




Unified Spatio-Temporal Attention Networks for Action Recognition in Videos

Dong Li , Ting Yao , Ling-Yu Duan , *Member, IEEE*, Tao Mei , *Senior Member, IEEE*,
and Yong Rui, *Fellow, IEEE*

Abstract—Recognizing actions in videos is not a trivial task because video is an information-intensive media and includes multiple modalities. Moreover, on each modality, an action may only appear at some spatial regions, or only part of the temporal video segments may contain the action. A valid question is how to locate the attended spatial areas and selective video segments for action recognition. In this paper, we devise a general attention neural cell, called *AttCell*, that estimates the attention probability not only at each spatial location but also for each video segment in a temporal sequence. With *AttCell*, a unified Spatio-Temporal Attention Networks (STAN) is proposed in the context of multiple modalities. Specifically, STAN extracts the feature map of one convolutional layer as the local descriptors on each modality and pools the extracted descriptors with the spatial attention measured by *AttCell* as a representation of each segment. Then, we concatenate the representation on each modality to seek a consensus on the temporal attention, *a priori*, to holistically fuse the combined representation of video segments to the video representation for recognition. Our model differs from conventional deep networks, which focus on the attention mechanism, because our temporal attention provides a principled and global guidance across different modalities and video segments. Extensive experiments are conducted on four public datasets; UCF101, CCV, THUMOS14, and Sports-1M; our STAN consistently achieves superior results over several state-of-the-art techniques. More remarkably, we validate and demonstrate the effectiveness of our proposal when capitalizing on the different number of modalities.

Index Terms—Action recognition, spatio-temporal attention, deep convolutional networks.

I. INTRODUCTION

TODAY'S digital contents are inherently multimedia: text, image, audio, video and so on. Video in particular has

Manuscript received January 8, 2018; revised May 16, 2018 and July 1, 2018; accepted July 7, 2018. Date of publication August 1, 2018; date of current version January 24, 2019. This work was supported in part by the PKU-NTU Joint Research Institute (JRI) sponsored by a donation from the Ng Teng Fong Charitable Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chang-Su Kim. (*Corresponding author: Tao Mei.*)

D. Li is with the University of Science and Technology of China, Hefei 230000, China (e-mail: dongli1995.ustc@gmail.com).

T. Yao is with the Microsoft Research, Beijing 100080, China (e-mail: tingyao.ustc@gmail.com).

L.-Y. Duan is with the National Engineering Lab for Video Technology, Peking University, Beijing 100080, China (e-mail: lingyu@pku.edu.cn).

T. Mei is with the JD AI Research, Beijing 100101, China (e-mail: tmei@live.com).

Y. Rui is with the Lenovo, Beijing 100085, China (e-mail: yongrui@lenovo.com).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMM.2018.2862341

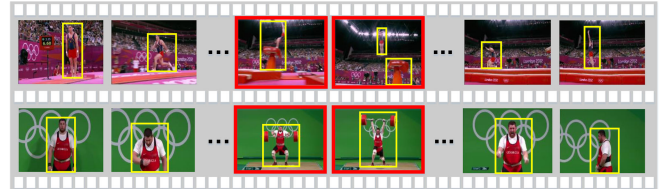


Fig. 1. Examples showing spatial and temporal attention over a video sequence from “vault” (upper row) and “weight lifting” (lower row) category, respectively. Yellow box indicates the spatial focus in a frame while red box indicates the crucial frame in the video for recognizing the action.

become a new method of communication between Internet users with the proliferation of sensor-rich mobile devices. This trend has encouraged the development of advanced techniques for a broad range of video-understanding applications. A fundamental issue that underlies the success of these technological advances is action recognition. Nevertheless, video is an information-intensive media and includes multiple modalities, e.g., frame, motion or audio, which makes the problem of action recognition very challenging. Moreover, taking the modality of frame as an example in Figure 1, the focus of an action (e.g., “vault” or “weight lifting”) may concentrate in spatial regions of interest (yellow boxes) instead of an entire frame. Similarly, different frames in the temporal sequence contribute differently to the recognition of an action and the red box indicates the crucial frame for recognition in the figure. Therefore, recognizing actions in videos should consider both the spatial and temporal focus of attention.

There has been extensive research on video action recognition, including hand-crafted feature-based methods [1]–[6] and deep models of video representation learning [7]–[11]. The first category of research focuses predominantly on the detection of spatio-temporal interest points followed by describing these points with local representation or further encoding local features into a vectorial representation. The direction of deep learning methods heavily relies on Convolutional Neural Networks (CNN) to learn the visual appearance or on Recurrent Neural Networks (RNN) to model the temporal dynamics in the video. Although most of these methods explore spatial appearances and temporal dynamics in videos for action recognition, the attention mechanism by jointly utilizing the spatio-temporal dependencies is not fully exploited. More importantly, there have been no rigorous studies on the joint exploration of spatio-temporal attention, particularly when there are multiple modalities to consider for recognition.

By consolidating the idea of exploring both spatial and temporal attention in the context of multiple modalities, we propose a unified Spatio-Temporal Attention Networks (STAN) for video action recognition. Specifically, a video is temporally decomposed into multiple video segments by even sampling. Each video segment is represented by multiple modalities (e.g., frame, motion (optical flow) and clip), and we model each modality as a single stream. In each stream, the activations of one convolutional layer in CNN form local descriptors on each modality. Our spatial Attention Neural Cell (*AttCell*) is devised to produce the attention distribution over all local regions to infer the action and spatially pool these local descriptors as a representation of each video segment. Then, the representations of each video segment on different modalities are concatenated and sequentially fed into a Long Short-Term Memory (LSTM) network to learn the temporal attention via temporal *AttCell*, which diverts the attention to the most indicative video segments. Thus, the temporal *AttCell* seeks a global and holistic consensus on the temporal attention across all modalities. The temporal attention is exploited to fuse the concatenated representations of video segments to a video-level representation, which is finally input into a softmax layer to predict the action in the video. The entire architecture is trainable in an end-to-end fashion.

The main contribution of this work is the proposal of a unified spatio-temporal attention architecture for recognizing actions in videos. By identifying the most distinctive regions in each video segment and the most indicative segments in a video to infer the action, our work takes a further step forward to enhance action recognition. Moreover, our solution leads to an elegant view of how the attention mechanism should be modeled and leveraged in the context of multiple modalities, which is a problem not yet fully understood in the literature. The remaining sections are organized as follows. Section II describes related works on video action recognition. Section III presents our proposed spatio-temporal attention networks for recognition. The implementation details are shown in Section IV. Section V provides the empirical evaluations, followed by the conclusions in Section VI.

II. RELATED WORKS

Video action recognition has attracted intensive research interests in recent years. We briefly group the methods for action recognition into two major categories: hand-crafted feature-based methods and deep learning-based models.

Hand-crafted feature-based methods can usually be decomposed into two phrases: detecting spatio-temporal interest points and describing the visual patterns of those points with local representation. Many video representations are derived from the image domain and extended to measure the temporal dimension of 3D volumes. For example, Laptev proposes space-time interest points (STIP) by extending the 2D Harris corner detector into 3D space [1]. The Histogram of Gradient (HOG) and Histogram of Optical Flow (HOF) [12], 3D Histogram of Gradient (HOG3D) [13], SIFT-3D [2], Extended SURF [14] and Cuboids [15] are all good descriptors as the local spatio-temporal

features. Recently, Wang *et al.* propose dense trajectory features, which densely sample local patches from each frame at different scales and then track them in a dense optical flow field [16]. Further improvements are achieved by the compensation of camera motion [17] and the use of advanced feature encoding methods such as Bag-of-Words (BOW) [12], [18] and Fisher Vectors [3], [19]. However, these hand-crafted descriptors are not particularly optimized for the video action recognition task and may lack discriminative capacity in this task.

To overcome the limitations of low-level local descriptors, several mid-level representations have been proposed for action recognition. These approaches attempt to decompose an action into “parts” designed to capture aspects of the local spatial or temporal structure in the data. These “parts” typically correspond to semantic entities such as humans and objects. For instance, in [20], object context and object reactions are used to recognize actions that may otherwise be too similar to distinguish or too difficult to observe. Wang *et al.* [21] propose a data-driven approach to discover those effective parts with high motion salience. Although these approaches attempt to generate a discriminative representation and learn the structure of the videos in terms of constituent objects, one of their inherent drawbacks is that they rely heavily on the success of object and action detection.

