, the weights are learned based on the features extracted from input videos, and are allocated to their corresponding features. However, such a "self-attention" scheme essentially lacks physical explanation and is not reliable enough to find the key person for activity recognition.

In this work, we move a new step towards the interaction of appearance domain and semantic domain, and propose a Semantics-Preserving Teacher-Student (SPTS) model for

Manuscript received September 3, 2018; revised March 2, 2019 and April 26, 2019; accepted April 29, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant U1813218, Grant 61822603, Grant U1713214, Grant 61672306, and Grant 61572271. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tolga Tasdizen. *(Corresponding author: Jiwen Lu.)*

Y. Tang, J. Lu, Z. Wang, and J. Zhou are with the State Key Laboratory of Intelligent Technologies and Systems, Beijing Research Center for Information Science and Technology (BNRist), Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: tys15@mails.tsinghua.edu.cn; lujiwen@tsinghua.edu.cn; wza15@mails.tsinghua.edu.cn; jzhou@tsinghua.edu.cn).

M. Yang is with Horizon Robotics, Inc., Beijing 100080, China (e-mail: ming.yang@horizon-robotics.com).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TIP.2019.2914577

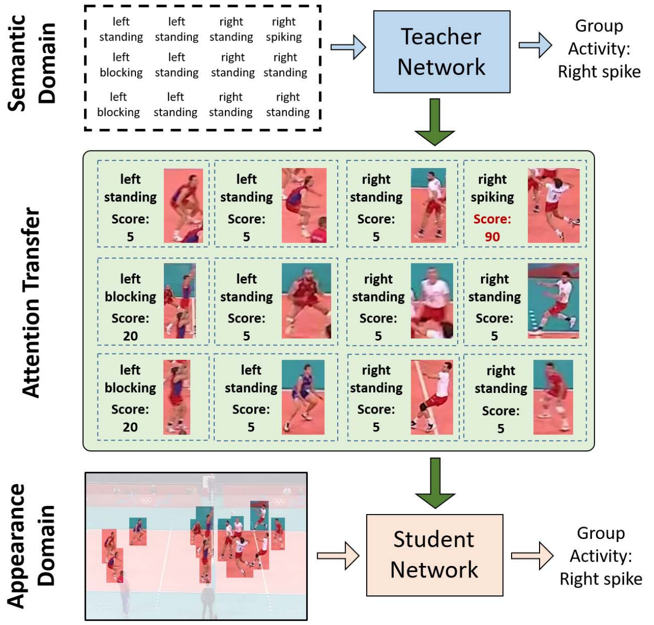


Fig. 1. The basic idea of the SPTS networks. In the semantic domain, the task is to map the *words* of individual actions, which can be treated as a caption of the video [4], to the *word* of group activity. In the appearance domain, we attempt to predict the label of group activity based on the corresponding input video. We first learn a Teacher Network in the semantic domain, and then employ the learned attention information, which represents the different importance of different people for recognizing the group activity, to guide a Student Network in the appearance domain. (Best viewed in color.)

group activity recognition. Fig. 1 shows the basic idea of our approach. Concretely, we first learn a high-performance model with typical attention mechanism (namely Teacher Network) to map the individual actions to group activity in the semantic domain. Next, we develop another model (namely Student Network), which predicts the group activity from the individual actions in the appearance domain. Then, we design a unified framework to utilize the attention knowledge in the Teacher Network to guide the Student Network. As the inputs of our Teacher Network are generated from the off-the-shelf single-action labels, our method requires no extra labelled data and only takes additional 2.70% computational time cost. Moreover, most conventional methods model the features of group people as regular tensor-based vectors, which ignore the intrinsic dependency among different people. To address this, we construct two types of graphs in semantic domain and appearance domain, respectively. The nodes of the graph contain the extracted features of the individual persons, while the adjacency matrices that encode their spatial coordinates are used to describe the relationship among different people. Since the graph of features lies in a non-Euclidean space, we further build two graph convolutional modules on both the Teacher Network and the Student Network to reason the relationship among different people. Besides, we propose a new approach for segmenting group activities in untrimmed videos, which is based on the intermediate features of our model and temporal convolutional networks [12]. We evaluate our approach on the Volleyball dataset, Collective Activity Dataset, Collective Activity Extended Dataset and Choi's Dataset, where the experimental results show that the SPTS networks outperform the state-of-the-arts for group activity analysis.

Our main contributions are summarized as follows:

1) In contrast to recent works for group activity recognition which utilize the appearance clues only, we have developed a Teacher Network to leverage the prior knowledge in the semantic domain, which requires no extra labelled data and a little additional computational time cost.
2) Different from existing self-attention based works, we have explored the discriminative information of different people by transferring the semantics-preserving attention learned by the Teacher Network to the Student Network in the appearance domain. Towards this, we equip the Teacher Network and Student Network with two attention modules and design an objective function which enforces the Student Network to mimic the Teacher Network. To our best knowledge, these are original efforts leveraging attention in both semantics and appearance clues, to perform group activity recognition.
3) Unlike most conventional works which model the features of people as regular tensors, we have constructed two types of graph for different people according to their spatial coordinates, and built two graph convolutional modules on the Teacher Network and Student Network to reason about the relationship of different people. Extensive experimental results on four widely used datasets have shown the effectiveness of our proposed method.
4) We have extended our method for action segmentation task based on its intermediate features. With the new designed model, the temporal intervals of group activities in an untrimmed sequence can be accurately segmented and our method achieves very competitive performance on this task.

It is to be noted that a preliminary conference version of this work was initially presented in [13]. As an extension, our SPTS with two new graph convolutional modules can better exploit the interaction information of different people. Moreover, we have conducted experiments on other two datasets and provided more in-depth analysis on the experimental results. Furthermore, we have extended our approach on action segmentation task for untrimmed videos and demonstrate its effectiveness. Besides, we have presented analysis on the computational time cost of our work.

II. Related Work

In this section, we briefly review four related topics: 1) group activity recognition, 2) attention-based models, 3) knowledge distillation, and 4) graph convolutional network.

A. Group Activity Recognition

Activity recognition is one of the most important issues in computer vision [14]–[18], where group activity recognition is an active sub-topic and various methods have been explored in recent years [1]–[7], [19]. These methods can be roughly divided into two categories: hand-crafted feature based and deep learning feature based methods. For the first category, a number of researchers fed hand-crafted features into graphical models to capture the structure of group activity. For example, Lan *et al.* [9] presented a latent variable framework to model the contextual information of person-person interaction and group-person interaction. Hajimirsadeghi *et al.* [1] developed a multi-instance model to count the instances in a video for group activity recognition. Shu *et al.* [10] employed AND-OR graph formalism to jointly group people, recognize event and infer human roles in aerial videos. However, these methods relied on hand-crafted features, which require strong prior knowledge and were short of discriminative power to capture the temporal cue.

For the deep learning based methods, numbers of works have been proposed to leverage the discriminative power of deep neural network for group activity recognition. For example, Ibrahim *et al.* [2] proposed a hierarchical model with two LSTM networks, where the first LSTM captured the dynamic cues of each individual person, and the second LSTM learned the information of group activity. Shu *et al.* [3] extended this work by replacing the softmax layer of the RNN with a new energy layer to improve reliability and numerical stability of inference. Wang *et al.* [6] built another LSTM network upon this work to capture the interaction context of different people. More recently, Ibrahim *et al.* [20] developed a Hierarchical Relational Network architecture to calculate the relational representation of people and describe their potential interactions. However, the works mentioned above mainly focused on the appearance domain, which ignored the semantic relationship between the individual actions and group activity. More recently, Li *et al.* [4] presented a SBGAR scheme, which generated the captions of each video and predicted the final activity label based on these captions. However, the generated captions were not always reliable, and the inferior captions will do harm to the final process of recognition. To this end, we simultaneously explore the contextual relationship of individual actions and group activity in both semantic and appearance domains, and employ the semantic knowledge to enhance the performance of vision task.

B. Attention-Based Models

Attention-based model is motivated by the attention mechanism of primate visual system [21], [22]. It aims to select the most informative parts from the global field. In the past two decades, attention-based models have been widely applied into the realm of natural language processing (*e.g.,* machine translation [23], [24]), computer vision (*e.g.,* video face recognition [25], [26], person re-identification [27], object localization [28]), and their intersection (*e.g.,* image caption [29], video caption [30] and visual question answering [31]). As for human action/activity recognition, Liu *et al.* [32] developed global context-aware attention LSTM networks to select the informative joints in skeleton-based videos. Furthermore, Song *et al.* [33] proposed a spatial-temporal attention-based model to learn the importance of different joints and different frames. Different from these two works [32], [33], we employ the attention model to allocate different weights to different people in a group for RGB-based activity recognition. Although a few works [5], [11] have exploited attention-based models for group activity recognition, they only applied "self-attention" scheme and were incapable to explain the physical meaning of the learned attention explicitly. Different from these methods, our SPTS networks distill the attention knowledge in the semantic domain to guide the appearance domain, which utilize the semantic information adequately and make the learned attention interpretable by further showing the visualization results.

C. Knowledge Distillation

The concept of "knowledge distillation" is originated from the work [34] by Hinton *et al.*, which aims to transfer the knowledge in a "teacher" network with larger architecture and higher performance to a smaller "student" network. They enforced a constraint on the softmax outputs of the two networks when optimizing the student network. After that, several works have been proposed to regularize the two networks based on the intermediate layers [29], [35], [36]. For example, Yim *et al.* [36] utilized flow of solution procedure (FSP) matrix, which were generated based on feature maps of two layers, to transfer knowledge in teacher network to student network. Chen *et al.* [37] employed technique of function-preserving transformations to accelerate the learning process of student network. The most related work to ours is [29], which also utilized the information across the attention modules of two networks. Different from [29], where the inputs of the two networks were both images and the networks architecture were similar, our work explores the knowledge in two different domains (semantic domain and appearance domain) and utilizes the additional recurrent neural network to address a more challenging task of group activity recognition.

D. Graph Convolutional Network

Recently, there has been progress in the formulation of convolutional neural network on graphs (*i.e.* graph convolutional network) [38]–[41] thanks to the development of graph signal processing (GSP) [42]. Given inputs on the nodes of the graph, the graph convolutional network (GCN) aims to learn representative features like standard CNN, which sheds lights on new possibilities to adopt data-driven method and perform convolutional operator on non-Euclidean space. Computer vision has also benefited from GCN in recent years [43], [44]. For example, Wang *et al.* [45] considered the semantic embeddings as different nodes of the knowledge graph, and adopted graph convolutional network to promote the problem of zero-shot recognition. Wang *et al.* [46] proposed a Graph Reasoning Model (GRM) to study the problem of social relationship understanding. For human action recognition, several works [47]–[49] have been proposed to develop graph convolutional network for skeleton-based action recognition. Unlike these works which regarded the coordinates of human joints as the nodes of the graph, we construct the nodes of the graph according to the features of individual person in both semantic domain and appearance domain. Then, we employ two graph convolutional modules to model the relationship of different people and enhance the recognition performance.

