Mining Semantics-Preserving Attention for Group Activity Recognition

Yansong Tang
Tsinghua University
tys15@mails.tsinghua.edu.cn

Zian Wang
Tsinghua University
wza15@mails.tsinghua.edu.cn

Peiyang Li
Tsinghua University
lipy15@mails.tsinghua.edu.cn

Jiwen Lu*
Tsinghua University
lujiwen@tsinghua.edu.cn

Ming Yang
Horizon Robotics, Inc.
ming.yang@horizon-robotics.com

Jie Zhou
Tsinghua University
jzhou@tsinghua.edu.cn

ABSTRACT

In this paper, we propose a Semantics-Preserving Teacher-Student (SPTS) model for group activity recognition in videos, which aims to mine the semantics-preserving attention to automatically seek the key people and discard the misleading people. Conventional methods usually aggregate the features extracted from individual persons by pooling operations, which cannot fully explore the contextual information for group activity recognition. To address this, our SPTS networks first learn a Teacher Network in semantic domain, which classifies the *word* of group activity based on the *words* of individual actions. Then we carefully design a Student Network in vision domain, which recognizes the group activity according to the input videos, and enforce the Student Network to mimic the Teacher Network during the learning process. In this way, we allocate semantics-preserving attention to different people, which adequately explores the contextual information of different people and requires no extra labelled data. Experimental results on two widely used benchmarks for group activity recognition clearly show the superior performance of our method in comparisons with the state-of-the-arts.

CCS CONCEPTS

• **Computer methodologies → Activity recognition and understanding**;

KEYWORDS

Semantic-Preserving; Attention; Group Activity Recognition; Teacher-Student Networks

ACM Reference Format:
Yansong Tang, Zian Wang, Peiyang Li, Jiwen Lu, Ming Yang, and Jie Zhou. 2018. Mining Semantics-Preserving Attention for Group Activity Recognition. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240576

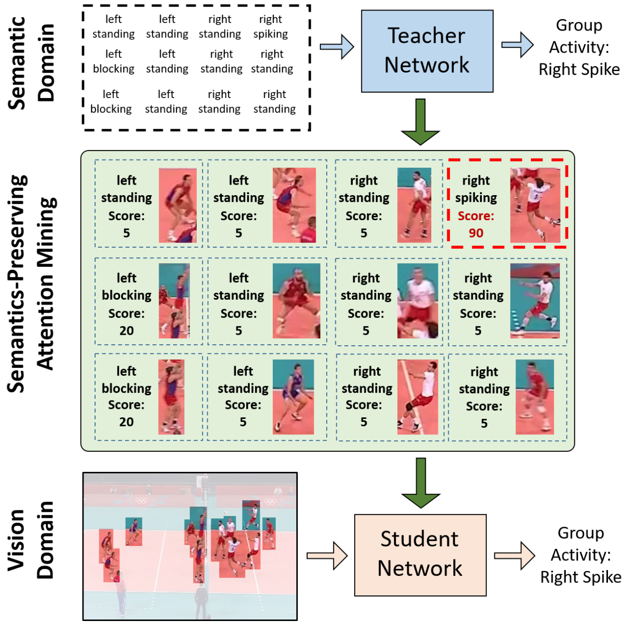


Figure 1: The basic idea of the SPTS networks. In the semantic domain, the task is to map the *words* of individual actions, which can be treated as a caption of the video [1], to the *word* of group activity. In the vision domain, we attempt to predict the label of group activity based on the corresponding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

1 INTRODUCTION

Group activity recognition (*a.k.a.* collective activity recognition), which refers to discern 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 widely real-world applications for group activity recognition including traffic surveillance, social role understanding and sports video analysis. Compared with

*The corresponding author is Dr. Jiwen Lu.

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240576

conventional action recognition which focuses on a single person, group activity recognition is a more challenging task as it requires further understanding the high-level relationship among different people. Hence, it is desirable to design a model to aggregate the individual dynamics across people and explore their contextual information for effective group activity recognition.

During the past few years, great efforts have been devoted to mine the contextual information for group activity recognition. In the early times, a typical series of approaches are developed to design graph-based structure models based on hand-crafted features [7–10]. However, these methods required strong prior knowledge and lack discriminative power to model the temporal evolution of group activity. In recent years, with the promising progress of deep learning methods, researchers attempt to build different deep neural networks [2, 4] 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 Figure 1 as an example. The bottom of Figure 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 works [5, 11] have proposed attention-based models to assign different weights to different people. Specifically, they learn the weights based on the features extracted from input videos, and allocate these weights to their corresponding features. However, such a "self-attention" scheme is essentially lack of physical explanation and is not reliable enough to find the key person for activity recognition in the video.