Recent approaches for action recognition are to devise **deep architectures** for learning video representation. [22] and [23] perform action recognition using Support Vector Machine with mean pooling of the CNN-based representations over frames. Xu *et al.* [24] exploit multi-scale pooling on the last pooling layer to obtain the latent concept descriptors and encode them by Vector of Locally Aggregated Descriptors (VLAD) for event detection. Qiu *et al.* [25] design a novel end-to-end deep quantization architecture by incorporating the Fisher Vector encoding strategy into deep generative models. To further model the motion information, the CNN-based architecture is extended by stacking visual features in a fixed size of windows and using spatio-temporal convolutions for video classification in [7]. Later in [9], Tran *et al.* employ 3D ConvNets trained on the Sports-1M dataset to learn spatio-temporal video descriptors, but these deep models achieve lower performance than do the shallow hand-crafted representations [3]. More effective ways of learning spatio-temporal relationships are proposed in [26]–[31]. Simonyan *et al.* [8] design two-stream ConvNets, which contains spatial and temporal nets to capture the discriminative appearance feature and motion feature, respectively. Ng *et al.* [32] highlight a drawback of the two-stream network that exploits a standard image CNN instead of a specialized network for training videos, which makes the two-stream network unable to capture long-term temporal information. To overcome this limitation, the LSTM-RNN networks have been successfully used to model long-range temporal dynamics in videos. Srivastava *et al.* [33] propose learning video representations with LSTM in an unsupervised manner. Li *et al.* [10], [34] propose a multi-granular deep architecture and employ LSTM to incorporate long-term temporal dynamics based on multiple granularity features. Mahasseni *et al.* [35] further use 3D human-skeleton sequences to regularize the learning of

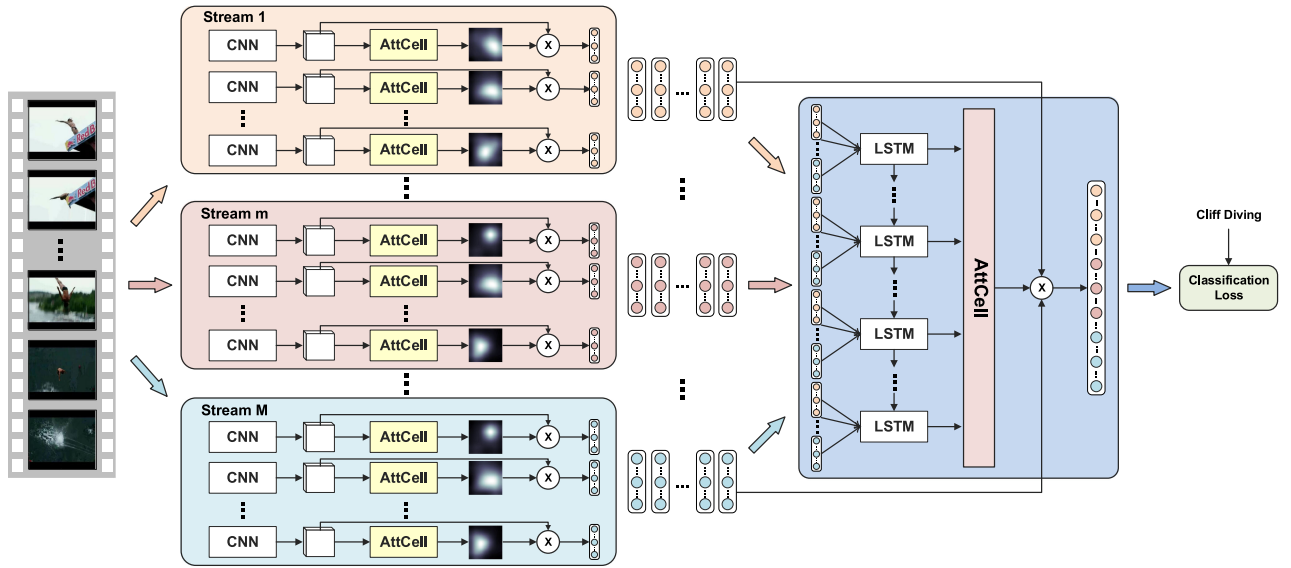


Fig. 2. An overview of our unified STAN architecture for action recognition (better viewed in color). The input video is first represented by multiple modalities and decomposed into multiple streams, followed by CNN to extract convolutional features as local descriptors on each modality. The local descriptors are further pooled with spatial attention as video segment representation by spatial *AttCell*. Then, the representations of each video segment on different modalities are concatenated and fed into LSTM to learn temporal attention with temporal *AttCell*, which is incorporated as a global and holistic guidance across all modalities to fuse the combined video segment representations into video representation. The aggregated video representation is finally exploited for action recognition.

LSTM, which is grounded via DCNN onto the video for action recognition.

More recently, there have been several works that explore attention mechanism for action recognition. For example, [11] presents a recurrent soft attention model by aggregating responses of a convolutional feature map with spatial attention at each step of LSTM to predict the spatial attention of the next frame. [36] extends [11] by upgrading LSTM to multi-scale RNN and implementing hard attention with the aid of Gumbel-softmax. The attention mechanism in the two papers are only applied to the spatial dimension. Furthermore, [37] and [38] improve the original LSTM network by proposing spatio-temporal attention module for action recognition from skeleton data, which represents a person by 3D coordinate positions of key joints and learns to selectively focus on discriminative joints in the skeleton sequence. Nevertheless, the input features of LSTM in these works are all extracted from RGB frames and the motion cues are overlooked. Attention in [39] and [40] then integrates appearance and motion into a unified framework. [39] employs motion features to measure attention and better guide attention towards the relevant locations. [40] incorporates a hierarchical LSTM structure and joint attention model into two-stream ConvNets. The most closely related work is [41], which recurrently learns spatio-temporal attention to explore video context for enhancing the representations of current frame. Ours is different from [41] in the way that we focus more on mining the spatio-temporal attention across different modalities to endow the video representations with more power. Technically, [41] particularly integrates the computation of attention into LSTM, whereas our method devises a general attention neural cell regardless of the task influence. Moreover, the spatial attention is contextually modeled but shared between frame stream and

flow stream in [41], while ours learns spatial attention independently for each frame, and the spatial attention is tailored to each stream. In addition, our work contributes by holistically seeking the temporal attention across different modalities.

III. A UNIFIED STAN

Video is essentially an information-intensive media and can be naturally represented by multiple modalities, e.g., frame, motion (optical flow) and clip. Given an input video, a set of video segments $\{S_1, S_2, \dots, S_T\}$ is delimited by a temporally uniform partition. We model each modality of a video segment as a single stream and construct a multi-stream architecture. The main goal of our Spatio-Temporal Attention Networks (STAN) is to locate the attended areas in each video segment and further highlight the indicative segments among all video segments with our designed Attention Neural Cell (*AttCell*) for action recognition. The training of STAN is performed by exploring both the spatial attention distribution over all local regions of each segment in each stream and the holistic temporal attention across different modalities on all video segments. Figure 2 shows an overview of our approach. In the following, we first define a general attention neural cell *AttCell*, that estimates the attention probability at each spatial location in one video segment and each temporal segment in a video. Then, the detailed process of STAN is presented, including the convolutional feature extraction from CNN on each segment in each stream, spatial attention over local descriptors to represent each video segment on each modality and temporal attention over all video segments modeled with LSTM across all streams to generate video representation. Finally, the generated video representation is fed into the classifiers for action recognition.

A. Attention Neural Cell

To leverage both spatial and temporal attention mechanisms into our action recognition framework, we devise a general attention neural cell (*AttCell*), which is treated as a key element in our architecture. The *AttCell* is designed to assign a positive weight score to each local descriptors extracted from a video segment in each stream or each segment in a video across multiple modalities. The score can be interpreted as either the probability that the spatial regions in a video segment or the temporal segments in a video that should be attended to recognize the target action, or the relative importance of each spatial location (local regions in a segment) or temporal location (specific segments in a video) to generate a segment or video representation, respectively.

Technically, given the input feature set $\mathbf{F} = \{\mathbf{f}_1, \dots, \mathbf{f}_N\}$ to *AttCell*, the corresponding attention score map $\alpha = [\alpha_1, \dots, \alpha_N]$ over the N feature vectors is computed by

$$\alpha = \text{Softmax}(\mathbf{s}) \text{ and } s_n = g_{att}(\mathbf{f}_n; \theta_{att}), \quad (1)$$

where $g_{att}(\cdot)$ is the attention network parameterized by θ_{att} and $\mathbf{s} = [s_n]$ is an intermediate attention score map that is further normalized with the softmax function to produce the positive attention probabilities $\alpha = [\alpha_n]$ in the range from 0 to 1. Note that $g_{att}(\cdot)$ could be any type of network, such as a multilayer perception. Specifically, the attention network utilized here is a convolutional layer with 1×1 kernel, and Eq. (1) can be rewritten as

$$s_n = \mathbf{W}^T \mathbf{f}_n + b, \quad (2)$$

where \mathbf{W} is the convolution weight and b is the bias.

B. Convolutional Features from CNN

In general, different layers in CNN express different information. For example, the outputs of fully-connected layers usually denote high-level concepts, whereas the activations of convolutional layers contain spatial information to represent objects and scenes. Inspired by recent works [11], [24] which exploit effective spatial information from convolutional and pooling layers for action recognition, we follow this elegant recipe and employ the outputs of pooling layers as the representation of video segment on each modality in our framework.

Suppose that there are M streams in our network and that each takes one modality of video segments as inputs. In each stream, we utilize a deep CNN to extract the convolutional features on the modality of video segments. By feeding each input into CNN, we choose the output convolutional features from the last pooling layer, which is denoted as pool_5 here and remains the spatial information of the original input. Moreover, to extend each feature vector from pool_5 with more local context information, we attach an additional convolutional layer (kernel size: 3×3 , stride: 1) on the top of pool_5 layer, named as conv_6 , for convolutional features extraction.