III. Approach

The motivation of this work is to adequately explore the information in both appearance domain and semantic domain

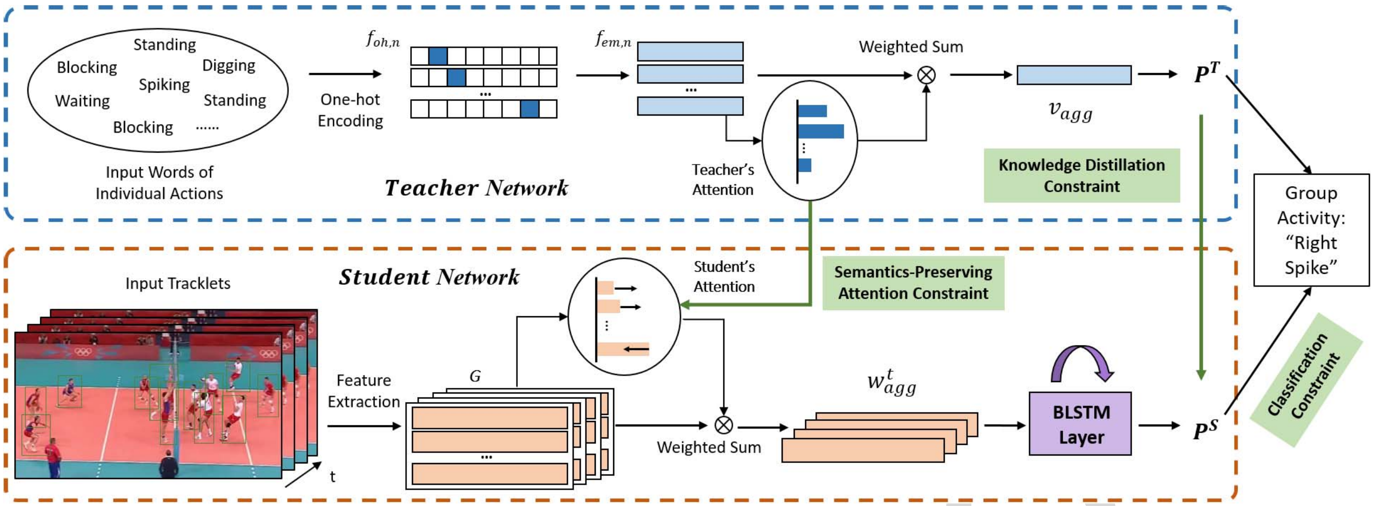


Fig. 2. A framework of our proposed SPTS networks, which contain two sub-networks. We first train the Teacher Network, which models relationship between words of individual actions and the word of group activity. Next, we train the Student Network, which takes a set of tracklets as input and predicts the label of group activity. We enforce three types of constraints during the training process of Student Network, *i.e.,* semantics-preserving attention constraint, knowledge distillation constraint and classification constraint.

for group activity recognition. In this section, we first formulate the problem, then we present the details of our SPTS networks and introduce how to build several graph convolutional modules on the SPTS. Finally we discuss the difference of our models with other related works.

A. Problem Formulation

We denote a tri-tuple (V, y, z) as a training sample for a video clip, where V is the specific video and z is the ground-truth label for group activity. Let $Y = \{y_n\}_{n=1}^N$ denote the labels of individual actions, where y_n represents the label corresponding to the nth person. The goal of group activity recognition is to infer the final label z corresponding to V during testing phase. Previously, researchers usually utilize a set of tracklets of the people in the video as inputs. The tracklets are denoted as $X = \{x_1^t, x_2^t, ...x_n^t, ...x_N^t\}_{t=1}^T$. where t represents the time stamp of the tth frame. We follow this problem setting in our work.

B. SPTS Networks

Our SPTS networks consist of two subnetworks, namely Student Network and Teacher Network. Fig. 2 illustrates the pipeline of SPTS networks. In this framework, the Student Network aims to predict the final label z given a set of tracklets from an input video in the appearance domain, while the Teacher Network aims to model the relationship between the *words* of individual actions $Y = \{y_n\}_{n=1}^N$ and the *word* of group activity z in the semantic domain. It is reasonable that Teacher Network tends to achieve comparable or better performance than Student Network, because individual action labels are powerful low-dimensional representations for the task of group action recognition, which is also demonstrated in the Experiments section. Additionally, we find the Teacher Network and Student Network are complementary in classification results, which indicates that jointly considering the semantic domain and appearance domain will help. However, the ground-truth individual labels $Y = \{y_n\}_{n=1}^N$ are not available during the testing stage. A natural way to address this issue is to employ the knowledge of the Teacher Network to guide the training process of the Student Network. We now detail the proposed SPTS networks as follows.

1) Student Network: The goal of our Student Network is to learn a model $z = \mathbf{S}(X; \theta_s)$ to predict the label of group activity given a set of tracklets in a video clip, where θ_s is the set of learnable parameters of the Student Network. For a fair comparison, we utilize the off-the-shelf tracklets provided by [2], [7].

In order to capture the appearance information and temporal evolution of each single person, we employ a DCNN network and an LSTM network to extract features of X, which is a similar scheme according to [2]. Then, we concatenate the features of the last fc layers of the DCNN and the LSTM network. The concatenation, denoted as $G = \{g_1^t, g_2^t, ...g_n^t, ..., g_N^t\}_{t=1}^T$, represents the temporal feature of each individual person. Sequentially, we calculate the score s_n^t which indicates the importance of the nth person as:

$$s_n^t = tanh(W_1 g_n^t + b_1)\,, \tag{1}$$

where W_1 and b_1 are the weighted matrix and biased term. The activation weight we allocate to each person is obtained as follow:

$$\beta_n^t = exp(s_n^t)/\sum_{j=1}^N exp(s_j^t)\,, \tag{2}$$

where β_n^t is the score normalized by a softmax function. Instead of conventional aggregation methods like max-pooling or mean-pooling, we fuse the feature of each individual person at time-step t as:

$$w_{agg}^t = \sum_{n=1}^N \beta_n^t \cdot g_n^t\,. \tag{3}$$

In this way, the set of activation factors $\{\beta_n^t\}_{n=1}^N$ control the contribution of each person to the aggregated feature w_{agg}^t.

Having obtained w^t_{agg}, the aggregated features of each frame, we feed them into another group-level bidirectional LSTM network. The output features are sent into an fc layer activated by a softmax function to obtain the final label of the group activity.

2) Teacher Network: As illustrated above, our Student Network can be regarded as an extension of the hierarchical deep temporal model [2] by adopting a typical self-attention mechanism. However, in such a scheme, the labels of individual actions and group activities are utilized to supervise the discriminative feature learning, while their corresponding relationship, which captures the dependency of the individual actions and group activities in the semantic domain, is rarely used. In this section, we introduce a Teacher Network, which aims to learn a model $z = \mathbf{T}(Y; \theta_t)$ to integrate the labels of individual actions $Y = \{y_n\}_{n=1}^N$ into a label of group activity z. Note that our Teacher Network essentially addresses an NLP-related task, where attention mechanism also shows its advantage. Based on this, we develop our Teacher Network by introducing an attention scheme, which is similar to our Student Network.

Given a set of individual action labels $Y = \{y_n\}_{n=1}^N$ as the input of our Teacher Network, we first encode them into a sequence of one-hot vectors $F_{oh} = \{f_{oh,n}\}_{n=1}^N$, where $f_{oh,n} \in R^C$ and C is the number of individual action category. Then we embed the $F_{oh} \in R^{P\times C}$ into a latent space as:

$$f_{em,n} = ReLU(W_2 f_n + b_2)\,, \tag{4}$$

where W_2 and b_2 are the weighted matrix and biased term, $ReLU$ denotes the nonlinear activation function [50]. Then another attention mechanism, which is corresponding to that of the Student Network, is derived as follow:

$$s_n = tanh(W_3 f_{em,n} + b_3)\,, \tag{5}$$

$$\alpha_n = exp(s_n) / \sum_{j=1}^{N} exp(s_j)\,, \tag{6}$$

$$v_{agg} = \sum_{n=1}^{N} \alpha_n \cdot f_{em,n}\,. \tag{7}$$

Having obtained the v_{agg}, we feed it into an fc layer followed by a softmax activation to predict the final label. We train the Teacher Network using the ground-truth labels of Y and z. It is relatively easy to classify a set of words in the semantic domain, thus the Teacher Network will achieve higher performance as illustrated in the Experiments section.

3) Semantics-Preserving Attention Learning: As we described, there are two attention modules in our method and they both work separately via a self-attention scheme. Noticing the fact that they both model the importance of different people, a valid question is why not jointly consider these two modules. More specially, as the Teacher Network directly takes the ground-truth label of individual actions as inputs, it is reasonable that its performance is better than the Student Network, which takes the tracklets as inputs and requires a more complex feature learning process before the attention module.

Based on this reason, we aim to use the attention knowledge of the Teacher Network to guide the Student Network.

Algorithm 1 SPTS

Input: Training samples: $\{X, Y, z\}$, Parameters: Γ (iterative number) and ϵ (convergence error).
Output: The weights of the Student Network θ_s.
// Teacher Network Training:
Optimize the parameter θ_t of the Teacher Network with (Y, z).
// Student Network Training:
Finetune the DCNN and the train first LSTM with (X, Y) [2].
Extract features G from X.
Initialize θ_s.
Perform forward propagation.
Calculate the initial $\boldsymbol{J}_0$ by (8).
for $i \leftarrow$ *1, 2, ...,* Γ **do**
  Update θ_s by back propagation through time (BPTT).
  Perform forward propagation.
  Compute the objective function $\boldsymbol{J}_i$ using (8).
  If $|\boldsymbol{J}_i - \boldsymbol{J}_{i-1}| < \epsilon$, go to **Return**.
end
Return: The parameters θ_s of the Student Network.