In this work, we move a new step towards the interaction of vision domain and semantic domain, and propose a Semantics-Preserving Teacher-Student (SPTS) model for group activity recognition. Figure 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 a similar model (namely Student Network), which predicts the group activity from the individual actions in vision domain. Then, we carefully 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 SPTS networks require no extra labelled data. We evaluate our approach on the widely used Volleyball dataset and Collective Activity Dataset, where the experimental results show that the SPTS networks outperform the state-of-the-arts for group activity recognition.

2 RELATED WORK

Group Activity Recognition: Activity recognition is a broadly researched field [12–27], where group activity recognition is an important topic. There have been many methods for group activity recognition in recent years [1–7], which 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 examples, 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.* [3] 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 human roles in aerial videos. However, these methods relied on hand-crafted features, which required strong prior knowledge and were short of discriminative power to capture the temporal cue. For the deep learning based methods, various 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.* [4] extended this work by introducing 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. However, these works mainly focused on the vision domain, which igonred the semantic relationship between the individual actions and group activity. More recently, Li *et al.* [1] 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 vision domains, and employ the semantic knowledge to enhance the performance of vision task.

Attention-based Models: Attention-based model, motivated by the attention mechanisms of primate visual system [28, 29], aims to select the most informative parts from the global field. During the past two decades, attention-based models have been widely applied into the realm of natural language processing (e.g., machine translation [30, 31]), computer vision (e.g., video face recognition [32, 33], person re-identification [34], object localization [35]), and their intersection (e.g., image caption [36], video caption [37] and visual question answering [38]). As for human action/activity recognition, Liu *et al.* [39] developed global context-aware attention LSTM networks to select the informative joints in skeleton-based videos. Further more, Song *et al.* [40] proposed a spatial-temporal attention-based model to learn the importance of different joints and different frames. Different from these two works [39, 40], 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 vision domain, which utilize the semantic information adequately and make the learned attention interpretable.

Knowledge Distillation: The concept of "knowledge distillation" is originated from the work [41] 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 network based on the intermediate layers [36, 42, 43]. For example, Yim *et al.* [43] 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.* [44] employed technique of function-preserving transformations to accelerate the learning process of student network. The most related work to ours is [36], which also utilized the information across the attention mechanisms of two networks. Different from [36], where the input 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 vision domain) and utilizes the additional recurrent neural network to address a more challenging task of group activity recognition.

3 APPROACH

The motivation of this work is to adequately explore the information in both vision domian and semantic domain for group activity recognition. In this section, we first formulate the problem, and then present the details of our SPTS networks.

3.1 Problem Formulation

We denote a tri-tuples (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 test phase. In many previous works, researchers utilize a set of cropped images of each single person at each frame (*i.e.,* tracklets) $X = \{x_1^t, x_2^t, ...x_n^t, ...x_N^t\}_{t=1}^{T}$ as inputs, where t represents the tth frame. We follow this problem setting in our work.

3.2 SPTS Networks

Our SPTS networks consist of two subnetworks, namely Student Network and Teacher Network. Figure 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 vision 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 a combination of semantic domain and vision 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 attention-transfer networks as follows.

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 learnable parameters of the student network. For 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 LSTM network to extract features of X, which is a similar scheme according to [2]. Then, we concatenate the last-fc-layer feature of DCNN and feature of LSTM layer. 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:

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

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

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

where β_n^t is a softmax normalization of the scores. Instead of conventional aggregation methods like max-pooling or mean-pooling, we fuse the feature of each individual person at timestep t as:

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

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_{agg}^t, 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.

Teacher Network: As illustrated above, our Student Network can be regarded as extending a typical self-attention mechanism on the hierarchical deep temporal models proposed in [2]. 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 another semantic space, 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 their 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:

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

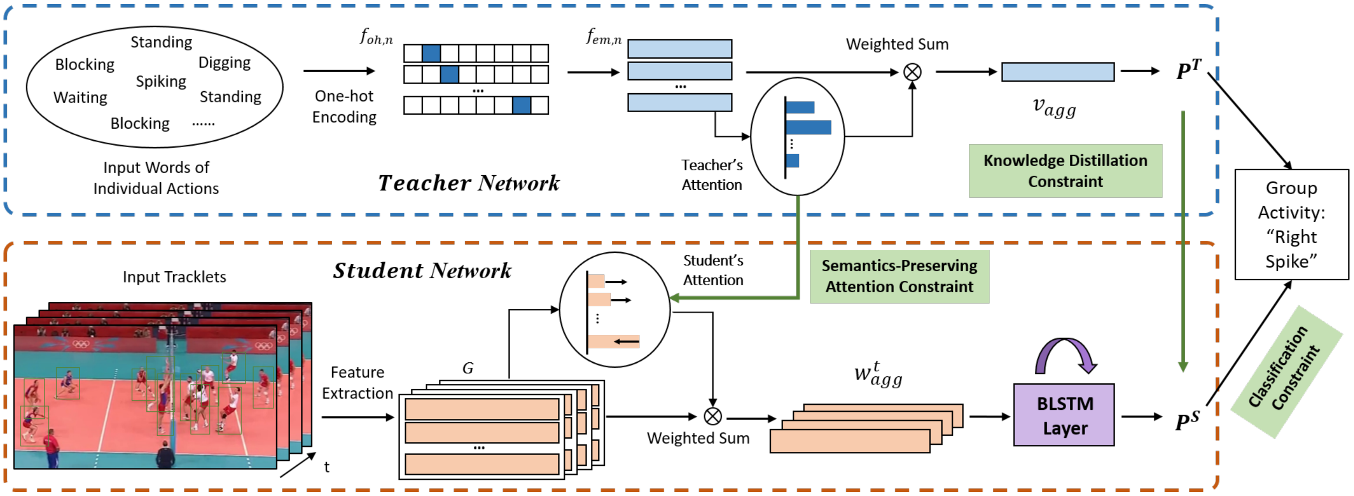


Figure 2: A framework of our proposed SPTS networks, which contains two sub-networks. We first train the Teacher network, which models relationship between words of individual actions and the word of group activity. Next, we train the student network, which takes a set of tracklets as input and predicts 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.

where W_2 and b_2 are the weighted matrix and biased term, $ReLU$ denotes the nonlinear activation function [45]. 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 this model using the ground-truth label z, and achieve high performance due to the discriminative power of the semantic space and the complementary property of semantic domain and vision domain.

Semantics-Preserving Attention Learning: As we described, there are two attention mechanism in this work and they both work separately via 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 mechanism. 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 mechanism.

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

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

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 of these two types of attention, and minimize it during the optimal process. The third term J_{KD} denotes the loss of knowledge distillation [41], in which P_T and P_S are the softmax output of the Teacher and Student Network respectively.

To optimize (8), we employ the back propagation through time (BPTT) algorithm 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 $Y = \{y_n\}_{n=1}^{N}$ are not available during the testing stage.

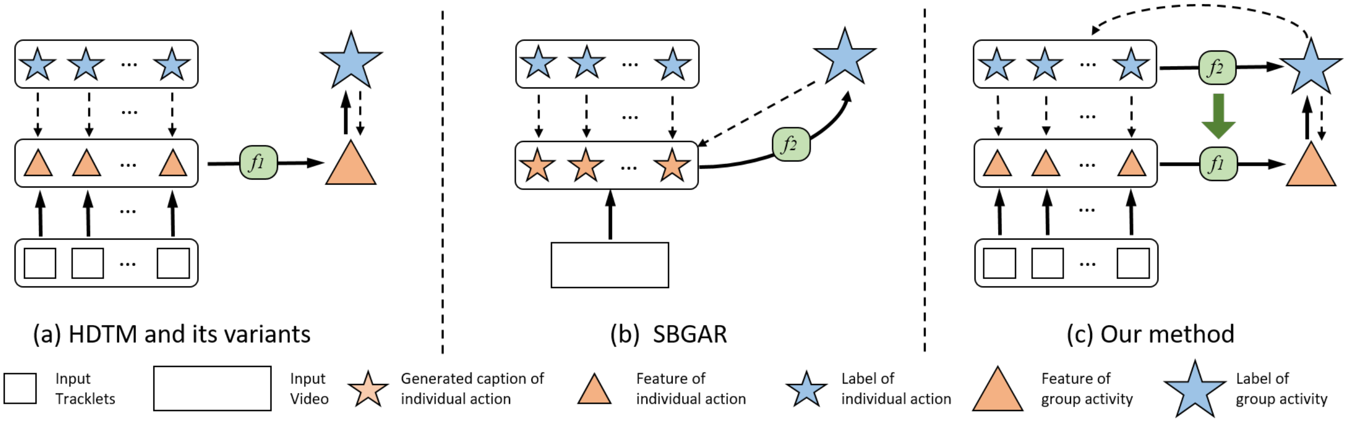

f1
f2
f2
f1
(a) HDTM and its variants
(b) SBGAR
(c) Our method
Input Tracklets
Input Video
Generated caption of individual action
Feature of individual action
Label of individual action
Feature of group activity
Label of group activity

Figure 3: 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 learns f_2 to model the relationship between individual action labels and group activity label. Then we employ the attention knowledge in f_2 to guide f_1 when aggregating features of individual actions to feature of group activity, and make the final prediction.