Hence, the convolutional features taken from conv_6 have a dimension of $K \times K \times D$. $K \times K$ is the number of regions in the input and D is the dimension of the feature vector of each region ($K = 7$ and $D = 1024$ in our experiments). Specifically,

in the m -th stream, we extract K^2 D -dimensional vectors as our local descriptors for the t -th video segment, which is represented as

$$\mathbf{F}^{m,t} = \{\mathbf{f}_{1,1}^{m,t}, \dots, \mathbf{f}_{i,j}^{m,t}, \dots, \mathbf{f}_{K,K}^{m,t}\}, \quad \mathbf{f}_{i,j}^{m,t} \in \mathbb{R}^D. \quad (3)$$

Each local descriptor slice maps to different overlapping regions in the original input. We refer to these local descriptors as the feature cube in Figure 2 and aim to model the spatial attention over the K^2 local descriptors for each segment.

C. Spatial Attention over Local Descriptors

In many cases, the target action may only relate to some spatial regions that contain special objects or motion in one segment. Therefore, directly using one global feature vector to predict the action could lead to suboptimal results due to the noises introduced from regions that are irrelevant to the action. To enable the framework to pinpoint the spatial regions that are most indicative for inferring the action, spatial *AttCell* is employed at the top of local descriptors for each segment to explore the spatial attention in each stream.

Given a video with T segments, let $\mathbf{F}^{m,t}$ be the local descriptors extracted from the t -th segment in the m -th stream, as defined in Eq. (3). First, we feed the local descriptors $\mathbf{F}^{m,t}$ into spatial *AttCell*; then, the spatial attention score map $\alpha^{m,t} = [\alpha_{i,j}^{m,t}]$ over all local descriptors of the segment on the specific modality is calculated by

$$\alpha^{m,t} = \text{Softmax}(\mathbf{s}^{m,t}) \text{ and } s_{i,j}^{m,t} = g_{spatial}(\mathbf{f}_{i,j}^{m,t}; \theta_{spatial}^m), \quad (4)$$

where $g_{spatial}(\cdot)$ is the spatial attention network parameterized by $\theta_{spatial}^m$. Based on the spatial attention distribution $\alpha^{m,t}$, we calculate the weighted sum of the local descriptors from all regions and compute the aggregated segment representation $\mathbf{f}_{att}^{m,t}$ as

$$\mathbf{f}_{att}^{m,t} = \sum_{i=1}^K \sum_{j=1}^K \alpha_{i,j}^{m,t} \mathbf{f}_{i,j}^{m,t}. \quad (5)$$

D. Temporal Attention over Video Segments

As video is a sequence of segments with large content variance and complexity, not every segment in the video is relevant to the target action, particularly when the videos are untrimmed and related to multiple topics [42], [43]. Therefore, it is natural to concentrate more on the segments where the action happens and generate the video representation by fusing segment representations with temporal attention. Here, we apply Long Short-Term Memory (LSTM) networks to model the temporal dynamics in videos. The standard LSTM is a variant of RNN, which could capture long-term temporal information in the sequential data by mapping sequences to sequences. More importantly, a consensus across different modalities is leveraged to holistically produce temporal attention. Figure 3 details the process of computing temporal attention for each segment and fusing the representations of all video segments with temporal attention to generate video representation. Specifically, for timestep t , we first concatenate the aggregated segment representations with

Algorithm 1: The training of Spatio-Temporal Attention Networks (STAN) architecture

- 1: Given the number of streams M and maximum training iteration I .
 - 2: **for** $i = 1$ to I **do**
 - 3: Fetch input batch with sampled videos, each consisting of T segments.
 - 4: **for** each video in the batch **do**
 - 5: **for** $t = 1$ to T **do**
 - 6: **for** $m = 1$ to M **do**
 - 7: Extract local descriptors $\mathbf{F}^{m,t}$ from t -th segment in m -th stream.
 - 8: Feed $\mathbf{F}^{m,t}$ into spatial *AttCell* and estimate the corresponding spatial attention score map $\alpha^{m,t}$ via Eq. (4).
 - 9: Compute the spatially attended segment representation $\mathbf{f}_{att}^{m,t}$ via Eq. (5).
 - 10: **end for**
 - 11: Concatenate the spatially attended segment representations from all M streams to obtain a global segment representation \mathbf{f}_{att}^t in Eq. (6).
 - 12: **end for**
 - 13: Feed the sequence of global segment representations into LSTM to model temporal dynamics.
 - 14: Feed the output sequence of LSTM into temporal *AttCell* and estimate the corresponding temporal attention score map β via Eq. (7).
 - 15: Compute the final video representation \mathbf{f}_{att} via Eq. (8).
 - 16: **end for**
 - 17: Measure the recognition softmax loss and update the network by stochastic gradient descent.
 - 18: **end for**
-

spatial attention from each stream as a global representation of each video segment, which is given by

$$\mathbf{f}_{att}^t = [\mathbf{f}_{att}^{1,t}, \dots, \mathbf{f}_{att}^{m,t}, \dots, \mathbf{f}_{att}^{M,t}]. \quad (6)$$

\mathbf{f}_{att}^t is denoted as the input \mathbf{x}^t in Figure 3. Given the input \mathbf{x}^t and output vector \mathbf{h}^{t-1} at the previous timestep, the LSTM unit updates its parameters as described in [10] and readers can refer to [10] for technical details.

With the output sequence from LSTM ($\mathbf{h}^1, \mathbf{h}^2, \dots, \mathbf{h}^T$), we utilize temporal *AttCell* over all T segments to estimate the temporal attention score map $\beta = [\beta^t]$ as

$$\beta = \text{Softmax}(\mathbf{k}) \text{ and } k^t = g_{temporal}(\mathbf{h}^t; \theta_{temporal}), \quad (7)$$

where $g_{temporal}(\cdot)$ is the temporal attention network parameterized by $\theta_{temporal}$. It is worth noting that different from the spatial attention which is tailored for each stream, the temporal attention is learned based on the concatenated video segment representations from all modalities to seek a global and holistic guidance across the modalities. Then, based on the temporal

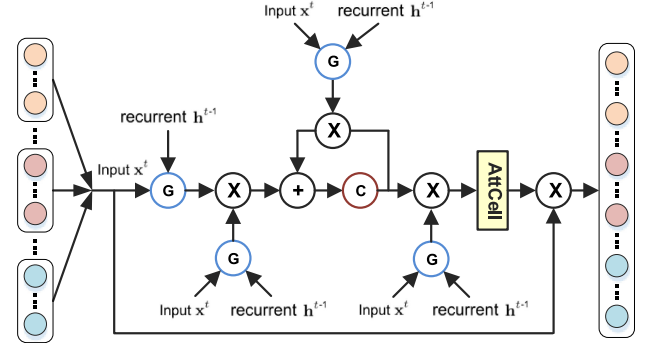


Fig. 3. The process of computing temporal attention for each segment and fusing the representations of all video segments to generate video representation.

attention distribution β , we calculate the weighted sum of the concatenated video segment representations of all T segments and compute the final video representation \mathbf{f}_{att} as

$$\mathbf{f}_{att} = \sum_{t=1}^T \beta^t \mathbf{f}_{att}^t. \quad (8)$$

Finally, the video representation \mathbf{f}_{att} is fed into a softmax layer for action recognition. The entire training process of our STAN is shown in Algorithm 1.

IV. IMPLEMENTATIONS

The proposed spatio-temporal attention networks (STAN) can be trained in an end-to-end manner by the standard backpropagation, since every operation within the network is differentiable. Our implementations are mainly based on the publicly available Caffe toolbox [44].

Network Architectures. In our experiments, STAN consists of three streams, i.e., frame stream, optical flow stream and clip stream. VGG-16 [45] is utilized for convolutional descriptors extraction in both frame and optical flow streams, which is a very deep 2D CNN with five groups of convolutional layers and demonstrates outstanding performance in image recognition. The VGG-16 networks are pre-trained on the ImageNet dataset [46], which consists of around 1.2 million labeled images from 1,000 different classes. Following [8], the frame stream operates on a single RGB frame and the optical flow stream takes a stack of 10 consecutive optical flow fields as the inputs. Here we exploit TVL1 optical flow algorithm [47] to extract optical flow and discretize the optical flow fields in the range of $[0, 255]$ by a linear transformation. For the clip stream, we capitalize on C3D [9] to extract local descriptors, which takes 16 consecutive frames as the inputs and is pre-trained on Sports-1M dataset [7] with 1.1 million sports videos. Given 224×224 frames as the inputs, the outputs of pool_5 layer of the three streams are all $7 \times 7 \times 512$.

Network Training. All additional layers at the top of pool_5 layer in the three streams are trained from scratch. Conv_6 layer contains 1,024 convolutional filters and the number of hidden states in LSTM is also set to 1,024. The LSTM weights are learned using the BPTT algorithm. A *softmax* layer is built on top

of the entire architecture to predict the action scores. Meanwhile, we add an extra dropout layer before the *softmax* layer to reduce over-fitting effect. The network weights are trained by stochastic gradient descent with 0.9 momentum and the mini-batch size is set to 128. The learning rate is initialized as 10^{-3} and decreases to its $\frac{1}{10}$ every 4,000 iterations. The whole training procedure stops at 15,000 iterations.