In practice, we first train the Teacher Network $\mathbf{T}(Y; \theta_t)$ with the provided labels of training samples. Then, we enforce the Student Network to absorb the teacher's knowledge during the learning process via a total loss function defined as below:

$$\begin{aligned} J &= J_{CLS} + \lambda_1\ J_{SPA} + \lambda_2\ J_{KD} \\ &= -\sum_{l=1}^{L} \mathbb{1}(z = l) log(P_S^l) \\ &\quad + \lambda_1 \frac{1}{N} \sum_{n=1}^{N} (\alpha_n - \frac{1}{T} \sum_{t=1}^{T} \beta_n^t)^2 \\ &\quad + \lambda_2 \| P_T - P_S \|_2^2 \end{aligned} \tag{8}$$

Here λ_1 and λ_2 are the hyper-parameters to balance the effects of two different terms to make a good trade-off. The physically interpretations of the J_{CLS}, J_{SPA} and J_{KD} are respectively explained as below.

The first term J_{CLS} represents classification loss for activity recognition. We calculate the categorical cross-entropy loss, where $\mathbb{1}$ is the indicator function which equals 1 when the prediction $z = l$ is true and 0 otherwise. Here l and L denote the predicted label and the number of the total activity categories. The softmax output P_S^l represents the corresponding class probability of the Student Network. The second term J_{SPA} aims to enforce the student's attention to preserve the teacher's semantics attention. We adopt the mean squared distance for these two types of attention. The third term J_{KD} denotes the loss of knowledge distillation [34], in which P_T and P_S are the softmax outputs of the Teacher and Student Network respectively.

To optimize (8), we employ the back propagation through time (BPTT) algorithm [51] for learning all the parameters θ_s of our Student Network. We summarize the pipeline of our SPTS method in **Algorithm 1**. Note that the Teacher Network only guides the Student Network during the training phase, as the ground-truth label $Y = \{y_n\}_{n=1}^{N}$ is not available during the testing stage.

C. SPTS + GCN

Since a group of people can be considered as a graph-based structure, where the node and edge represents each individual person and the relationship between two people respectively, we further build two graph-based modules upon our SPTP networks to adequately explore the contextual information of different people for group activity recognition.

1) Graph Construction: We construct a graph $\mathcal{G}(U, A)$ to model each frame, where U and A are the nodes sets and adjacency matrix respectively. On the one hand, we denote $U = \{u_1, u_2, ..., u_N\}$, where $u_n \in D$ is corresponding to the feature of the nth person. On the other hand, motivated by the fact that, the relationship of different people are highly correlated to the distance among them, we define the adjacency matrix A according to the spatial coordinates of different people as follow:

$$a_{mn} = exp(-\frac{||c_m - c_n||_2^2}{2}), \tag{9}$$

where c_m represent the central location of the mth person:

$$c_m = (\gamma \frac{x_{m,mid}}{W\_I}, \gamma \frac{y_{m,mid}}{H\_I}). \tag{10}$$

Here, W_I and H_I are the width and height of each frame respectively. $x_{m,mid}$ and $y_{m,mid}$ are the central positions of the input tracklets at the x axis and y axis. The γ is a scale factor, where we set it to be 10 empirically. In this way, we embed the spatial information into the adjacency matrix A. If two people m and n approach each other in the space, the corresponding a_{mn} will have a large value, and vice versa.

2) Graph Convolutional Layer: Since the graph of people lie in a non-Euclidean space, we leverage the graph-based convolutional Networks (GCN) [39] to learn the spatial dependency between different people. Mathematically, we can represent a layer of the graph convolution as:

$$Z = AUW, \tag{11}$$

where W are the learned parameters. Unlike conventional convolutional operator that reasons about the regular structure locally, the graph convolutional layer passes messages among different nodes and updates each nodes according to the predefined adjacency matrix A, which allows us to better capture the contextual information among different people. Moreover, we can stack multiple layers of graph convolution to better model the non-linear structure among people.

3) Building GCN Upon SPTS: Fig. 3 displays the illustration of building GCN upon our SPTS. For the Teacher Network, we perform graph convolution on the one-hot vector F_{oh} of each video clip:

$$Z_{teacher} = AF_{oh}W_{teacher}, \tag{12}$$

where A is obtained based on the middle frame of the video clip. The output feature $Z_{teacher}$ is then fed into the attention mechanism of the Teacher Network.

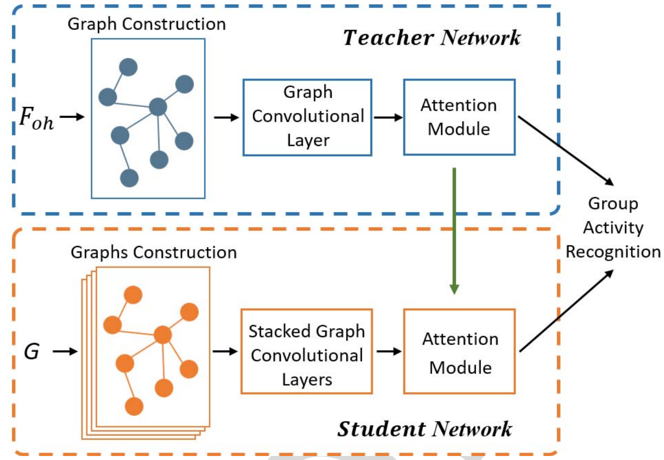


Fig. 3. **Flowchart of building graph convolutional modules upon the SPTS networks.** We develop two graph convolutional modules for better exploring the contextual information of different people. We construct two types of graph according to the spatial coordinates of different people. The graph for the Teacher Network is built based on the one-hot encoding vector F_{oh}, while the graph for the Student Network is constructed according to the extracted feature G from the input tracklets. The two graphs are sent into two graph convolutional modules to pass messages of different nodes. The output features are then fed into the two attention modules of the SPTS networks, respectively.

For the Student Network, we feed $G^t = \{g_1^t, g_2^t, ..., g_N^t\}$, the features of N people at the time stamp t, into the graph convolutional layer:

$$Z_{student}^t = A^t G^t W_{student}, \tag{13}$$

where A^t is calculated based on the tracklets of the tth frame. We also perform instance-normalization [52] and non-linear activation (ReLU) on the output feature $Z_{student}^t$ before it is sent into the next layer. We stack three graph convolutional layers for the Student Network, as the input G^t lies in a high-dimension space. The G^t at different time stamps t share the same parameter $W_{student}$, we concatenate $Z_{student}^t$ from 1 to T as $Z_{student} = (Z_{student}^1, ..., Z_{student}^T)$, and then sent $Z_{student}$ into the attention module of the Student Network. The effects of the number of graph convolutional layer will be explored in the Experiments section.

D. Discussions

We discuss the difference of our methods with other two categories of DNN-based methods in this subsection.

The first category, such as HTDM [2] and its variants [3] shown in Fig. 4(a), mainly focus on the appearance domain. They first learn features of individual person with an LSTM network, then aggregate them into group representations with a function f_1, and finally recognize the activity based on the group representations with another LSTM network. The labels of individual actions Y and group activity z were respectively used to supervise the training process of the first and second LSTM networks. But the corresponding relationship of Y and z have not been utilized explicitly. Moreover, the function f_1 turned to be max-pooling or mean-pooling, which lacks physical meaning.

The second category, such as SBGAR [4] displayed in Fig. 4(b), focuses on the semantic domain. This method

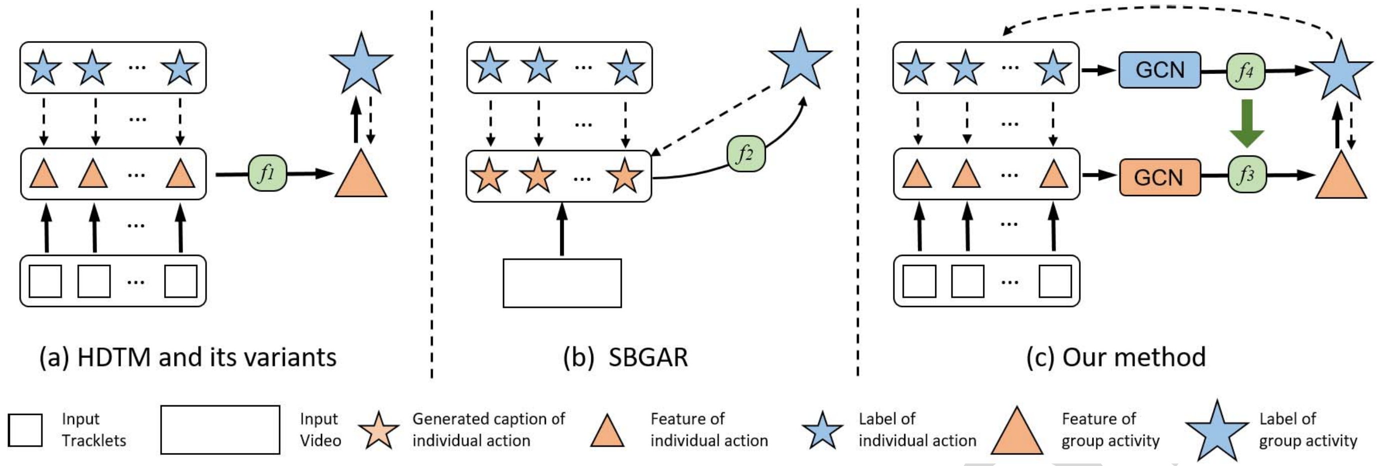


Fig. 4. Comparison of different DNN-based frameworks for group activity recognition. The solid lines, dashed lines and green arrow denote the process of forward propagation, backward propagation and semantics-preserving attention learning respectively. Method in (a) first extracts features of individual action, then aggregates them into group representations with f_1, and finally recognizes the activity based on the group representations. Approach in (b) first generates captions (*i.e.,* individual action labels) of video frames, and recognizes the activity based on these captions by f_2. Our method in (c) first employs two graph convolutional modules to capture the contextual information of features in both semantic and appearance domain. Then we learn f_4 to classify the group activity label based on the learned features in the semantic domain. Finally, we employ the attention knowledge in f_4 to guide f_3 when aggregating features in the appearance domain to make the final prediction.

directly generates the caption to describe the video frames, and utilizes the captions to classify the group activity with a function f_2. The individual actions Y were used to supervise the process of caption generation and the group activity z was utilized to supervise the learning process of f_2. However, as the group label is sensitive to the captions, the inaccurate generated captions will do harm to the final recognition results.

Different from these methods, our approach in Fig. 4(c), adequately leverage the information in the appearance domain and the semantic domain for group activity recognition. We distill the knowledge in f_4 learned in the semantic domain to guide the training process of f_3 in the appearance domain. Moreover, we have employed two graph convolutional modules to further reason the dependency of different people and enhanced the final recognition performance.

E. Exploration on Temporal Segmentation for Group Activity

Temporal segmentation (*a.k.a.* action segmentation) aims to segment actions in untrimmed videos and recognize their action labels. Although it has attracted growing attention in recent years [12], [53]–[56], few attempts on temporal segmentation for group activity have been devoted due to the scarcity of annotated datasets and complicated relationship of different people. In order to see how our method performs on this task, we have made explorations as follows.