Algorithm 1: SPTS

```
Input: Training samples: $\{X, Y, z\}$, Parameters: $\Gamma$ (iterative
        number) and $\epsilon$ (convergence error).
Output: the weights of Student Network $\theta_s$.
// Teacher Network Training:
Optimize the parameter $\theta_t$ of Teacher Network with $(Y, z)$.
// Student Network Training:
Finetuned the DCNN and the train first LSTM with $(X, Y)$ [2].
Extract features $G$ from $X$.
Initialize $\theta_s$.
Perform forward propagation.
Calculate the initial $\boldsymbol{J}_0$ by (8).
for $i \leftarrow$ 1, 2, ..., $\Gamma$ do
    Update $\theta_s$ 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 $\theta_s$ of Student Network.
```

3.3 Discussion

This section clarifies the difference of our SPTS networks in comparison with other two categories DNN-based methods. The first category, such as HTDM [2] and its variants [4] shown in Figure 3(a), mainly focus on the vision domain. They first learn features of individual person with an LSTM network, 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 their 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 is lack of physical meaning. The second category, such as SBGAR [1] shown in Figure 3(b), focuses on the semantic domain. This method 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 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 SPTS networks in Figure 3(c), adequately leverage the information in the vision domain and semantic domain for group acitivity recognition. We distill the knowledge in f_2 learned in semantic domain to guide the training process of f_1 in vision domain.

4 EXPERIMENTS

4.1 Datasets and Experiment Settings

Volleyball Dataset [46]: 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 [46] 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.

Collective Activity Dataset (CAD) [47]: The Collective Activity Dataset is a widely used benchmark for the task of group activity recognition. It comprises of 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 [47] 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 CAD dataset, which can better evaluate the performance of the classifiers.

4.2 Implementation Details and Baselines

Our SPTS was 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 SPTS 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 embeded features were weighted and sumed 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 [48] pretrained on ImageNet [49] 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 embeded features of each person were then sumed 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 [25], we computed the optical flow between two adjacent video frames using Flownet 2.0 [50]. We extracted DCNN and LSTM features of optical flow tracklets, and concatenate them with the features extracted from orignal RGB tracklets before the attention mechanism 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 tracklet 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 Teachers 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 mechanism upon the HDTM [2] .
- Ours-SPA (semantics-preserving attention): A version of model which employ 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 [41] with Ours-SPA.
- Ours-SPA (+OF), Ours-SPA (+OF), Ours-SPA+KD (+OF): Models of combining the optical flow input based on the Ours-SPA, Ours-SPA and Ours-SPA+KD, respectively.

Table 1: Comparison of the group activity recognition accuracy (%) on the volleyball dataset

Method	MCA	MPCA
CERN-2 [4]	83.3	83.6
SSU [5]	89.9	–
SRNN [51]	83.5	–
Ours-teacher*	88.3	84.4
Ours-teacher	69.3	66.8
Baseline-HDTM [2]	81.9	82.9
Ours-SA	87.1	86.1
Ours-SPA	89.3	89.2
Ours-SPA + KD	89.3	89.0
Ours-SA (+OF)	87.7	87.0
Ours-SPA (+OF)	89.6	89.5
Ours-SPA + KD (+OF)	**90.7**	**90.0**

4.3 Results on the Volleyball Dataset

We first evaluate our method on the Volleyball dataset. We follow [2] to seperate 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 positions. Table 1 presents the comparison with different approaches, where our SPTS networks achieve 90.7% MCA and 90.0% MPCA, outperforming existing state-of-the-art methods for group activity recognition. 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