V. EXPERIMENTS

A. Datasets

We empirically and thoroughly evaluate our spatio-temporal attention networks (STAN) mainly on three public datasets: UCF101 [48], CCV [49] and THUMOS14 [50].

The **UCF101** dataset is one of the most popular action recognition benchmarks. It consists of 13,320 videos from 101 action categories. The action categories are divided into five groups: Human-Object Interaction, Body-Motion Only, Human-Human Interaction, Playing Musical Instruments, and Sports. Three training/testing splits are provided by the dataset organisers and each split includes about 9.5K training and 3.7K testing videos. Our experiments follow the standard evaluation scheme by using three training/testing splits and we report average accuracy over these splits.

The **Columbia Consumer Videos (CCV)** dataset is collected to stimulate research on consumer video analysis. Consumer videos are captured by ordinary consumers without any professional post-editing, which increases the difficulty of action recognition in videos. The dataset contains 9,317 YouTube videos with the annotation of 20 classes, most of which are events such as “basketball,” “parade” and “graduation ceremony.” We follow the convention and utilize a training set of 4,659 videos and a testing set of 4,658 videos. We adopt average precision (AP) of each class as the performance metric and the mean AP (mAP) over all the classes is reported as the overall measure.

The **THUMOS14** dataset is proposed for THUMOS Challenge 2014. It contains 13,320 videos for training, 1,010 videos for validation, and 1,574 videos for testing. Unlike UCF101, the videos in validation and testing set of THUMOS14 are untrimmed. The fact that there are many irrelevant frames has made the training and testing more challenging. We employ training and validation datasets to train our CNNs. The official evaluation tool is utilized to validate the performance. Following the standard setup of this dataset, we report the mean Average Precision (mAP) on testing set.

The three datasets possess very different characteristics. Although there are fewer semantic categories in CCV than UCF101, the average video duration in CCV is approximately 80 seconds, which is about 10 times longer than that of UCF101. Moreover, CCV and THUMOS14 have more intra-class variations than UCF101. Testing on all three datasets can be more comprehensive for verifying both the effectiveness and the generalization capability of our networks. Following the standard settings in [25], [51] at test time, we uniformly sample 25 RGB frames (or optical flow stacks or clips) per video in UCF101 and CCV datasets. As testing videos in THUMOS14 are untrimmed

and have long durations, we follow [52] and sample one single frame every 30 frames.

B. Evaluation of STAN on Single Modality

To validate the effectiveness of the proposed framework, we first examine our spatio-temporal attention on single modalities, i.e., frame, optical flow or clip, and compare it with the following baseline methods: (1) fc_6 . The output of the fc_6 layer is utilized as representation of each video segment. The video representation is then produced by average pooling over the representation of all segments. This is one of the most widely used video representations [10], [22], [53]. (2) $pool_5_25088$. The outputs of the last pooling layer are flattened and concatenated into a super vector as segment-level representation, whose dimension is 25,088 ($7 \times 7 \times 512$). All of these super vectors are then averaged to obtain the video representation. Obviously, the feature dimension is very high, and spatial information contained in $pool_5$ is ignored. (3) $pool_5_512+ave$. In this run, $pool_5$ layer is considered as 49 512-dimensional local descriptors and the segment representation is produced by average fusing the 49 descriptors. Similarly, the video representation is obtained by mean pooling all the segment representations. (4) $pool_5_512+VLAD$. It is proposed by Xu *et al.* in [24] as a discriminative CNN video representation. The only difference from $pool_5_512+ave$ is that all local descriptors are encoded by VLAD instead of average pooling. Specifically, in our experiments, the dimension of local descriptors is reduced to 256 by PCA and 128 components are used for k -means. Thus, the dimension of the final video representation is 32,768 (256×128). (5) Spatial Attention Networks (SAN). This is a variant of our STAN, which employs only spatial attention. (6) Spatio-Temporal Attention Networks (STAN) refers to our proposal. In all these baselines, we exploit pre-trained VGG-16/C3D as the basic CNN architecture and linear SVM [54] as the classifier.

The individual performances of each single modality on UCF101 (split1), CCV and THUMOS14 are summarized in Table I. Overall, the results across three datasets and three types of modalities consistently indicate that recognizing actions by exploring spatio-temporal attention leads to a performance boost. There is a performance gap among three runs fc_6 , $pool_5_25088$ and $pool_5_512+ave$. Although all three runs originate from $pool_5$, they are fundamentally different in the way of generating segment representations. The representation of fc_6 is a result of flattening all kernel maps in $pool_5$ to the neurons in a fully-connected layer, whereas $pool_5_25088$ and $pool_5_512+ave$ directly concatenate the local descriptors or average fuse them in $pool_5$ layer. As our results indicate, fc_6 can constantly lead to better performance than $pool_5_25088$ and $pool_5_512+ave$. It is not surprising that the performance of $pool_5_25088$ is only slightly better than that of $pool_5_512+ave$ in most cases. This somewhat reveals the weakness of ignoring spatial information in $pool_5$ layer. Furthermore, SAN outperforms $pool_5_512+VLAD$. Although both runs involve the use of spatial information, the learning strategies are different. $pool_5_512+VLAD$ separates the process of local descriptor learning and encoding, making the representation learning

TABLE I
PERFORMANCES OF SINGLE MODALITY (FRAME, OPTICAL FLOW, OR CLIP) ON UCF101 (SPLIT1), CCV AND THUMOS14. (FRA: FRAME; OPF: OPTICAL FLOW)

Model	UCF101 (split1)			CCV			THUMOS14		
	FRA	OPF	CLIP	FRA	OPF	CLIP	FRA	OPF	CLIP
fc ₆	75.4	84.4	83.5	74.1	63.5	76.2	55.8	57.0	65.8
pool ₅ _25088	74.3	83.8	81.8	70.4	62.4	75.3	55.1	57.2	64.7
pool ₅ _512+ave	73.3	82.8	80.8	69.8	61.7	73.6	49.2	50.2	58.5
pool ₅ _512+VLAD	80.2	84.9	83.7	77.8	64.2	77.2	61.8	60.4	68.2
SAN	80.5	85.8	84.6	78.0	64.6	77.8	62.4	59.6	68.4
STAN	81.5	86.6	85.0	78.8	65.3	78.3	64.0	61.5	68.7

TABLE II
PERFORMANCES OF TWO MODALITIES (FRAME+OPTICAL FLOW) ON UCF101 (SPLIT1), CCV AND THUMOS14. (EF: EARLY FUSION; LF: LATE FUSION)

Model	UCF101 (split1)		CCV		THUMOS14	
	EF	LF	EF	LF	EF	LF
SAN	90.1	90.4	79.1	79.5	68.9	69.3
STAN	91.6	91.2	80.7	80.3	71.6	71.2

TABLE III
PERFORMANCES OF THREE MODALITIES ON UCF101 (SPLIT1), CCV AND THUMOS14. (EF: EARLY FUSION; LF: LATE FUSION)

Model	UCF101 (split1)		CCV		THUMOS14	
	EF	LF	EF	LF	EF	LF
SAN	92.0	92.4	81.4	81.7	74.1	74.5
STAN	93.2	92.7	82.9	82.3	76.7	76.1

sub-optimal as each step is optimized independently, whereas SAN learns the representation in an end-to-end deep architecture. More importantly, the dimension of SAN (1,024) is much lower than that of pool₅_512+VLAD (32,768). By additionally incorporating temporal attention, STAN exhibits better performance than SAN. Compared to UCF101 and CCV, our STAN is benefited more from the mechanism of temporal attention on THUMOS14 dataset since the untrimmed videos in THUMOS14 contain more irrelevant snippets. Therefore, using temporal attention is more effective to automatically highlight the most informative segments and rule out irrelevant segments such as static background or non-action poses.

C. Evaluation of STAN on Two Modalities

Next, we turn to verify the merit of our spatio-temporal attention on two modalities, i.e., frame and optical flow. Two combination strategies, i.e., early fusion (feature concatenation) and late fusion (weighted sum of score), are compared. The former concatenates the representation of each modality and then seeks a global consensus on the temporal attention, which is the method exploited in STAN, while the latter models the temporal attention on each modality separately and linearly fuses the prediction scores at the decision stage. Please note that no temporal attention is measured in SAN and thus we simply employ average pooling to fuse segment representation. Table II details the performance on three datasets, respectively. Interestingly, late fusion can constantly lead to better performance gain than early fusion across three datasets in SAN. In comparison, early fusion in STAN benefits from the mechanism of exploring the modality consensus in modeling the temporal attention and exhibits better performance than does late fusion in STAN.

D. Evaluation of STAN on Three Modalities

To further validate the high capability of our network, we then generalize our spatio-temporal attention networks to three modalities, i.e., frame, optical flow and clip. Following the

comparisons on two modalities, two fusion strategies, early fusion and late fusion, are applied to SAN and STAN. The performances are summarized in Table III. Similar to the observations on two modalities, early fusion consistently outperforms late fusion in STAN. The results again indicate the advantage of exploring consensus among all modalities to produce temporal attention.