Fig. 5 presents the illustration of incorporating our method with temporal convolutional networks (TCN) [12] for group activity segmentation. Since our method takes the tracklets of N people in T frames as input, we first divide the input video into L clips and the length of each clip is T frames. Then we employ faster-RCNN [57] to detect people in each frames, and align the cropped people in T frames according to their locations. Through this pre-process, we obtain a set of tracklets and choose N of them according to the top-N detection scores in the first frames of the clip. Then we adopt a DCNN and LSTM network to extract the features $\{F_1^l\}_{l=1}^L$ of the input tracklets, where F_1^l is a tensor with the shape of $N \times T \times d$. Here d is the summed dimension of the last fc layers in the DCNN and LSTM networks. The features of individual actions are fed into our model (SPTS Network + GCN). Finally, we concatenate the output features $\{F_2^l\}_{l=1}^L$ into a video-based feature $F_3 = concat(F_2^1, F_2^2, ..., F_2^L)$ and sent it into the TCN model to obtain the segmentation results.

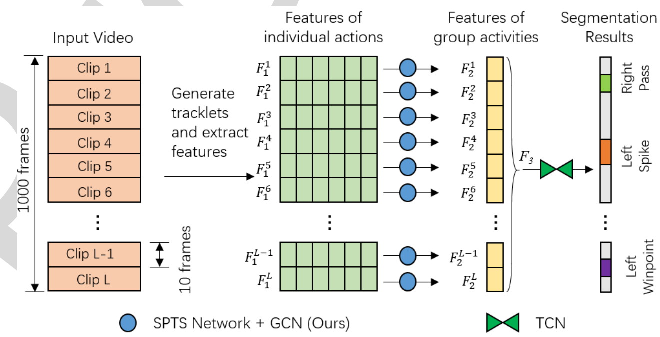


Fig. 5. Flowchart of combining our method with temporal convolutional networks (TCN) [12] for group activity segmentation. The input of the approach is an untrimmed video with L_{total} (L_{total} = 1000) frames, we first divide it into L (L=100) clips and the length of each clip is T (T =10) frames. Then we generate the tracklets based on the mask-rcnn detector and the locations of different people. Similar with the trimmed setting, the tracklets are feed into a DCNN and LSTM network to extract features of individual actions. The extracted features are sent into our model (SPTS Network + GCN) and generate the features of group activities for each clips. Finally, we concatenate these clip-based features to a video-based feature and utilize TCN model to learn the segmentation results. For the l-th clips, the F_1^l and F_2^l are corresponding to the G and $\{w_{agg}^t\}_{t=1}^T$ in Fig.2.

IV. Experiments

In this section, we conducted experiments on three datasets for group activity recognition, including volleyball dataset [59], collective activity (CA) dataset [60] and collective activity extended (CAE) dataset [8]. The experimental results and analysis are described in details as follows.

(a) (b) (c) (d)

Fig. 6. Examples of the pair-wise representative frames from three different datasets we used. For each group, the RGB-based pictures are presented on the left, while the corresponding optical flows extracted by Flownet 2.0 [58] are shown on the right. (a) Volleyball dataset. (b) Collective activity dataset. (c) Collective activity extended dataset. (d) Choi's dataset.

A. Datasets and Experiment Settings

1) Volleyball Dataset [59]: The Volleyball dataset is currently the largest dataset for group activity recognition. It contains 55 volleyball videos with 4830 annotated frames. There are 9 individual action labels (waiting, setting, digging, falling, spiking, blocking, jumping, moving and standing) and 8 group activity categories (right set, right spike, right pass, right winpoint, left winpoint, left pass, left spike and left set) in this dataset. We employ the evaluation protocol in [59] to separate the training/testing sets. We employ the metrics of Multi-class Classification Accuracy (MCA) and Mean Per Class Accuracy (MPCA) on this dataset.

2) Collective Activity (CA) Dataset [60]: The Collective Activity Dataset is a widely used benchmark for the task of group activity recognition. It comprises 44 video clips, annotated with 6 individual action classes (NA, crossing, walking, waiting, talking and queueing) and 5 group activity labels (crossing, walking, waiting, talking and queueing). There are also 8 pairwise interaction labels, which we do not utilize in this paper. We split the training and testing sets following the experimental setup in [9].

As suggested in [60] that originally presented the dataset, the "walking" activity is rather an individual action than a collective activity. To address this, we follow the experimental setup in [6], to merge the class of "walking" and "crossing" as a new class of "moving". We report the Mean Per Class Accuracy (MPCA) of the four activities on the CA dataset, which can better evaluate the performance of the classifiers.

3) Collective Activity Extended (CAE) Dataset [8]: The Collective Activity Extended Dataset contains 7 individual action labels and 6 group activities categories. It replaces the "walking" activity with other two activities of "dancing" and "jogging" in the CA Dataset. We adopted the training and testing splits used in [61] to train our models.

4) Choi's Dataset [7]: The Choi's dataset comprises 32 videos, which are annotated with 3 individual actions (walking, standing still, and running), and 6 group activities (gathering, talking, dismissal, walking together, chasing, and queueing). The dataset also provided 8 pose labels and 9 interaction labels which we did not utilize. We followed the standard experimental protocol of the 3-fold cross validation, which was adopted in [7].

5) Untrimmed Volleyball Dataset [59]: The untrimmed Volleyball dataset consists of 54 long videos of Volleyball datasets,[1] which is for temporal segmentation. The duration of each video varies from 76.76 minutes to 185.13 minutes. Since the length of these videos are too long for analysis and only numbers of temporal intervals have been annotated in [2]. We proceed them in to 837 clips according to the annotation [2], where each clips has 1000 frames. We chose this length as it is comparable with the duration of video clips in GTEA dataset [62] and 50 Salads dataset [63] evaluated by TCN [12]. We finally obtained 612 clips for training and 225 clips for testing. There are 8 group activity labels (the same with [2]) and a background label. We report the F1 score at frame level, which is computed as:

$$F1 = \frac{2 \times precision \times recall}{precision + recall}. \quad (14)$$

[1]The original volleyball dataset provided trimmed clips and the names of 55 long videos. However, the 21-th video cannot be found according to its names. Moreover, due to the changes of frame rate on YouTube, 8 videos are incorrectly aligned with the temporal annotation provided in [2]. To address this, we spent 2 days refining the annotations to ensure their correctness.

B. Implementation Details and Baselines

1) Group Activity Recognition: Our proposed methods were built on the Pytorch toolbox and implemented on a system with the Intel(R) Xeon(R) E5-2660 v4 CPU @ 2.00Ghz. We trained our model with two Nvidia GTX 1080 Ti GPUs and tested it with one GPU.

For the Teacher Network, we took the ground-truth label of each individual action as input, and the one-hot vectors were projected through an fc layer. The embedded features were weighted and summed based on different weights learned by the self-attention mechanism, which indicates the importance of different people. The aggregated features were then fed into an fc layer for classification. The Teacher Network was trained with the Adam optimization method with 16 as the batch size. And the initial learning rate was 0.003.

For the Student Network, we first finetuned VGG network [64] pretrained on ImageNet [65] to extract CNN features of the tracklets. The features of the last fc layer were fed into a LSTM network with 3000 nodes. The concatenated features of VGG and LSTM networks were then fed into an fc layer with the size of 512 to cut down the dimension. The importance of each person on each frame was generated by the attention mechanism, and the embedded features of each person were then summed by weight. The weighted features were then fed into a bidirectional LSTM network with the hidden size of 128. The output features were fed into an fc layer for classification. During the Teacher guided training process, the Student Network was optimized with Adam and the initial learning rate was 0.00003. As for ratio of different parts of losses, we set $\lambda_1 = \lambda_2 = 1$. The batch size was set to be 16.

In order to better explore the motion information of the video and inspired by the success of two-stream network architecture [18], we computed the optical flow between two adjacent video frames using Flownet 2.0 [58]. We extracted the DCNN and LSTM features of optical flow tracklets, and concatenated them with the features of the original RGB tracklets before the attention module of the Student Network.

We report the performance of the following baseline methods and different versions of our approach:

- HDTM [2]: A hierarchical framework with two LSTM models. The first LSTM network took the features extracted from the tracklets of each person as input, and was trained with the supervision of the individual action label. The input of the second LSTM network was the aggregation of features learned by the first LSTM, and was trained with the supervision of the group activity label.
- Ours-teacher*: The Teacher Network directly took the ground-truth labels of the individual actions as input during both training and testing phases. Hence, it is not fair to directly compare the performance of Teacher Network with other methods, which are inaccessible to the ground-truth labels of the individual actions during testing phase. We report the performance of Ours-teacher* only for reference.
- Ours-teacher: During the training phase, we used the ground-truth label of each individual action as input to train the Teacher Network. During the testing stage, we used the individual action label learned from the first LSTM of HDTM to predict the final group activity label.
- Ours$_{-SA}$ (self-attention): An original model of our Student Network, which can be regarded as adding a self-attention module upon the HDTM [2].
- Ours$_{-SPA}$ (semantics-preserving attention): A version of model which employed the attention knowledge in Teacher Network to help the training of Student Network.
- Ours$_{-SPA+KD}$ (knowledge distillation): A model of combining the knowledge distillation loss [34] with Ours$_{-SPA}$.
- Ours†-x: Models of combining the optical flow input based on the original Ours-x.
- Ours-teacher* + GCN: Building the graph convolutional module upon the Teacher Network.
- Ours+GCN$_{-SA}$, Ours+GCN$_{-SPA+KD}$, Ours† +GCN$_{-SA}$ and Ours† +GCN$_{-SPA+KD}$: Models of equipping the graph convolutional module with Ours$_{-SA}$, Ours$_{-SPA+KD}$, Ours$^\dagger_{-SA}$ and Ours$^\dagger_{-SPA+KD}$.

2) Temporal Segmentation for Group Activity: During experiments, we first pretrained our model on the trimmed Volleyball dataset, and finetuned it on the untrimmed dataset to extract features. We report the segmentation results of comparing methods in two categories: image-level methods and person-level methods. *The first category* consists of two methods, which took the whole images as input directly: (1) VGG16 [64]: We employed VGG16 network pretrained on ImageNet [65], and finetuned it on the training set of untrimmed Volleyball to predict the frame-level labels. (2) TCN [12]: We used the features of the fc7 layer in VGG16 to train the TCN models. *The second category* comprises three approaches, which were based on the tracklets of different persons: TCN$_{-SA}$, TCN$_{-SPA+KD}$, TCN-GCN$_{-SPA+KD}$. They denote using the methods Ours$_{-SA}$, Ours$_{-SPA+KD}$, Ours-GCN$_{-SPA+KD}$ for feature extraction respectively.