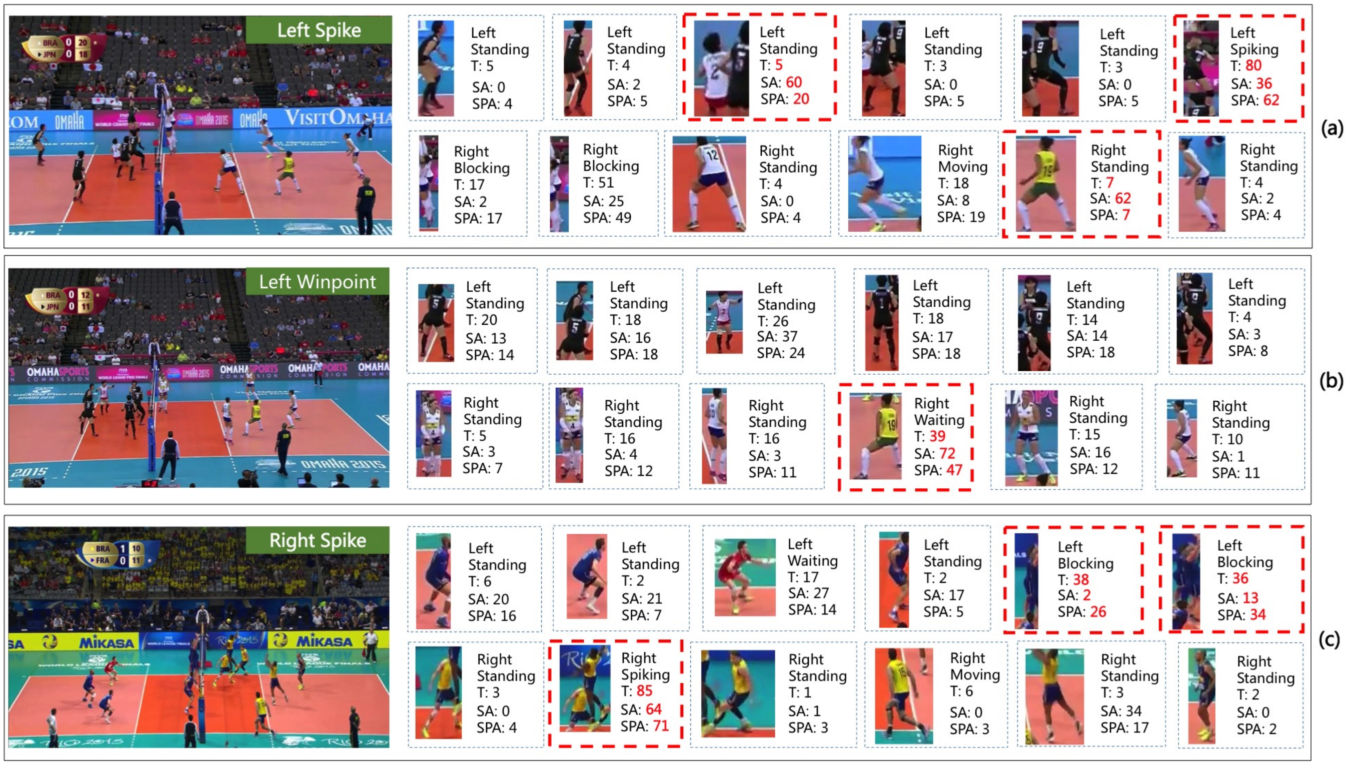

Left Spike
Left Standing T: 5 SA: 0 SPA: 4
Left Standing T: 4 SA: 2 SPA: 5
Left Standing T: 5 SA: 60 SPA: 20
Left Standing T: 3 SA: 0 SPA: 5
Left Standing T: 3 SA: 0 SPA: 5
Left Spiking T: 80 SA: 36 SPA: 62
Right Blocking T: 17 SA: 2 SPA: 17
Right Blocking T: 51 SA: 25 SPA: 49
Right Standing T: 4 SA: 0 SPA: 4
Right Moving T: 18 SA: 8 SPA: 19
Right Standing T: 7 SA: 62 SPA: 7
Right Standing T: 4 SA: 2 SPA: 4
(a)
Left Winpoint
Left Standing T: 20 SA: 13 SPA: 14
Left Standing T: 18 SA: 16 SPA: 18
Left Standing T: 26 SA: 37 SPA: 24
Left Standing T: 18 SA: 17 SPA: 18
Left Standing T: 14 SA: 14 SPA: 18
Left Standing T: 4 SA: 3 SPA: 8
Right Standing T: 5 SA: 3 SPA: 7
Right Standing T: 16 SA: 4 SPA: 12
Right Standing T: 16 SA: 3 SPA: 11
Right Waiting T: 39 SA: 72 SPA: 47
Right Standing T: 15 SA: 16 SPA: 12
Right Standing T: 10 SA: 1 SPA: 11
(b)
Right Spike
Left Standing T: 6 SA: 20 SPA: 16
Left Standing T: 2 SA: 21 SPA: 7
Left Waiting T: 17 SA: 27 SPA: 14
Left Standing T: 2 SA: 17 SPA: 5
Left Blocking T: 38 SA: 2 SPA: 26
Left Blocking T: 36 SA: 13 SPA: 34
Right Standing T: 3 SA: 0 SPA: 4
Right Spiking T: 85 SA: 64 SPA: 71
Right Standing T: 1 SA: 1 SPA: 3
Right Moving T: 6 SA: 0 SPA: 3
Right Standing T: 3 SA: 34 SPA: 17
Right Standing T: 2 SA: 0 SPA: 2
(c)