Figure 4 illustrates both spatial and temporal attention on two video examples from the categories “high jump” and “long jump,” respectively. We sample ten segments in the video and one frame is selected to represent each segment in the top row. The white regions show the spatial locations on which our STAN focuses in each frame (second row), optical flow image (third row) and clip (fourth row), respectively. The brightness indicates the strength of the focus. We observe that all core objects or regions that infer the action in the video, e.g., person and high jump bar or person and sand pit, have high spatial attention. Furthermore, the temporal attention probability of each segment is given in the bottom row. Clearly, the temporal segments including crucial motion, e.g., jump over the bar or jump into the sand pit, are assigned high attention probabilities.¹

E. Comparison with State-of-the-Art

We compare our technique with several state-of-the-art techniques on UCF101, CCV and THUMOS14. For fair comparison with most state-of-the-art methods, we adopt the standard 10-crop argumentation method by cropping and flipping four corners and center of the frame in the testing stage. The final prediction score is obtained by averaging the scores across the crops. It is also worth noting that most of the recent works on UCF101 and CCV are based on two-stream architecture. Therefore, only the performances of frame, optical flow and the fusion of the two modalities from state-of-the-art models are reported on UCF101 and CCV. As shown in Table IV, our STAN consistently achieves the best performance on UCF101 across different

¹Updated results of [11] from [39].

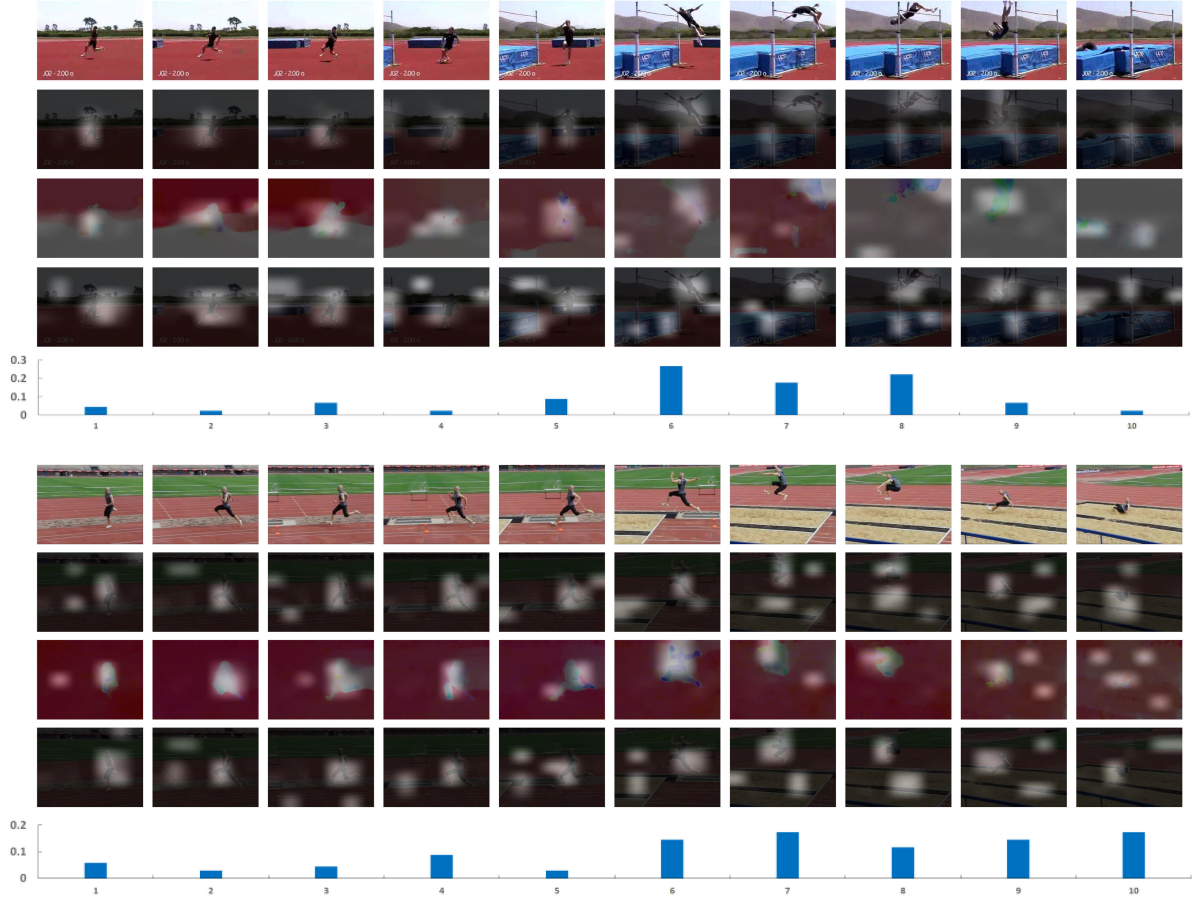


Fig. 4. Two video examples of the spatial and temporal attention learned by our STAN (top row: sampled segments from the video, where one frame is selected to represent each segment; second row: spatial attention map of each frame, where the brightness indicates the strength of focus; third row: spatial attention map of each optical flow, where the optical flow images are adjusted by color transitions for clearer showcase; fourth row: spatial attention map of each clip; bottom row: temporal attention probability of each segment). STAN can concentrate attention to the regions of interest, e.g., person and high jump bar or person and sand pit, which highly infer the action “high jump” or “long jump” in the video, respectively. Moreover, STAN can predict the contributions of different temporal segments for particular actions. For example, the segments with the motion of “jump over the bar” or “jump into the sand pit” contribute more to the action recognition.

TABLE IV
ACCURACY COMPARISONS ON UCF101 (ALL THREE SPLITS). (FRA: FRAME;
OPF: OPTICAL FLOW; TWO: FRA+OPF)

Model	FRA	OPF	TWO
Slow Fusion Network [7]	65.4	-	-
CNN-hidden6 [54]	79.3	-	-
MDI-end-to-end + Static-rgb [55]	76.9	-	-
Soft Attention [11] ¹	75.8	-	-
HM-AN [35]	78.2	-	-
Multi-granular Streams [10]	80.2	74.6	83.5
Two-stream + LSTM [31]	73.3	-	88.6
Two-stream [8]	72.6	83.6	87.6
LRCN [56]	71.1	77.0	82.9
Composite LSTM [32]	75.8	77.7	84.3
Siamese Network [57]	80.8	87.8	92.4
VideoLSTM [38]	79.6	82.1	88.9
Long-term Temporal ConvNet [58]	82.4	85.2	91.7
AdaScan [59]	78.6	83.4	89.4
HAN [39]	75.1	85.4	92.7
Two-stream Fusion [60]	-	-	92.5
Recurrent Attention [40]	-	-	92.5
STAN	82.8	88.2	92.8
STAN (FRA+OPF+CLIP)		93.6	

modalities. Specifically, on the modality of frame, compared to two-stream model [8], Composite LSTM [33] and Multi-granular Streams [10] performs better by exploring temporal dynamics with LSTM. On the modality of optical flow, [59] which utilizes 3D CNN to model long-term temporal information leads to better performance than [8]. Soft Attention [11] and VideoLSTM [39] which exploit spatio-temporal context to recurrently predict only spatial attention of the next frame and strengthen video representations, further improve LRCN [57] in which no attention mechanism is involved. HAN [40] and Recurrent Attention [41] by additionally mining hierarchical temporal structures and temporal attention are superior to Soft Attention and VideoLSTM particularly on the two streams of frame and optical flow, but the performances are still lower than STAN. The result indicates that STAN benefits from the holistic consensus of temporal attention across different modalities, and capable of distilling temporal information in the temporal attention to affect the recognition. It is also worth pointing out that as our STAN is a unified multi-stream framework, it can easily capitalize on the recently advanced networks to represent the modality in each stream.

TABLE V
MAP PERFORMANCES ON CCV. (FRA: FRAME; OPF: OPTICAL FLOW;
TWO: FRA+OPF)

Model	FRA	OPF	TWO
RADM [61]	63.0	-	-
SrMM+MDN+TP [62]	71.7	-	-
TGFL [63]	75.3	-	-
Two-stream CNN [64]	71.2	58.8	73.3
Spatial & Motion CNN [50]	75.0	59.1	75.8
STAN	79.2	65.5	81.4
STAN (FRA+OPF+CLIP)	83.7		

TABLE VI
MAP PERFORMANCES ON THUMOS14

Model	mAP
iDT+FV [3]	63.1
MAEs [65]	65.4
Apearance+Motion [22]	70.8
Two-stream [8]	66.1
RGB+EMV [51]	61.5
Objects+Motion [66]	71.6
STAN (FRA+OPF)	72.4
STAN (FRA+OPF+CLIP)	77.3

The performance comparisons on CCV are shown in Table V. Overall, there is a performance gap between six deep learning-based methods and one hand-crafted feature-based model, i.e., RADM [62]. Moreover, the results constantly indicate that our STAN outperforms others and exhibits an absolute improvement over the best competitor [51] by 4.2%, 6.4% and 5.6% on the frame, optical flow and the two modalities, respectively. By combining three modalities together, our final performance is boosted to 83.7%. Since most state-of-the-art methods on THUMOS14 still exploit hand-crafted descriptors (e.g., iDT [3]) instead of a deep architecture to capture the motion information, we only show the final performance of all the compared methods in Table VI. Similar to the observations on the other two datasets, STAN leads to a performance boost against the baselines, e.g., [22] which won first place in THUMOS Challenge 2014 and [67] which combines deep-learned object encodings with hand-crafted motion descriptors. The results again demonstrate the advantage of incorporating spatio-temporal attention to boost action recognition.