TABLE I

COMPARISON OF THE GROUP ACTIVITY RECOGNITION ACCURACY (%) ON THE VOLLEYBALL DATASET. † DENOTES THAT THE MODEL TAKES BOTH RGB IMAGES AND OPTICAL FLOWS AS INPUTS

Method	MCA	MPCA
CERN-2 [3]	83.3	83.6
SSU [5]	89.9	–
SRNN [66]	83.5	–
RCRG [20]	89.5	–
Ours-teacher*	88.3	84.4
Ours-teacher* + GCN	92.3	90.7
Ours-teacher	69.3	66.8
Baseline-HDTM [2]	86.8	85.8
Ours $_{-SA}$	87.1	86.1
Ours $_{-SPA}$	89.3	89.2
Ours $_{-SPA+KD}$	89.3	89.0
Ours† $_{-SA}$	87.7	87.0
Ours† $_{-SPA}$	89.6	89.5
Ours† $_{-SPA+KD}$	90.7	90.0
Ours + GCN $_{-SA}$	89.2	88.8
Ours + GCN $_{-SPA+KD}$	90.4	89.3
Ours† + GCN $_{-SA}$	90.4	90.5
Ours† + GCN $_{-SPA+KD}$	**91.2**	**91.4**

C. Results on the Volleyball Dataset

We first evaluate our proposed methods on the Volleyball dataset. We follow [2] to separate players into two groups on the left and right, and extend the individual action labels to 18 categories (*e.g.,* "left standing", "right waiting", etc.) according to their spatial coordinates.

1) Comparison With the State-of-the-Arts: Table I presents the comparison performance with different approaches. We observe that our final model (Ours† + GCN$_{-SPA+KD}$) achieves 91.2% MCA and 91.4% MPCA, outperforming existing state-of-the-art methods for group activity recognition.

2) Analysis on the SPTS Networks: Here we analyze our semantics-preserving learning scheme. Compared with the 0.3% (MCA and MPCA) improvement by the self-attention scheme over the baseline method, our attention-guided approach achieves 2.5% (MCA) and 3.2% (MPCA) improvement, which demonstrates the effectiveness of our proposed method. We also discover that, combining with the optical flow can lead to a slight improvement on this dataset. While besides, Our-teacher*, which takes the ground-truth of individual actions as the testing inputs of the Teacher Network, reaches performance of 88.3% MCA, Our-teacher, which utilizes the predicted individual actions as the testing inputs, only attains 69.3% MCA. This is because, the Teacher Network is sensitive to the inputs and the incorrected predicted individual actions will greatly harm the performance of the final recognition.

We also show several visualization results of the learned attention in Fig. 7. The group activity label of Fig. 7(a) is "left spike". For the self-attention model of the Student Network, the model most likely focuses on those people wearing different clothes in a group, *e.g.,* the white person (SA:60) in the black team, and the yellow person (SA:62) in the white team. However, these people are not exactly key people for recognizing the group activity. When we employ the attention model of Teacher Network, we can focus on those words, which are essentially important in the semantic

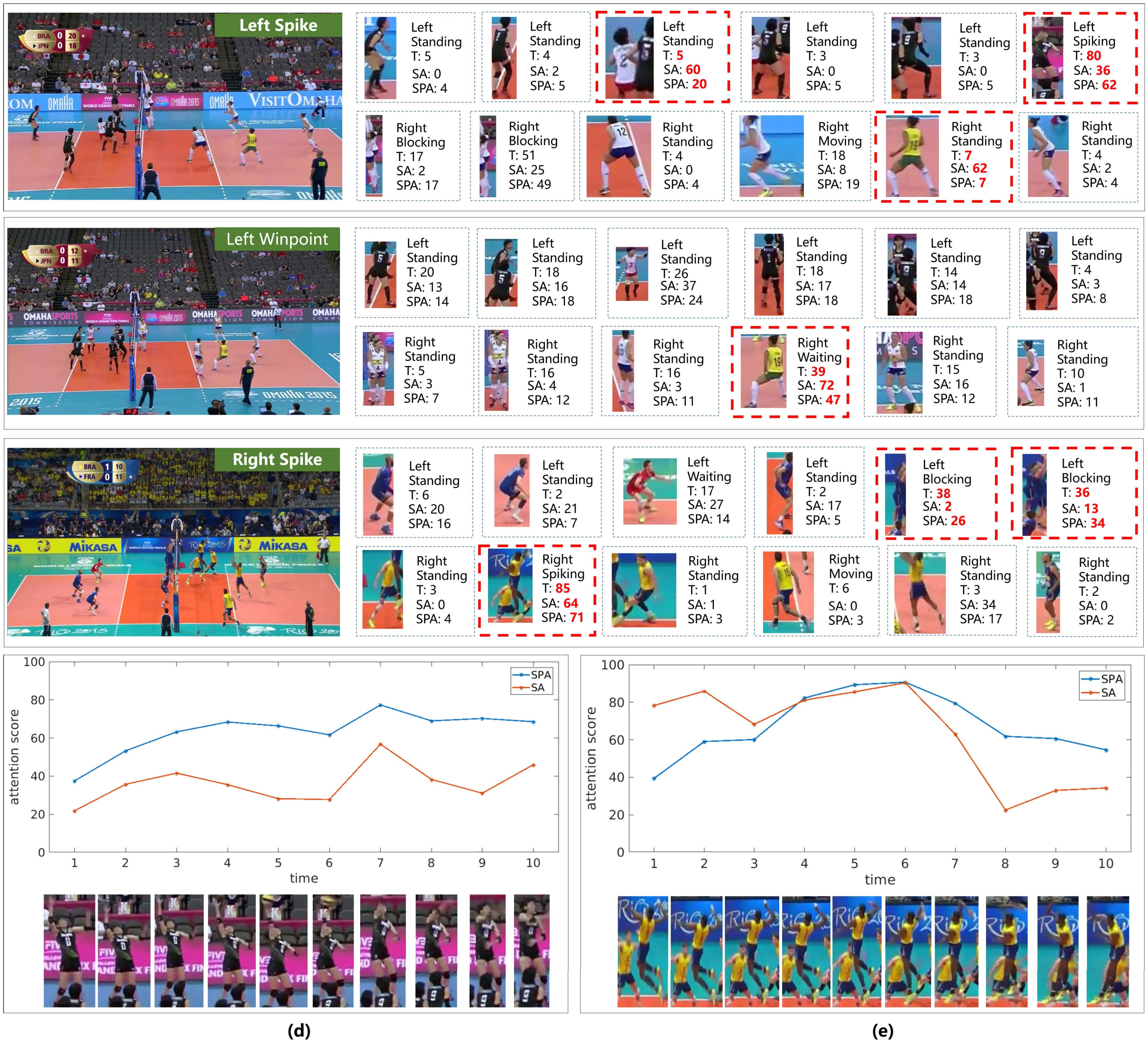


Fig. 7. Visualization of the learned attention on the Volleyball dataset. In (a)(b)(c), for each video clip, we show the representative frame on the left, while the cropped people are shown on the right. In each dash box, we display the labels of individual actions and three types of attention score: T (Teacher Network), SA (Student Network with self-attention scheme) and SPA (Student Network with semantics-preserving attention method). The SA and SPA scores in (a)(b)(c) are averaged scores over a clips (10 frames). In (d)(e), we present the attention scores and the corresponding people in temporal domain.

domain, *e.g.,* the spiking (T:80), and the blocking (T:51). And after employing our SPTS networks, we will transfer this attention knowledge from the semantic domain to the appearance domain, and guide the Student Network to focus on the "left spiking" person (SPA:62), who contributes most to recognizing the final activity. The group activity label of Fig. 7(b) is "left winpoint", where there is no special people for recognizing this activity. However, the self-attention scheme assign the highest score to the yellow person (SA:72), which does not carry key information. After employing the SPTS networks, the score of this person is decreased to 47, and extra attention is allocated to other people. Fig. 7(c) illustrates similar results to Fig. 7(a).

We further present the learned attention scores on temporal domain in Fig. 7(d) and Fig. 7(e). For the "spiking" people in volleyball dataset, our SPA scores (blue ones) go up to climaxes when the players wave their hands to spike the ball, which assigns more attention to the discriminative frames.

3) Analysis on the Graph Convolutional Modules: As shown in Table I, when applying the graph convolutional modules, the Teacher Network achieves 4.0% and 6.3% improvement on the MCA and MPCA metrics respectively. For the Student Network, Ours Ours† + GCN$_{-SA}$ and Ours† + GCN$_{-SPA+KD}$ attain 2.7% and 0.5% improvement on MCA, and 3.5% and 1.4% improvements on MPCA, which consistently demonstrates the effectiveness of the graph convolutional modules.

Moreover, we have conducted experiments on adopting different layers for the Teacher Network and Student Network. As presented in Table II, the peaks of the Teacher Network and

TABLE II

COMPARISON OF THE GROUP ACTIVITY RECOGNITION ACCURACY (%) OF DIFFERENT NUMBER OF GRAPH CONVOLUTIONAL LAYERS ON THE VOLLEYBALL DATASET

Number of Graph Convolutional Layers	1	3	5	7
Ours-teacher* + GCN (semantic domain)	**92.3**	91.3	90.9	90.4
Ours† + GCN$_{-SA}$ (appearance domain)	89.6	**90.4**	90.3	90.2

TABLE III

COMPARISON OF THE GROUP ACTIVITY RECOGNITION ACCURACY (%) ON THE CA DATASET. † IS DEFINED IN THE CAPTION OF TABLE I

Method	MPCA
Cardinality kernel [1]	88.3
CERN-2 [3]	88.3
RMIC [6]	89.4
SBGAR [4]	89.9
MTCAR [7]	90.8
Ours-teacher*	97.6
Ours-teacher* + GCN	97.6
Ours-teacher	88.2
baseline-HDTM [2]	89.7
Ours $_{-SA}$	91.5
Ours $_{-SPA}$	92.3
Ours $_{-SPA+KD}$	92.5
Ours† $_{-SA}$	94.3
Ours† $_{-SPA}$	95.6
Ours† $_{-SPA+KD}$	95.7
Ours + GCN $_{-SA}$	91.8
Ours + GCN $_{-SPA+KD}$	92.9
Ours† + GCN $_{-SA}$	95.4
Ours† + GCN $_{-SPA+KD}$	**95.8**

Student Network appear at one layer and three layers respectively. This is because, the dimension of input feature to the Teacher Network is relatively low and one graph convolutional layer is proper. For the Student Network, the dimension of input feature is much higher, thus deeper structure is needed to achieve a better result.