Figure 4: Visualization of the learned attention on the Volleyball dataset. 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 indvidual 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).

also discover that, combining with the optical flow can lead to a slight improvement on this dataset. Besides, though the Our-teacher*, which takes the groud-truth of individual actions as the input of Teacher Network, reaches high performance of 88.3% MCA. Our-teacher only attains 69.3% MCA, which utilizes the predicted individual actions as inputs. This is because, the Teacher Network is sensitive to the inputs and the incorrectedly predicted individual acitions will greatly harm the performance of the final recognition.

We also show several visualization results of the learned attention in Figure 4. The group activity label of Figure 4(a) is "left spike". For the self-attention model of 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 space, e.g., the spiking (T: 80), and the blocking(51). And after employing our SPTS networks, we will transfer this attention knowledge from semantic space to the vision space, 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 Figure 4(b) is "left winpoint", where there's 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. Figure 4(c) also illustrates some similar results to Figure 4(a).

4.4 Results on the CAD dataset

Table 2 shows the comparison with different methods on the CAD dataset. The MPCA results of other approaches are computed based on the original confusion matrices in [1–4, 6, 7]. We observe that, our method achieves 95.7% MPCA, outperforming the state-of-the-arts by 4.9%. Moreover, our method have improved the baseline method HDTM [2] by 6.1%. 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 attention-guided method brings 1.4% improvement over self-attention(SA) model. This improvement is less significant than that on the Volleyball dataset because the setting of the CAD dataset is to assign what the major people are doing to the label of group activity. Hence, attention model is not so important. Besides, compared with SBGAR and Ours-teacher, which directly utilized the semantic information to predict the final labels, our methods achieves 5.8% and 7.5% improvement, which demonstrates its effectiveness.

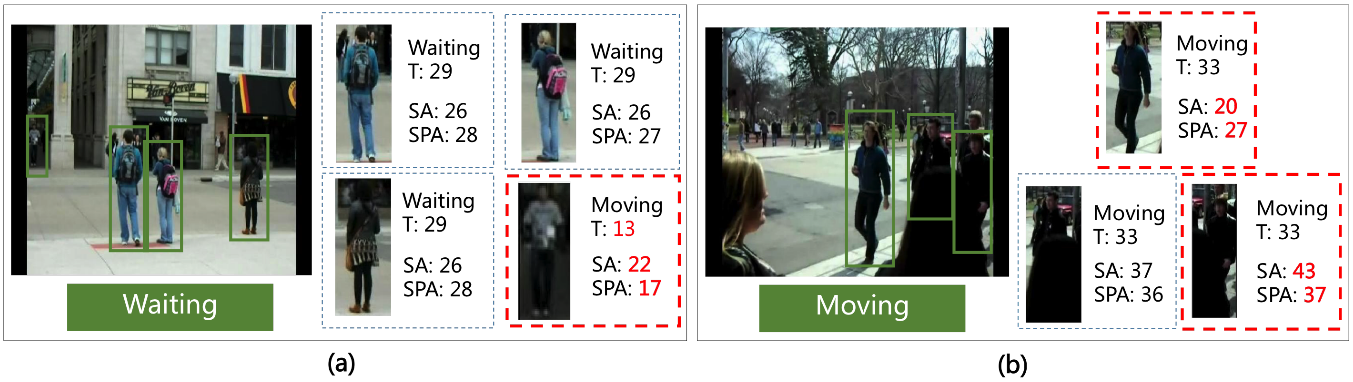
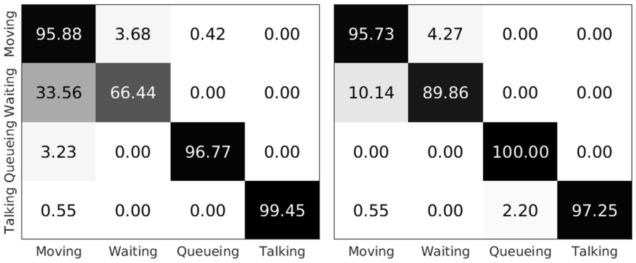


Figure 5: Visualization of the learned attention on the CAD dataset.