F. Evaluation of Parameters in STAN

To validate the effect of parameters involved in STAN towards recognition performance, we conduct experiments with two modalities (frame and optical flow) on UCF101 dataset.

Effect of size K of the convolutional feature map. We extract and compare the convolutional feature maps from different layers of VGG-16 backbone network, i.e., pool_5 ($K = 7$), $\text{conv}_{5,3}$ ($K = 14$), pool_4 ($K = 14$), and pool_3 ($K = 28$). The results are summarized in Table VII and performing STAN on feature map of pool_5 layer achieves the highest accuracy. This result is expected because the feature map of pool_5 layer has larger receptive fields and denotes higher-level concepts, which benefit the recognition.

TABLE VII
ACCURACY ON UCF101 WHEN STAN IS PERFORMED ON DIFFERENT
CONVOLUTIONAL FEATURE MAPS

Layer	pool_5	$\text{conv}_{5,3}$	pool_4	pool_3
K	7	14	14	28
Acc	92.8	91.2	89.1	85.3

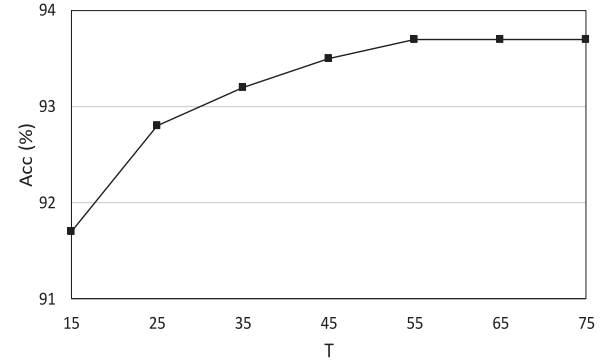


Fig. 5. Accuracy on UCF101 when leveraging different numbers (T) of segments in STAN.

Effect of number T of segments. In previous experiments, we follow most related works and set the number T to 25 for fair comparisons. Furthermore, we extract different numbers of frames per video in the inference to validate the number T of segments towards recognition accuracy. Figure 5 illustrates the accuracy of action recognition when leveraging different numbers of segments. As shown in the figure, the performance generally improves with an increase in number of segments. When the number exceeds a point (55 in this case), the accuracy changes smoothly. This result is also expected as only a few segments may not be sufficient to present the entire video, whereas a certain number of segments could already provide a good coverage.

G. Evaluation on Large-Scale Sports-1M Dataset

To examine the performance of STAN on large-scale datasets of action recognition, we further perform evaluations on large-scale Sports-1M dataset [7], which is one of the largest video classification benchmark. It contains approximately 1.13 million videos annotated with 487 Sports labels. There are 1K-3K videos per label. Please note that about 9.2% of the video URLs were dead when we downloaded the videos. Hence, we conducted experiments on the remaining 1.02 million videos and followed the official split, i.e., 70%, 10% and 20% for training, validation and test set, respectively. For simplicity, experiments are only implemented on single frame modality in this subsection.

Network Training. For efficient training on the large Sports-1M training set, we randomly sample 20 frames from each video as the network input. The settings of the network architecture and mini-batch are identical to those in Section IV. The learning rate is also initialized as 0.001 and divided by 10 after every 100 K iterations. The optimization will be complete after 250 K batches.

TABLE VIII
PERFORMANCE COMPARISONS ON SPORTS-1M DATASET

Model	Pre-train	Hit @1	Hit @5
Deep Video (Single Frame) [7]	ImageNet	59.3	77.7
Deep Video (Slow Fusion) [7]	ImageNet	60.9	80.2
C3D [9]	-	60.0	84.4
C3D [9]	I380K	61.1	85.2
Conv Pooling [31]	ImageNet	72.3	90.8
P3D ResNet [27]	ImageNet	66.4	87.4
VGG-16 [44]	ImageNet	61.3	83.9
STAN (Frame)	ImageNet	63.7	86.0

Network Testing. We evaluate the performance of STAN by measuring video classification accuracy on the test set. Specifically, we uniformly sample 20 frames from each video and adopt a single center crop per video, which is propagated through the network to obtain the video-level prediction score.

We compare the following approaches for performance evaluation: (1) Deep Video (Single Frame) and (Slow Fusion) [7]. To produce predictions for an entire video, the two methods randomly sample 20 clips and present each clip individually to the network. The former performs a CNN on one single frame from each clip to predict a clip-level score and averages individual clip-level predictions of each video to produce video-level predictions. (2) C3D [9] utilizes 3D convolutions on a clip volume to model the temporal information and the whole architecture could be trained on Sports-1M dataset from scratch or fine-tuned from the pre-trained model on I380 K internal dataset collected in [9]. (3) Conv Pooling [32] exploits max-pooling over the final convolutional layer of GoogleNet [68] across frames in each clip. (4) P3D ResNet [27] randomly samples 20 clips from each video and feeds each clip into the network to obtain a clip-level prediction score. The video-level score is computed by averaging all the clip-level scores of a video. (5) VGG-16 [45]. This run is similar with Deep Video (Single Frame), the only difference is that VGG-16 is adopted as backbone network in this run. The performances and comparisons are summarized in Table VIII. Overall, our STAN leads to a performance boost over VGG-16 (2D CNN) and C3D (3D CNN) by 2.4% and 2.6% in terms of top-1 video-level accuracy, respectively. The results indicate the advantage of exploring spatio-temporal attention for action recognition on large-scale video dataset. It is not surprise that the performance of STAN is lower than P3D ResNet [27] and Conv Pooling [32], since superior CNN architecture and/or more data are utilized in [27] and [32]. Specifically, [27] is built upon ResNet-152 backbone networks and takes 20 16-frame clips for each video into account. Instead, VGG-16 is exploited as backbone in our STAN and we only employ 20 single frames per video. When using ResNet-152 as our backbone, the top-1 accuracy of STAN could be improved from 63.7% to 65.8% and STAN is comparable to P3D ResNet in that case. Furthermore, [32] samples 240 video clips for each video and performs temporal pooling on 120 frames per video clip with frame rate of 1 fps. As such, the duration of each video clip is over 120s. In contrast, we only select 20 frames for each video and capitalize on much less information, making our STAN with better generalization capability. Meanwhile, compared to [27] and [32], the computation cost of STAN is much lower.

TABLE IX
ACTION DETECTION RESULTS ON THUMOS14 TEST SET

IoU	0.5	0.4	0.3	0.2	0.1
iDT+CNN [69]	8.3	11.7	14.0	17.0	18.2
motion+context [70]	14.4	20.8	27.0	33.6	36.6
LAF [71]	4.4	5.2	8.5	11.0	12.4
UntrimmedNet [72]	13.7	21.1	28.2	37.7	44.4
STAN	14.5	22.3	30.9	38.0	44.7

H. Evaluation of Temporal Attention on Action Detection

As described in Section III-D, our STAN not only outputs a video-level recognition score, but also produces a temporal attention score for each video segment. Naturally, this attention score could also be exploited for temporal action detection in untrimmed videos. Here we report the mean average precision (mAP) performance of action detection for different intersection over union (IoU) values on THUMOS14 dataset, where the detection task is limited to 20 classes (compared to 101 classes for the recognition task). As this section focuses on temporal detection in untrimmed videos, we exploit the untrimmed validation data (1,010 videos) to train our model and the test data (1,574 videos) to evaluate the performance. Please note that THUMOS14 dataset also provides temporal annotations of action instances in validation data, but we do not utilize these specific temporal annotations when training our networks and only capitalize on the video-level category information. As a result, the learning of our STAN for action detection is in a weakly-supervised manner.

For more precise localization, we sample a video segment every 5 frames, which is then fed into STAN to obtain the aggregated segment representation and compute the temporal attention score via Eq. (5) and Eq. (7), respectively. Then, the aggregated segment representation is fed into a separate classifier to obtain the classification score of the current segment. Considering that the differences of temporal attention scores between action instances and background segments will decrease through the normalization in the *softmax* operation, we employ the original attention score k^t here instead of the *softmax* attention score β^t in Eq. (7). To generate temporal proposals, we strictly follow the implementations in [72] and first remove the segments whose temporal attention scores are lower than the bar of 0.01. After that, we apply thresholding to all remaining segments and select the segments with classification scores above 0.4. The temporally consecutive segments in the selected ones are then grouped to form the final proposals. Naturally, the method could generate multiple proposals in this way. In addition, the score of each proposal is the weighted summation of the classification score of each segment within the proposal, and the weight is the temporal attention score.