D. Results on the CA Dataset

1) Comparison With the State-of-the-Arts: Table III shows the comparison with different methods on the CA dataset. The MPCA results of other approaches are computed based on the original confusion matrices in [1]–[4], [6], [7]. We observe that, our final model (Ours† + GCN$_{-SPA+KD}$) achieves 95.8% MPCA, outperforming the state-of-the-art [7] by 5.0%. Moreover, our method have improved the baseline method HDTM [2] by 6.0%. Fig. 9 presents the confusion matrices of the baseline methods and our SPTS networks. It is clear that SPTS networks attain superior results, especially for distinguishing the activity of "moving" and "waiting". Besides, compared with SBGAR and Ours-teacher, which directly utilized the semantic information to predict the final labels, our method achieves 5.9% and 7.6% improvement, which demonstrates its effectiveness. Objectively speaking, we should own the major contribution to the combination of the optical flow, which explicitly captures the motion information of the scene. Based on this, our two semantics-preserving learning method and graph convolutional module have further enhanced the recognition performance, which will be discussed as follow.

2) Analysis on the SPTS Networks: From Table III, our attention-guided method brings 1.0%, 1.4% and 0.4% improvements on the self-attention scheme of Ours$_{-SA}$, Ours$^\dagger_{-SA}$ and Ours+GCN$^\dagger_{-SA}$. We notice that these improvements are less significant than those on the Volleyball dataset. This is because the setting of the CA dataset is to assign what the major people are doing to the label of group activity. Hence, attention model is not so important.

We also show the visualization of the learned attention in Fig. 8. As shown in Fig. 8(a), the group activity label is "waiting", hence the Teacher Network allocates more attention to the words "waiting" (29) and less attention to the word "moving". Guided by this information, the Student Network decreases the attention (from 22 to 17) of the "moving" person, which can be regarded as a noise for recognizing the group activity. For Fig. 8(b), the group activity is "moving", and it is reasonable that the Teacher Network allocates averaged score to the three individual words "moving". Taught by this attention knowledge, the Student Network increases the attention of the top person from 20 to 27, and decreases the attention of the right person from 43 to 37, so that the information of three people can be utilized equally.

The temporal attention scores are shown in Fig. 8(c) and Fig. 8(d). For the "spiking" people in volleyball dataset, our SPA scores (blue ones) go up to climaxes when the players wave their hands to spike the ball, which assigns more attention to the discriminative frames. For the "waiting" and "moving" people in CA dataset, the learned SPA scores vary little over time because there is no part of particular significance during these actions.

3) Analysis on the Graph Convolutional Modules: When we apply graph convolutional modules to the SPTS networks, the MPCA increases 1.1% and 0.1% over Ours$^\dagger_{-SA}$ and Ours$^\dagger_{-SPA+KD}$ respectively, which also shows its effectiveness. However, we observe that the improvements are not novel as the results on the volleyball dataset. The reason is that the volleyball dataset is the currently largest dataset for group activity recognition, while the CA dataset is relatively small. Since the graph convolutional module is a data-driven model, more training data can bring more benefits.

E. Results on the CAE Dataset

We further conducted experiments on the CAE dataset. Table IV presents the comparison with different methods, where our final model reaches a performance of 98.1%, outperforming the existing state-of-the-art methods. The self-attention scheme achieves 95.0% and 95.9% recognition accuracy on the RGB inputs and combining optical flows respectively, where we obtains 0.9% and 1.7% improvements when applying our SPTS network. Moreover, Ours-teacher* +GCN, Ours† +GCN$_{-SA}$ and Ours† +GCN$_{-SPA+KD}$ obtained 1.3%, 0.9% and 0.5% improvements benefiting from the graph convolutional modules, which further shows the effectiveness of the proposed approaches.

Fig. 9 presents the comparison of confusion matrices on the baseline method and our final model. For the baseline method, "waiting" is sometimes confused with the activity "crossing", and "dancing" is likely to be misclassified as "jogging". When applying our method, we clearly show the

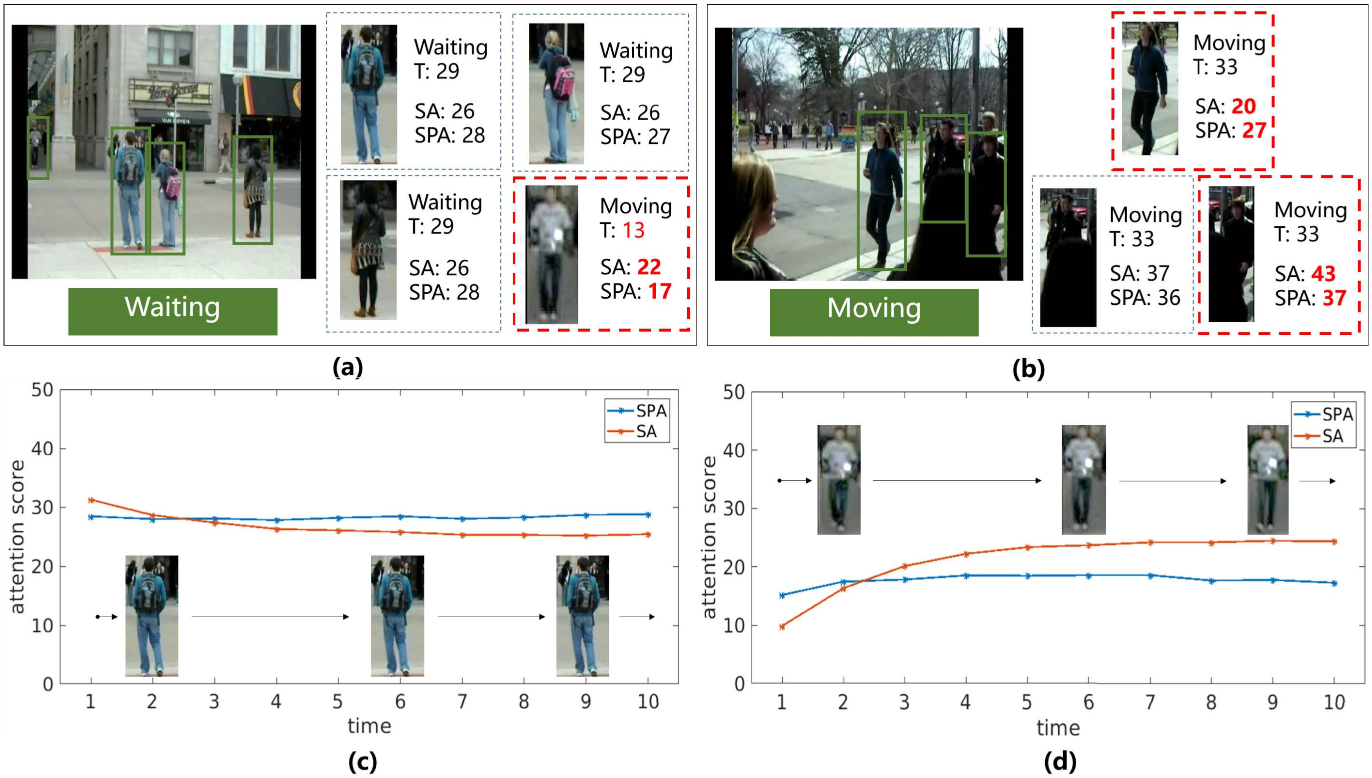


Fig. 8. Visualization of the learned attention on the CA dataset. The definitions of T, SA and SPA are the same with those in Fig. 8.

TABLE IV

COMPARISON OF THE GROUP ACTIVITY RECOGNITION ACCURACY (%) ON THE COLLECTIVE ACTIVITY EXTENDED DATASET. † IS DEFINED IN THE CAPTION OF TABLE I

Method	Accuracy
CRF+CNN [61]	86.8
Structural SVM + CNN [61]	87.3
Structure Inference Machines [61]	90.2
Image Classification Model [11]	92.3
Person Classification Model [11]	95.1
Latent Embeddings Model [11]	97.9
Ours-teacher*	97.8
Ours-teacher* + GCN	99.1
Ours-teacher	96.0
baseline-HDTM [2]	94.2
Ours $_{-SA}$	95.0
Ours $_{-SPA}$	95.8
Ours $_{-SPA+KD}$	95.9
Ours† $_{-SA}$	95.9
Ours† $_{-SPA}$	97.2
Ours† $_{-SPA+KD}$	97.6
Ours + GCN $_{-SA}$	95.6
Ours + GCN $_{-SPA+KD}$	96.2
Ours† + GCN $_{-SA}$	96.8
Ours† + GCN $_{-SPA+KD}$	**98.1**

advantages on discriminating these activities and obtain the promising recognition results.

F. Results on the Choi's Dataset

Table V presents the experimental results. In this dataset, our final model Ours† + GCN$_{-SPA+KD}$ achieves 78.1% accuracy, which is comparable with existing methods [2], [7], [60]. Objectively speaking, the performance of our method is not novel as those in the volleyball [59], CA [60] and CAE [8] datasets, and the reasons are two folds: (1) The methods [7], [60] utilize the pose labels and interaction labels, which are not used in our methods. (2) Our methods are data-driven based, while the methods [7], [60] use hand-crafted features. So they have more advantages on the Choi's dataset, which is the smallest compared with the other three datasets. Besides, we observe that combining optical flow can bring a large improvement in this dataset. This is because the individual action labels of this dataset are "walking", "standing still", and "running", so the features obtained with the input of optical flow have much more discriminative power. Moreover, we find the GCN and semantics-preserving attention scheme can further lead to improvements, which demonstrates the effectiveness of our proposed approaches.

G. Results on the Untrimmed Volleyball Dataset

We evaluate our method for action segmentation on this dataset and Table VI presents the experimental results. First, in the image-level category, we find that utilizing TCN can improve the performance over the frame level method, which demonstrates the effectiveness of TCN in modelling temporal dependency. Second, the person-level methods perform better than the whole frame based methods. This is because the later ones can better focus on the action performer, which provides more discriminative power of action. Finally, we observe that adopting our semantic-preserving attention and GCN model can further improve the performance, which indicates the discriminative power of features learned by our proposed method. We also show several action segmentation results in supplementary material for visualization.

(a) (b) (c) (d)

Fig. 9. Comparison of Confusion Matrices on CA [60] and CAE dataset [8]. † is defined in the caption of Table I. For the CA datset, we merge the class of *Walking* and *Crossing* as the same class of *Moving* as suggested in [6]. (a) Baseline - HDTM. (b) Ours† + GCN$_{-SPA+KD}$. (c) Baseline - HDTM. (d) Ours† + GCN$_{-SPA+KD}$.