(a) Baseline [2]

(b) Our Method

Figure 6: Comparison of Confusion Matrices on CAD [47]. We merge the class of *Walking* and *Crossing* as the same class of *Moving* as suggested in [6].

Table 2: Comparison of the group activity recognition accuracy (%) on the CAD dataset

Method	MPCA
Cardinality kernel [3]	88.3
CERN-2 [4]	88.3
RMIC [6]	89.4
SBGAR [1]	89.9
MTCAR [7]	90.8
Ours-teacher*	97.6
Ours-teacher	88.2
baseline-HDTM [2]	89.6
Ours-SA	91.5
Ours-SPA	92.3
Ours-SPA + KD	92.5
Ours-SA (+OF)	94.3
Ours-SPA (+OF)	**95.6**
Ours-SPA + KD (+OF)	**95.7**

We also show the visualization of the learned attention in Figure 5. As shown in Figure 5(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) to the "moving" person, which can be regarded as a noise for recognizing the group activity. For Figure 5(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 increase the attention of the top person from 20 to 27, and decrease the attention of the right person from 43 to 37, so that the information of three people can be utilized equally. Figure 6 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".

5 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 vision domain, which explicitly exploits the attention information of the group people. Both quantitative and qualitative experimental results on the widely-used CAD dataset and Volleyball dataset have shown the superior performance of our proposed method in comparison with the state-of-the-arts. To our best knowledge, these are original efforts leveraging attention in multimedia clues, *i.e.,* both semantic and vision clues, performing group activity recognition. In the future, it is an interesting direction to employ 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.

ACKNOWLEDGEMENT

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 61672306, Grant U1713214, Grant 61572271, and Grant 61527808, in part by the National 1000 Young Talents Plan Program, and in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564. We would like to thank Tianmin Shu from VCLA@UCLA for his generous help on reimplementing the baseline model.

REFERENCES

[1] Xin Li and Mooi Choo Chuah. SBGAR: semantics based group activity recognition. In *ICCV*, pages 2895–2904, 2017.

[2] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, pages 1971–1980, 2016.

[3] Hossein Hajimirsadeghi, Wang Yan, Arash Vahdat, and Greg Mori. Visual recognition by counting instances: A multi-instance cardinality potential kernel. In *CVPR*, pages 2596–2605, 2015.

[4] Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. CERN: confidence-energy recurrent network for group activity recognition. In *CVPR*, pages 4255–4263, 2017.

[5] Timur M. Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *CVPR*, pages 3425–3434, 2017.

[6] Minsi Wang, Bingbing Ni, and Xiaokang Yang. Recurrent modeling of interaction context for collective activity recognition. In *CVPR*, pages 7408–7416, 2017.

[7] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *ECCV*, pages 215–230, 2012.

[8] Wongun Choi, Khuram Shahid, and Silvio Savarese. Learning context for collective activity recognition. In *CVPR*, pages 3273–3280, 2011.

[9] Tian Lan, Yang Wang, Weilong Yang, Stephen N. Robinovitch, and Greg Mori. Discriminative latent models for recognizing contextual group activities. *TPAMI*, 34(8):1549–1562, 2012.

[10] Tianmin Shu, Dan Xie, Brandon Rothrock, Sinisa Todorovic, and Song-Chun Zhu. Joint inference of groups, events and human roles in aerial videos. In *CVPR*, pages 4576–4584, 2015.

[11] Yongyi Tang, Peizhen Zhang, Jianfang Hu, and Wei-Shi Zheng. Latent embeddings for collective activity recognition. In *AVSS*, pages 1–6, 2017.

[12] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *CVPR*, pages 3169–3176, 2011.

[13] Alejandro Hernandez Ruiz, Lorenzo Porzi, Samuel Rota Bulò, and Francesc Moreno-Noguer. 3d cnns on distance matrices for human action recognition. In *ACM MM*, pages 1087–1095, 2017.