We compare our detection results with those of several state-of-the-art methods in Table IX. The first two methods are the winners of THUMOS Challenge 2014, which are both fully-supervised approaches. The other two methods are recently published approaches based on weakly-supervised learning. Our method clearly leads to better performance than the weakly supervised baselines and is even competitive with

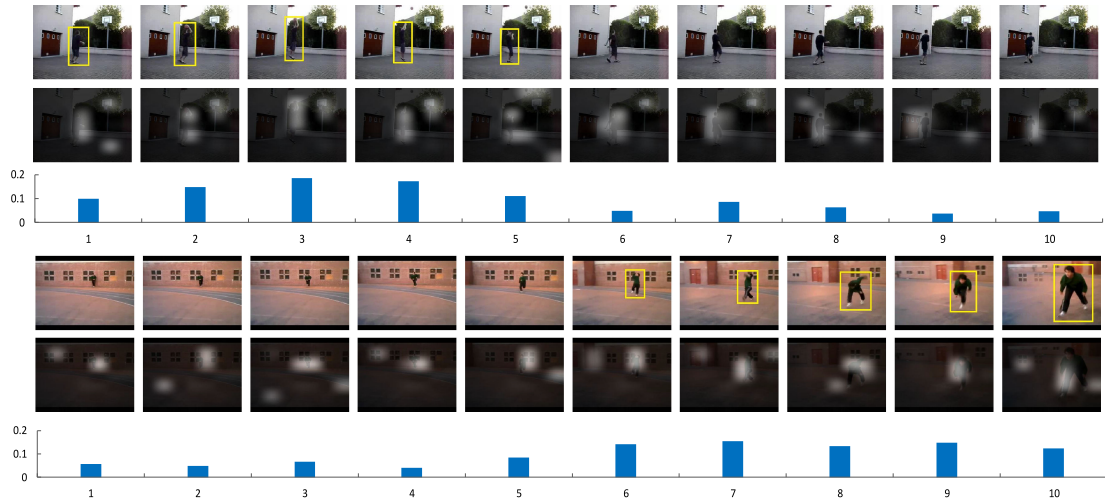


Fig. 6. Two video examples from the categories “Basketball Dunk” and “Cricket Bowling” in THUMOS14 dataset, which showcases the learnt attention by our STAN on spatio-temporal action detection task. Top row: the frame selected from each segment with its ground-truth bounding box, middle row: spatial attention by averaging three attention maps from three modalities, and bottom row: temporal attention.

fully supervised approaches across different IoUs. As indicated by the results, the learned temporal attention is also helpful in the action detection task. We could certainly boost the performance by further replacing the simple detection strategy with more advanced ones, which will be one of our future works on video action detection task.

I. Evaluation of Spatio-Temporal Attention on Action Detection

To examine our STAN on spatio-temporal action detection, Figure 6 illustrates two video examples from “Basketball Dunk” and “Cricket Bowling” in THUMOS14 dataset. The figure includes the frame selected from each segment with its ground-truth bounding box (top row), spatial attention by averaging three attention maps from three modalities (middle row), and temporal attention (bottom row). As shown in the figure, the learnt spatial attention is always well aligned with the ground-truth bounding boxes in the frames of both video examples. Meanwhile, the temporal segments in which the action happens receive higher temporal attention scores. Such results basically validate our STAN on the spatio-temporal action detection task.

VI. CONCLUSIONS

We have presented a Unified Spatio-Temporal Attention Networks (STAN) which explores both spatial and temporal attention to enhance action recognition in videos. In particular, we study the problem in the context of multiple modalities. To verify our claim, we devise an attention neural cell to learn the spatial attention on each modality and exploit the attention to pool local descriptors from one convolutional layer in CNN as video segment representation. Then, the representations on different modalities are concatenated and fed into LSTM to model the temporal attention across modalities, which is incorporated as a priori to holistically fuse all segment representations to a video representation for recognition. Experiments conducted on four datasets validate our proposal and analysis. The performance improvements are clearly observed compared to other action

recognition techniques, and more remarkably, the merit of our STAN is constantly proven when applying to different numbers of modalities. Our future works are as follows. First, we will extend our spatio-temporal attention to other modalities, e.g., audio. Second, the strategy of learning temporal attention could be explored by using RNN in an encoder-decoder manner.

REFERENCES

- [1] I. Laptev, “On space-time interest points,” *Int. J. Comput. Vis.*, vol. 64, no. 2/3, pp. 107–123, 2005.
- [2] P. Scovanner, S. Ali, and M. Shah, “A 3-dimensional sift descriptor and its application to action recognition,” in *Proc. ACM Multimedia*, 2007, pp. 357–360.
- [3] H. Wang and C. Schmid, “Action recognition with improved trajectories,” in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 3551–3558.
- [4] M. Hasan and A. K. Roy-Chowdhury, “A continuous learning framework for activity recognition using deep hybrid feature models,” *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1909–1922, Nov. 2015.
- [5] Y.-G. Jiang, Q. Dai, T. Mei, Y. Rui, and S.-F. Chang, “Super fast event recognition in internet videos,” *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1174–1186, Aug. 2015.
- [6] W. Xu, Z. Miao, X.-P. Zhang, and Y. Tian, “A hierarchical spatio-temporal model for human activity recognition,” *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1494–1509, Jul. 2017.
- [7] A. Karpathy *et al.*, “Large-scale video classification with convolutional neural networks,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 1725–1732.
- [8] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 568–576.
- [9] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in *Proc. Int. Conf. Comput. Vision*, 2015, pp. 4489–4497.
- [10] Q. Li *et al.*, “Action recognition by learning deep multi-granular spatio-temporal video representation,” in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2016, pp. 159–166.
- [11] S. Sharma, R. Kiros, and R. Salakhutdinov, “Action recognition using visual attention,” in *Proc. Workshop Int. Conf. Learn. Represent.*, 2015.
- [12] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2008, pp. 1–8.
- [13] A. Klaser, M. Marszalek, and C. Schmid, “A spatio-temporal descriptor based on 3D-gradients,” in *Proc. Brit. Mach. Vision Conf.*, 2008, pp. 275–1–275–10.
- [14] G. Willems, T. Tuytelaars, and L. Van Gool, “An efficient dense and scale-invariant spatio-temporal interest point detector,” in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 650–663.

- [15] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE Int. Workshop Visual Surveillance Perform. Eval. Tracking Surveillance*, 2005, pp. 65–72.
- [16] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2011, pp. 3169–3176.
- [17] M. Jain, H. Jegou, and P. Boutheymy, "Better exploiting motion for better action recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 2555–2562.
- [18] Z. Zhou, F. Shi, and W. Wu, "Learning spatial and temporal extents of human actions for action detection," *IEEE Trans. Multimedia*, vol. 17, no. 4, pp. 512–525, Apr. 2015.
- [19] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.
- [20] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1775–1789, Oct. 2009.
- [21] L. Wang, Y. Qiao, and X. Tang, "Motionlets: Mid-level 3D parts for human motion recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 2674–2681.
- [22] M. Jain *et al.*, "University of Amsterdam at THUMOS Challenge 2014," in *Proc. THUMOS Challenge Workshop Eur. Conf. Comput. Vis.*, 2014.
- [23] Z. Qiu, Q. Li, T. Yao, T. Mei, and Y. Rui, "MSR ASIA MSM at THUMOS Challenge 2015," in *Proc. THUMOS Challenge Workshop Conf. Comput. Vis. Pattern Recognit.*, 2015.
- [24] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1798–1807.
- [25] Z. Qiu, T. Yao, and T. Mei, "Deep quantization: Encoding convolutional activations with deep generative model," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6759–6768.
- [26] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 4597–4605.
- [27] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 5534–5542.
- [28] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7445–7454.
- [29] C. Feichtenhofer, A. Pinz, and R. Wildes, "Spatiotemporal residual networks for video action recognition," in *Proc. Neural Inf. Process. Syst.*, 2016, pp. 3468–3476.
- [30] T. Yao *et al.*, "MSR ASIA MSM at ActivityNet Challenge 2017," in *Proc. ActivityNet Challenge Workshop Conf. Comput. Vision Pattern Recognit.*, 2017.
- [31] D. Li, Z. Qiu, Q. Dai, T. Yao, and T. Mei, "Recurrent tubelet proposal and recognition networks for action detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018.
- [32] J. Yue-Hei Ng *et al.*, "Beyond short snippets: Deep networks for video classification," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4694–4702.
- [33] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using LSTMs," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 843–852.
- [34] Q. Li *et al.*, "Learning hierarchical video representation for action recognition," *Int. J. Multimedia Inf. Retrieval*, vol. 6, no. 1, pp. 85–98, 2017.
- [35] B. Mahasseni and S. Todorovic, "Regularizing long short term memory with 3D human-skeleton sequences for action recognition," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3054–3062.
- [36] S. Yan, J. S. Smith, W. Lu, and B. Zhang, "Hierarchical multi-scale attention networks for action recognition," *Signal Process. Image Commun.*, vol. 61, pp. 73–84, 2018.
- [37] 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 Conf. Artif. Intell.*, 2017, pp. 4263–4270.
- [38] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention LSTM networks for 3D action recognition," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1647–1656.
- [39] Z. Li, K. Gavriluk, E. Gavves, M. Jain, and C. G. Snoek, "VideoLSTM convolves, attends and flows for action recognition," *Comput. Vis. Image Understanding*, vol. 166, pp. 41–50, 2018.
- [40] Y. Wang *et al.*, "Hierarchical attention network for action recognition in videos," 2016, *arXiv:1607.06416*.
- [41] 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.
- [42] G. A. Sigurdsson, S. Divvala, A. Farhadi, and A. Gupta, "Asynchronous temporal fields for action recognition," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 585–594.
- [43] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage CNNs," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1049–1058.
- [44] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Multimedia*, 2014, pp. 675–678.
- [45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1245–1258.
- [46] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
- [47] C. Zach, T. Pock, and H. Bischof, "A duality based approach for real-time TV-L1 optical flow," in *Proc. 29th DAGM Symp. Pattern Recognit.*, 2007, pp. 214–223.
- [48] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.
- [49] Y.-G. Jiang *et al.*, "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2011.
- [50] Y. Jiang *et al.*, "THUMOS Challenge: Action recognition with a large number of classes," in *THUMOS Challenge Workshop Eur. Conf. Comput. Vis.*, 2014.
- [51] Y. Sun *et al.*, "Exploiting objects with LSTMs for video categorization," in *Proc. ACM Multimedia*, 2016, pp. 142–146.
- [52] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, "Real-time action recognition with enhanced motion vector CNNs," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2718–2726.
- [53] Z. Wu, Y.-G. Jiang, J. Wang, J. Pu, and X. Xue, "Exploring inter-feature and inter-class relationships with deep neural networks for video classification," in *Proc. ACM Multimedia*, 2014, pp. 167–176.
- [54] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.
- [55] S. Zha, F. Luisier, W. Andrews, N. Srivastava, and R. Salakhutdinov, "Exploiting image-trained CNN architectures for unconstrained video classification," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 60–1–60–13.
- [56] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3034–3042.
- [57] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2625–2634.
- [58] X. Wang, A. Farhadi, and A. Gupta, "Actions ~ transformations," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2658–2667.
- [59] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1510–1517, Jun. 2018.
- [60] A. Kar, N. Rai, K. Sikka, and G. Sharma, "Adascan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3376–3385.
- [61] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1933–1941.
- [62] A. J. Ma and P. C. Yuen, "Reduced analytic dependency modeling: Robust fusion for visual recognition," *Int. J. Comput. Vis.*, vol. 109, no. 3, pp. 233–251, 2014.
- [63] M. Nagel *et al.*, "Event fisher vectors: Robust encoding visual diversity of visual streams," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 178–1–178–12.
- [64] Y. Pan *et al.*, "Learning deep intrinsic video representation by exploring temporal coherence and graph structure," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 3832–3838.
- [65] H. Ye *et al.*, "Evaluating two-stream CNN for video classification," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2015, pp. 435–442.
- [66] T. Lan, Y. Zhu, A. Roshan Zamir, and S. Savarese, "Action recognition by hierarchical mid-level action elements," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 4552–4560.