TABLE V

COMPARISON OF THE GROUP ACTIVITY RECOGNITION ACCURACY (%) ON THE CHOI'S DATASET. † DENOTES THAT THE MODEL TAKES BOTH RGB IMAGES AND OPTICAL FLOWS AS INPUTS. ‡ AND $^{\wr}$ REPRESENT THAT THE EXTRA POSE AND INTERACTION ANNOTATIONS ARE FURTHER USED

Method	Accuracy
STL‡ [60]	77.4
MTCAR$^{\ddagger\ \wr}$ [7]	83.0
Ours-teacher*	79.3
Ours-teacher* + GCN	79.8
Ours-teacher	70.2
baseline-HDTM [2]	57.0
Ours $_{-SA}$	57.3
Ours $_{-SPA}$	58.3
Ours $_{-SPA+KD}$	58.5
Ours† $_{-SA}$	76.2
Ours† $_{-SPA}$	77.3
Ours† $_{-SPA+KD}$	77.5
Ours + GCN $_{-SA}$	57.9
Ours + GCN $_{-SPA+KD}$	58.6
Ours† + GCN $_{-SA}$	76.8
Ours† + GCN $_{-SPA+KD}$	**78.1**

TABLE VI

COMPARISON OF THE GROUP ACTIVITY SEGMENTATION ACCURACY (%) ON THE UNTRIMMED VOLLEYBALL DATASET

Method	Category	F1 score
VGG16 [64]	Image level	41.74
TCN [12]	Image level	45.17
TCN$_{-SA}$	Person level	56.06
TCN$_{-SPA+KD}$	Person level	57.59
TCN-GCN$_{-SPA+KD}$	Person level	**59.49**

TABLE VII

ANALYSIS ON THE INFLUENCE OF INFERIOR CAPTIONS ON THE SPLIT2 OF CHOI'S DATASET

Method	Accuracy (%)	Influence (%)
Teacher*	79.2	-
Teacher*-new	58.5	-20.7 (Stage 2)
Student	74.4	-
Student-new	60.8	-13.6 (Stage 1)
Student-new$_{-SPA+KD}$	59.1	-1.7 (Stage 3)

H. Analysis on the Influence of Caption Quality

Captions, which are a sets of individual words of actions in this paper, are utilized during three stages in our method:

Stage 1: Finetuning the DCNN and LSTM network, and extracting the features of individual actions.

Stage 2: Training the Teacher network.

Stage 3: Guiding the training process of the Student network.

The Stage 1 is a common process in most deep-learning based methods [2], [3], [6] and the Stage 2 is an intermediate process of our method. The Stage 3 is what we should pay more attention to, as it is the core step of our method and directly influences the final recognition result.

In order to further analyze the influence of the caption quality, we conducted the experiments on the split2 of Choi's dataset. We randomly selected 50% captions in the training sets and assigned random single action labels to them. In this way, the caption quality will become inferior.

Table VII presents the comparison between results on the original setting (Teacher*, Student) and the new setting (Teacher*-new, Student-new, Student-new$_{-SPA+KD}$). We observe that the captions will heavily influence Stage 1 and Stage 2 (The accuracy drop from 74.4% (Student) to 60.8% (Student-new) because the extracted features became inferior). In comparison, the decrease caused by our method (Stage 3) is slight, which shows its robustness to the low quality captions. The intuition of our method's robustness lies in two folds. First, as the Teacher Network is trained with noisy input labels, the semantics-preserving attention would tend to learn to deal with such noise. Second, knowledge distillation from Teacher Network provides additional soft labels for training Student Network, which will inevitably cause the decrease of the Student Network if the Teacher Network is noisy. But with ground-truth group activity label as direct supervision, this decrease in performance is relieved and won't hurt the final result too much.

I. Analysis on the Computational Time

There are some real-world applications for group activity recognition, *e.g.*, sports video analysis and traffic surveillance, which require recognizing the activity in real time. Therefore, we are motivated to investigate the time cost of our approach. Table VIII shows the computational time

TABLE VIII

COMPUTATIONAL TIME ANALYSIS ON THE VOLLEYBALL DATASET. † IS DEFINED IN THE CAPTION OF TABLE I

Training Process (Based on Dataset)	**Time (h)**
Train Teacher Network	0.36
Train DCNN and LSTM for RGB Images	11.50
Extract Features for RGB Images	0.46
Train GCN, Attention Module and BLSTM	1.00
Compute Optical Flow	61.48
Train DCNN and LSTM for Optical Flow	11.50
Extract Features(OF)	0.46
†Train GCN, Attention Module and BLSTM	1.16
Testing Process (Based on Single Frame)	**Time (ms)**
Extract Features for RGB Images	8.01 × 12 (people)
Activity Recognition (10 Frames)	13.93
Compute Optical Flow	434.65
Extract Features for Optical Flow	8.01 × 12 (people)
†Activity Recognition (10 Frames)	26.45

TABLE IX

COMPARISON OF THE COMPUTATIONAL TIME (s) OF DIFFERENT METHODS ON THE VOLLEYBALL DATASET. THE RESULTS ARE BASED ON A CLIP WITH 10 FRAMES. † DENOTES THAT THE RESULTS ARE BASED ON THE INPUTS WITH RGB IMAGES AND OPTICAL FLOWS

SBGAR [4]	HDTM [2]	Ours$_{-SPA+KD}$	Ours+GCN$_{-SPA+KD}$
-	0.950	0.968	0.983
1.0966†	6.207†	6.227†	6.295†

analysis of our method. The training data were based on one run while the testing data were averaged over five runs on the Volleyball dataset. We did not include the time to detect individual players as we utilized the off-the-shelf tracklets provided by [2].

Without utilizing optical flow, it required about $0.36 + 11.50 + 0.46 + 1.00 = 13.32h$ to train the SPTS + GCN. For a video clip with 10 frames, it took $10 \times (8.01 \times 12) + 13.93 = 983.14ms(0.983sec)$ to predict the group activity label. Moreover, training the Teacher Network was about 0.36 h, only 2.70% of the entire training time.

When combining the optical flow, the training phase lasted about $0.36 + 61.48 + 2 \times (11.5 + 0.46) + 1.16 = 86.92h$ while predicting the label of a video clip took $10 \times (434.65 + 8.01 \times 12 \times 2) + 26.45 = 6295.35ms(6.295sec)$. The reason why combining the optical flow is relatively slow is that, we employed the Flownet 2.0 model with the best performance and highest computational time cost in [58].

Table IX presents the computational time comparison with state-of-the-arts. The result of SBGAR is reported from [4], and the others are based on our implementation. On one hand, we find that when combining optical flow, the SBGAR is more efficient and the reason are two folds. (1) The optical flow computation time of SBGAR on a single image is much faster than ours (0.022s vs 0.435s) due to the difference between the methods for calculating optical flow. (2) SBGAR directly takes the whole frames as inputs while our method is based on the a set of tracklets. On the other hand, compared with the baseline approach HDTM [2], the increased time cost of Ours$_{-SPA+KD}$ and Ours+GCN$_{-SPA+KD}$ are slight, which illustrates the efficiency of our methods.

V. FUTURE WORKS

There are some interesting directions for future works:

1) Designing different formulations of GCN for group activity recognition. For example, one is to use a single graph with temporal information. Concretely, we can first perform temporal pooling (e.g., max-pooling or attention-pooling) over the features of individual person and adjacency matrices of different frames, and then construct a single graph and feed it into the GCN model. Another one, which is inspired by [47], is to build a spatial-temporal graph. In this way, features of different people in different frames will be organized in a unified graph, and the final bidirectional LSTM layer in our model can be removed. However, as the scale of the spatial-temporal graph is much larger, other efforts on efficient modeling need to be devoted.
2) Transferring knowledge in the graph between the Student and Teacher network.[2]
3) Employing our method for the tasks like image/video caption or visual question answering (VQA), which lie in the interaction area of the natural language domain and computer vision domain.
4) Exploring different variants in [58] and other optical flow estimation algorithms to achieve a better trade-off between the accuracy and efficiency.

VI. CONCLUSIONS

In this paper, we have presented a Semantics-Preserving Teacher-Student (SPTS) architecture for group activity recognition in videos. The proposed method has explored the attention knowledge in the semantic domain and employed it to guide the learning process in appearance domain, which explicitly exploits the attention information of the group people. Moreover, we have strengthened our SPTS by incorporating with two graph convolutional modules to reason the relationship among different people. Furthermore, we have extended our approach on action segmentation task for untrimmed videos and demonstrated its effectiveness. Extensive experimental results on four datasets have shown the superior performance of our proposed method in comparison with the state-of-the-arts.

ACKNOWLEDGEMENT

The authors would like to thank Peiyang Li, Danyang Zhang, Yu Zheng, Simin Wang, Yongming Rao, and Tianmin Shu for their generous help.

REFERENCES

[1] H. Hajimirsadeghi, W. Yan, A. Vahdat, and G. Mori, "Visual recognition by counting instances: A multi-instance cardinality potential kernel," in *Proc. CVPR*, Jun. 2015, pp. 2596–2605.

[2] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proc. CVPR*, Jun. 2016, pp. 1971–1980.

[2] We have made some attempts on this direction, see supplementary material for details.

[3] T. Shu, S. Todorovic, and S.-C. Zhu, “CERN: Confidence-energy recurrent network for group activity recognition,” in *Proc. CVPR*, Jul. 2017, pp. 4255–4263.

[4] X. Li and M. C. Chuah, “SBGAR: Semantics based group activity recognition,” in *Proc. ICCV*, Oct. 2017, pp. 2895–2904.

[5] T. M. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese, “Social scene understanding: End-to-end multi-person action localization and collective activity recognition,” in *Proc. CVPR*, Jul. 2017, pp. 3425–3434.

[6] M. Wang, B. Ni, and X. Yang, “Recurrent modeling of interaction context for collective activity recognition,” in *Proc. CVPR*, Jul. 2017, pp. 7408–7416.

[7] W. Choi and S. Savarese, “A unified framework for multi-target tracking and collective activity recognition,” in *Proc. ECCV*, 2012, pp. 215–230.

[8] W. Choi, K. Shahid, and S. Savarese, “Learning context for collective activity recognition,” in *Proc. CVPR*, Jun. 2011, pp. 3273–3280.

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

[10] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S.-C. Zhu, “Joint inference of groups, events and human roles in aerial videos,” in *Proc. CVPR*, Jun. 2015, pp. 4576–4584.