[14] Xinyu Li, Yanyi Zhang, Jianyu Zhang, Yueyang Chen, Huangcan Li, Ivan Marsic, and Randall S. Burd. Region-based activity recognition using conditional GAN. In *ACM MM*, pages 1059–1067, 2017.

[15] Congqi Cao, Yifan Zhang, and Hanqing Lu. Spatio-temporal triangular-chain CRF for activity recognition. In *ACM MM*, pages 1151–1154, 2015.

[16] Zheng Zhou, Kan Li, and Xiangjian He. Recognizing human activity in still images by integrating group-based contextual cues. In *ACM MM*, pages 1135–1138, 2015.

[17] Wenchao Jiang and Zhaozheng Yin. Human activity recognition using wearable sensors by deep convolutional neural networks. In *ACM MM*, pages 1307–1310, 2015.

[18] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *TPAMI*, 35(1):221–231, 2013.

[19] Jen-Yin Chang, Kuan-Ying Lee, Yu-Lin Wei, Kate Ching-Ju Lin, and Winston Hsu. Location-independent wifi action recognition via vision-based methods. In *ACM MM*, pages 162–166, 2016.

[20] Yi Tian, Qiuqi Ruan, Gaoyun An, and Yun Fu. Action recognition using local consistent group sparse coding with spatio-temporal structure. In *ACM MM*, pages 317–321, 2016.

[21] Pichao Wang, Zhaoyang Li, Yonghong Hou, and Wanqing Li. Action recognition based on joint trajectory maps using convolutional neural networks. In *ACM MM*, pages 102–106, 2016.

[22] Lei Wang, Xu Zhao, Yunfei Si, Liangliang Cao, and Yuncai Liu. Context-associative hierarchical memory model for human activity recognition and prediction. *TMM*, 19(3):646–659, 2017.

[23] Wanru Xu, Zhenjiang Miao, Xiao-Ping Zhang, and Yi Tian. A hierarchical spatio-temporal model for human activity recognition. *TMM*, 19(7):1494–1509, 2017.

[24] Yanan Guo, Dapeng Tao, Jun Cheng, Alan Dougherty, Yaotang Li, Kun Yue, and Bob Zhang. Tensor manifold discriminant projections for acceleration-based human activity recognition. *TMM*, 18(10):1977–1987, 2016.

[25] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014.

[26] Yansong Tang, Yi Tian, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Action recognition in rgb-d egocentric videos. In *ICIP*, pages 3410–3414, 2017.

[27] Yansong Tang, Yi Tian, Jiwen Lu, Peiyang Li, and Jie Zhou. Deep progressive reinforcement learning for skeleton-based action recognition. In *CVPR*, 2018.

[28] John K. Tsotsos, Sean M. Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545, 1995.

[29] Ronald A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7(1-3):17–42, 2000.

[30] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2014.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 6000–6010, 2017.

[32] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *CVPR*, pages 5216–5225, 2017.

[33] Yongming Rao, Jiwen Lu, and Jie Zhou. Attention-aware deep reinforcement learning for video face recognition. In *ICCV*, pages 3931–3940, 2017.

[34] Albert Haque, Alexandre Alahi, and Li Fei-Fei. Recurrent attention models for depth-based person identification. In *CVPR*, pages 1229–1238, 2016.

[35] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *CVPR*, pages 685–694, 2015.

[36] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Paying more attention to saliency: Image captioning with saliency and context attention. In *ICLR*, 2017.

[37] Zhao Guo, Lianli Gao, Jingkuan Song, Xing Xu, Jie Shao, and Heng Tao Shen. Attention-based LSTM with semantic consistency for videos captioning. In *ACM MM*, pages 357–361, 2016.

[38] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. Stacked attention networks for image question answering. In *CVPR*, pages 21–29, 2016.

[39] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C. Kot. Skeleton-based human action recognition with global context-aware attention LSTM networks. *TIP*, 27(4):1586–1599, 2018.

[40] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, pages 4263–4270, 2017.

[41] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. 2014.

[42] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2014.

[43] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, pages 7130–7138, 2017.

[44] Tianqi Chen, Ian J. Goodfellow, and Jonathon Shlens. Net2net: Accelerating learning via knowledge transfer. In *ICLR*, 2015.

[45] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

[46] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. Hierarchical deep temporal models for group activity recognition. *CoRR*, abs/1607.02643, 2016.

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

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[49] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[50] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, pages 1647–1655, 2017.

[51] Sovan Biswas and Juergen Gall. Structural recurrent neural network (srnn) for group activity analysis. In *WACV*, 2018.