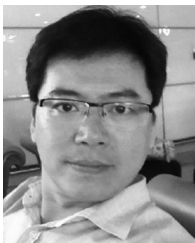
- [67] M. Jain, J. C. van Gemert, and C. G. Snoek, "What do 15,000 object categories tell us about classifying and localizing actions?" in *Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 46–55.
- [68] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
- [69] L. Wang, Y. Qiao, and X. Tang, "Action recognition and detection by combining motion and appearance features," in *Proc. THUMOS Challenge Workshop Eur. Conf. Comput. Vis.*, 2014.
- [70] D. Oneata, J. Verbeek, and C. Schmid, "The LEAR submission at THUMOS 2014," in *Proc. THUMOS Challenge Workshop Eur. Conf. Comput. Vis.*, 2014.
- [71] C. Sun, S. Shetty, R. Sukthankar, and R. Nevatia, "Temporal localization of fine-grained actions in videos by domain transfer from web images," in *Proc. ACM Multimedia*, 2015, pp. 371–380.
- [72] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "Untrimmednets for weakly supervised action recognition and detection," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4325–4334.



Dong Li received the B.E. degree in 2016 from the University of Science and Technology of China (USTC), Hefei, China. He is currently working toward the Ph.D. degree in the Department of Electronic Engineering and Information Science, USTC. He has participated in several large-scale video analysis competitions such as the ActivityNet Large Scale Activity Recognition Challenge 2017 and 2016. His research interests include large-scale video classification, action detection, and multimedia understanding.



Ting Yao received the Ph.D. degree in computer science from City University of Hong Kong, Kowloon, Hong Kong. He is currently a Researcher with the Multimedia Search and Mining group, Microsoft Research, Beijing, China. He is the principal designer of several top-performing multimedia analytic systems in worldwide competitions such as COCO image captioning, the ActivityNet Large Scale Activity Recognition Challenge 2017 and 2016, and THUMOS Action Recognition Challenge 2015. His research interests include video understanding, large-scale multimedia search, and deep learning. Dr. Yao is one of the organizers of the MSR Video to Language Challenge 2017 and 2016. For his contributions to Multimedia Search by Self, External and Crowdsourcing Knowledge, he was awarded the 2015 SIGMM Outstanding Ph.D. Thesis Award.



Ling-Yu Duan (M'06) received the Ph.D. degree in information technology from The University of Newcastle, Callaghan, NSW, Australia, in 2008. He is currently a Full Professor with the National Engineering Laboratory of Video Technology (NELVT), School of Electronics Engineering and Computer Science, Peking University (PKU), Beijing, China, and was the Associate Director of the Rapid-Rich Object Search Laboratory (ROSE), a joint lab between Nanyang Technological University (NTU), Singapore, and Peking University (PKU), China, since 2012. Before he joined PKU, he was a Research Scientist with the Institute for Infocomm Research (I2R), Singapore, from March 2003 to August 2008. His research interests include multimedia indexing, search, and retrieval, mobile visual search, visual feature coding, and video analytics, etc. Dr. Duan is the recipient of the *EURASIP Journal on Image and Video Processing* Best Paper Award in 2015, the Ministry of Education Technology Invention Award (First Prize) in 2016, the National Technology Invention Award (Second Prize) in 2017, the China Patent Award for Excellence (2017), and the National Information Technology Standardization Technical Committee "Standardization Work Outstanding Person" Award in 2015. He was a Co-Editor of MPEG Compact Descriptor for Visual Search (CDVS) Standard (ISO/IEC 15938-13), and is a Co-Chair of MPEG Compact Descriptor for Video Analytics (CDVA). He is currently an Associate Editor of *ACM Transactions on Intelligent Systems and Technology* and *ACM Transactions on Multimedia Computing, Communications, and Applications*.



Tao Mei (M'07–SM'11) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively. He is a Technical Vice President with JD.com, Beijing, China, and the Deputy Managing Director of JD AI Research, Beijing, China, where he is also the Director of Computer Vision and Multimedia Lab. Prior to joining JD.com in 2018, he was a Senior Research Manager with Microsoft Research Asia, Beijing, China, where he contributed 20 inventions and technologies to Microsoft's products and services. He has authored or co-authored more than 150 publications (with 11 best paper awards) and holds 20 US granted patents. He is or has been an Editorial Board Member of *IEEE TRANSACTIONS ON IMAGE PROCESSING*, *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*, *IEEE TRANSACTIONS ON MULTIMEDIA*, *ACM Transactions on Intelligent Systems and Technology*, and *ACM Transactions on Multimedia Computing, Communications, and Applications*. Dr. Mei is the General Co-chair of the IEEE ICME 2019, the Program Co-chair of ACM Multimedia 2018, the IEEE ICME 2015, and IEEE MMSP 2015. He was elected as a Fellow of IAPR and a Distinguished Scientist of ACM in 2016, for his contributions to large-scale video analysis and applications. He is also a Distinguished Industry Speaker of the IEEE Signal Processing Society.



Yong Rui (SM'04–F'10) received the B.S. degree from Southeast University, Nanjing, China, the M.S. degree from Tsinghua University, Beijing, China, and the Ph.D. degree from the University of Illinois at Urbana-Champaign, Champaign, IL, USA. He is currently the Chief Technology Officer and Senior Vice President of Lenovo Group, Beijing, China. He is responsible for overseeing Lenovo's corporate technical strategy, research and development directions, and Lenovo Research organization, which covers intelligent devices, big data analytics, artificial intelligence, cloud computing, 5G, and smart lifestyle-related technologies. He has authored 2 books, 12 book chapters, and 260 refereed journal and conference papers. With more than 22 000 citations, and an h-Index of 69, his publications are among the most referenced. He holds 62 issued U.S. and international patents. Dr. Rui is a recipient of many awards, including the 2016 IEEE Computer Society Technical Achievement Award, the 2016 IEEE Signal Processing Society Best Paper Award, and the 2010 Most Cited Paper of the Decade Award from the *Journal of Visual Communication and Image Representation*. He is an Associate Editor of the *ACM Transactions on Multimedia Computing, Communication and Applications*, and a founding Editor of the *International Journal of Multimedia Information Retrieval*. He was the Editor-in-Chief of the *IEEE Multimedia Magazine* (2013–2017), an Associate Editor of the *IEEE TRANSACTIONS ON MULTIMEDIA* (2004–2008), the *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGIES* (2006–2010), the *ACM/Springer Multimedia Systems Journal* (2004–2006), the *International Journal of Multimedia Tools and Applications* (2004–2006), and *IEEE ACCESS* (2013–2016). He is a Fellow of ACM, IAPR, and SPIE.