[11] Y. Tang, P. Zhang, J.-F. Hu, and W.-S. Zheng, “Latent embeddings for collective activity recognition,” in *Proc. AVSS*, Aug./Sep. 2017, pp. 1–6.

[12] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, “Temporal convolutional networks for action segmentation and detection,” in *Proc. CVPR*, Jul. 2017, pp. 1003–1012.

[13] Y. Tang, Z. Wang, P. Li, J. Lu, M. Yang, and J. Zhou, “Mining semantics-preserving attention for group activity recognition,” in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 1283–1291.

[14] H. Wang, H. Kläser, A. Schmid, and C.-L. Liu, “Action recognition by dense trajectories,” in *Proc. CVPR*, Jun. 2011, pp. 3169–3176.

[15] W. Du, Y. Wang, and Y. Qiao, “Recurrent spatial-temporal attention network for action recognition in videos,” *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1347–1360, Mar. 2018.

[16] S. Ji, W. Xu, M. Yang, and K. Yu, “3D convolutional neural networks for human action recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[17] W. Hu, B. Wu, P. Wang, C. Yuan, Y. Li, and S. J. Maybank, “Context-dependent random walk graph kernels and tree pattern graph matching kernels with applications to action recognition,” *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5060–5075, Oct. 2018.

[18] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Proc. NIPS*, 2014, pp. 568–576.

[19] X. Chang, W.-S. Zheng, and J. Zhang, “Learning person-person interaction in collective activity recognition,” *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1905–1918, Jun. 2015.

[20] M. S. Ibrahim and G. Mori, “Hierarchical relational networks for group activity recognition and retrieval,” in *Proc. ECCV*, 2018, pp. 721–736.

[21] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, “Modeling visual attention via selective tuning,” *Artif. Intell.*, vol. 78, nos. 1–2, pp. 507–545, 1995.

[22] R. A. Rensink, “The dynamic representation of scenes,” *Vis. Cognition*, vol. 7, nos. 1–3, pp. 17–42, Oct. 2010.

[23] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proc. ICLR*, 2014, pp. 1–15.

[24] A. Vaswani *et al.*, “Attention is all you need,” in *Proc. NIPS*, 2017, pp. 6000–6010.

[25] J. Yang *et al.*, “Neural aggregation network for video face recognition,” in *Proc. CVPR*, Jul. 2017, pp. 5216–5225.

[26] Y. Rao, J. Lu, and J. Zhou, “Attention-aware deep reinforcement learning for video face recognition,” in *Proc. ICCV*, Oct. 2017, pp. 3951–3960.

[27] A. Haque, A. Alahi, and L. Fei-Fei, “Recurrent attention models for depth-based person identification,” in *Proc. CVPR*, Jun. 2016, pp. 1229–1238.

[28] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Is object localization for free?—Weakly-supervised learning with convolutional neural networks,” in *Proc. CVPR*, Jun. 2015, pp. 685–694.

[29] S. Zagoruyko and N. Komodakis, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” in *Proc. ICLR*, 2017, PP. 1–13.

[30] Z. Guo, L. Gao, J. Song, X. Xu, J. Shao, and H. T. Shen, “Attention-based LSTM with semantic consistency for videos captioning,” in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 357–361.

[31] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, “Stacked attention networks for image question answering,” in *Proc. CVPR*, Jun. 2016, pp. 21–29.

[32] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, “Skeleton-based human action recognition with global context-aware attention LSTM networks,” *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586–1599, Apr. 2018,

[33] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, “An end-to-end spatio-temporal attention model for human action recognition from skeleton data,” in *Proc. AAAI*, 2017, pp. 4263–4270.

[34] G. E. Hinton, O. Vinyals and J. Dean, “Distilling the knowledge in a neural network,” in *Proc. NIPSW*, 2014.

[35] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “FitNets: Hints for thin deep nets,” in *Proc. ICLR*, 2014, pp. 1–13.

[36] J. Yim, D. Joo, J. Bae, and J. Kim, “A gift from knowledge distillation: Fast optimization, network minimization and transfer learning,” in *Proc. CVPR*, Jul. 2017, pp. 7130–7138.

[37] T. Chen, I. J. Goodfellow, and J. Shlens, “Net2Net: Accelerating learning via knowledge transfer,” in *Proc. ICLR*, 2015, pp. 1–12.

[38] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, “The graph neural network model,” *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

[39] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *Proc. ICLR*, 2017, pp. 1–14.

[40] M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” in *Proc. NIPS*, 2016, pp. 3844–3852.

[41] J. Bruna, W. Zaremba, A. Szlam, and Y. Lecun, “Spectral networks and locally connected networks on graphs,” in *Proc. ICLR*, 2014, pp. 1–14.

[42] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, “The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains,” *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.

[43] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, “3D graph neural networks for RGBD semantic segmentation,” in *Proc. ICCV*, Oct. 2017, pp. 5209–5218.

[44] X. Chen, L. Li, L. Fei-Fei, and A. Gupta, “Iterative visual reasoning beyond convolutions,” in *Proc. CVPR*, Jun. 2018, pp. 7239–7248.

[45] X. Wang, Y. Ye, and A. Gupta, “Zero-shot recognition via semantic embeddings and knowledge graphs,” in *Proc. CVPR*, Jun. 2018, pp. 6857–6866.

[46] Z. Wang, T. Chen, J. Ren, W. Yu, H. Cheng, and L. Lin, “Deep reasoning with knowledge graph for social relationship understanding,” in *Proc. IJCAI*, 2018, pp. 1021–1028.

[47] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *Proc. AAAI*, 2018, pp. 7444–7452.

[48] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, “Deep progressive reinforcement learning for skeleton-based action recognition,” in *Proc. CVPR*, Jun. 2018, pp. 5323–5332.

[49] C. Li, Z. Cui, W. Zheng, C. Xu, and J. Yang, “Spatio-temporal graph convolution for skeleton based action recognition,” in *Proc. AAAI*, 2018, pp. 3482–3489.

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proc. NIPS*, 2012, pp. 84–90.

[51] P. J. Werbos, “Backpropagation through time: What it does and how to do it,” *Proc. IEEE*, vol. 78, no. 10, pp. 1550–1560, Oct. 1990.

[52] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Instance normalization: The missing ingredient for fast stylization,” *CoRR*, abs/1607.08022, Jul. 2016.

[53] X. Xiang, Y. Tian, A. Reiter, G. D. Hager, and T. D. Tran, “S3D: Stacking segmental P3D for action quality assessment,” in *Proc. ICIP*, Oct. 2018, pp. 928–932.

[54] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, “Temporal action detection with structured segment networks,” in *Proc. ICCV*, Oct. 2017, pp. 2933–2942.

[55] H. Xu, A. Das, and K. Saenko, “R-C3D: Region convolutional 3D network for temporal activity detection,” in *Proc. ICCV*, Oct. 2017, pp. 5794–5803.

[56] Y. Tang *et al.*, “COIN: A large-scale dataset for comprehensive instructional video analysis,” in *Proc. CVPR*, 201, pp. 1–10.

[57] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 91–99.

[58] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. CVPR*, Jul. 2017, pp. 1647–1655.

[59] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "Hierarchical deep temporal models for group activity recognition," *CoRR*, abs/1607.02643, Jul. 2016.

[60] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *Proc. ICCVW*, Sep./Oct. 2009, pp. 1282–1289.

[61] Z. Deng, A. Vahdat, H. Hu, and G. Mori, "Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition," in *Proc. CVPR*, Jun. 2016, pp. 4772–4781.

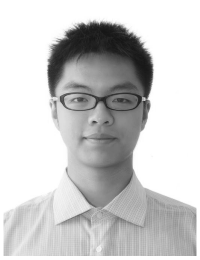
[62] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in *Proc. CVPR*, Jun. 2011, pp. 3281–3288.

[63] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2013, pp. 729–738.

[64] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.

[65] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

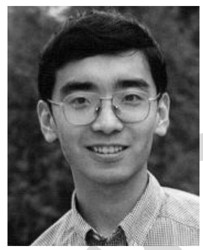
[66] S. Biswas and J. Gall, "Structural recurrent neural network (SRNN) for group activity analysis," in *Proc. WACV*, Mar. 2018, pp. 1625–1632.



Yansong Tang received the B.S. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2015, where he is currently pursuing the Ph.D. degree with the Department of Automation. His research lies in computer vision, especially multi-modal action recognition and egocentric vision analytics.

Jiwen Lu (M'11–SM'15) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision, pattern recognition, and machine learning. He has authored/coauthored over 200 scientific papers in these areas, where over 60 of them are the IEEE TRANSACTIONS papers (including 13 T-PAMI papers) and 50 of them are CVPR/ICCV/ECCV/NIPS papers. He is a member of the Multimedia Signal Processing Technical Committee and the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society, and a member of the Multimedia Systems and Applications Technical Committee and the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society. He was a recipient of the National 1000 Young Talents Program of China in 2015, and the National Science Fund of China for Excellent Young Scholars in 2018, respectively. He serves as the Co-Editor-of-Chief for the *Pattern Recognition Letters*, an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE, and *Pattern Recognition*.

Zian Wang is currently pursuing the B.S. degree with the Department of Automation, Tsinghua University, Beijing, China. His research interests include computer vision and machine learning.



Ming Yang (M'08) received the B.E. and M.E. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2001 and 2004, respectively, and the Ph.D. degree in electrical and computer engineering from Northwestern University, Evanston, IL, USA, in 2008. From 2004 to 2008, he was a Research Assistant with the Computer Vision Group, Northwestern University. After his graduation, he joined NEC Laboratories America, Cupertino, CA, USA, where he was a Senior Researcher. He was a Research Scientist in AI Research at Facebook (FAIR) from 2013 to 2015. He is currently the Co-Founder and VP of software at Horizon Robotics, Inc. His research interests include computer vision, machine learning, face recognition, large scale image retrieval, and intelligent multimedia content analysis. He is the author of over 50 peer-reviewed publications in prestigious international journals and conferences, which have been cited over 9400 times.

Jie Zhou (M'01–SM'04) received the B.S. and M.S. degrees from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology (HUST), Wuhan, China, in 1995. Since then, he has served as a Post-Doctoral Fellow at the Department of Automation, Tsinghua University, Beijing, China, until 1997. Since 2003, he has been a Full Professor with the Department of Automation, Tsinghua University. His research interests include computer vision, pattern recognition, and image processing. In recent years, he has authored over 100 papers in peer-reviewed journals and conferences. Among them, over 30 papers have been published in top journals and conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and CVPR. He is an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and two other journals. He received the National Outstanding Youth Foundation of China Award.