User-Ranking Video Summarization with Multi-Stage Spatio-Temporal Representation

Siyu Huang, Xi Li, Zhongfei Zhang, Fei Wu, and Junwei Han

***Abstract*—Video summarization is a challenging task, mainly due to the difficulties in learning complicated semantic structural relations between videos and summaries. In this paper, we present a novel supervised video summarization scheme based on three-stage deep neural networks. The scheme takes a divide-and-conquer strategy to resolve the complicated task of 3D video summarization into a set of easy and flexible computational subtasks, and then to sequentially perform 2D CNNs, 1D CNNs, and LSTM to address the subtasks in an hierarchical fashion. The hierarchical modeling of spatio-temporal structure leads to high performance and efficiency. In addition, we propose a simple but effective user-ranking method to cope with the labeling subjectivity problem of user-created video summarization, leading to the labeling quality refinement for robust supervised learning. Experimental results show that our approach outperforms the state-of-the-art video summarization methods on two benchmark datasets.**

***Index Terms*—Video summarization, recurrent neural network, convolutional neural network, multi-user inconsistency, user ranking.**

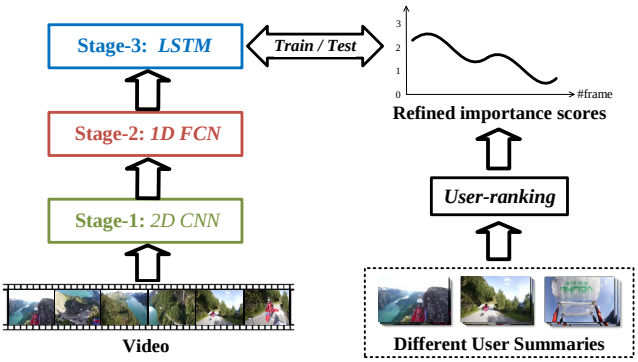


Fig. 1. **A brief illustration of our approach.** We propose a novel video summarization scheme based on three-stage deep neural networks consisting of 2D CNNs, 1D CNNs, and LSTM. The scheme learns the mapping between videos and frame-level importance scores in a hierarchical fashion. In addition, we propose a novel user-ranking method to refine ground truth importance scores by ranking the summary quality of users.

I. Introduction

As an important and challenging problem in computer vision, automatic video summarization aims at generating a short synopsis to extract the most informative parts of the video, which is crucial for human to effectively and efficiently browse and understand large amounts of video data in a user-friendly manner. Typically, it is formulated as a supervised learning problem that learns a spatio-temporal mapping function to select key frames or subshots from a video sequence, as illustrated in Fig. 1. Video summarization requires the modeling of the long-term temporal context, e.g., the topic or event, and also the short-term context, e.g., the motion and activity. Therefore, how to effectively perform spatio-temporal context modeling in an hierarchical fashion is a key issue to solve. Moreover, due to the subjective factors in annotating the ground truth by different persons, there exists a label inconsistency or bias problem, which may contaminate or confuse the supervised learning process to some degree. As a result, how to refine the label qualities for robust video summarization is another focus.

In general, the learning-based video summarization process is composed of two stages: 1) spatio-temporal video representation; and 2) key frame selection based sequence summarization. As a result, the difficulties in effective learning-based video summarization lie in the following three aspects: a) 3D video context modeling in both spatial and temporal dimensions; b) precise 1D key frame selection taking into account long-range and local-range temporal dependency; and c) joint modeling of video representation and sequence summarization.

Motivated by the above observations, we concentrate on designing a simple yet effective 3D deep learning scheme for

Manuscript received March 04, 2018; revised August 09 and November 07, 2018; accepted December 11, 2018. Date of publication XXX, 2019; date of current version XXX, 2019. This work is supported in part by NSFC (U1509206, 61472353, 61672456, and 61751209), Zhejiang Lab (2018EC0ZX01-2), the fundamental research funds for central universities in China (No. 2017FZA5007), the Key Program of Zhejiang Province, China (No. 2015C01027), ZJU Converging Media Computing Lab, Zhejiang Provincial Natural Science Foundation of China under Grant LR19F020004, the National Basic Research Program of China under Grant 2015CB352302, Zhejiang University K.P.Chao's High Technology Development Foundation, the funding from HIKVision, Artificial Intelligence Research Foundation of Baidu Inc., and Tencent AI Lab Rhino-Bird Joint Research Program(No. JR201806) The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Julian Fierrez. *(Corresponding author: Xi Li.)*

S. Huang is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: siyuhuang@zju.edu.cn).

X. Li is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China, and also with the Alibaba-Zhejiang University Joint Institute of Frontier Technologies, Hangzhou 310027, China. (email: xilizju@zju.edu.cn).

Z. Zhang is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China, and also with the Computer Science Department, Watson School, The State University of New York Binghamton University, Binghamton, NY 13902 USA (e-mail: zhongfei@zju.edu.cn).

F. Wu is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China (email: wufei@cs.zju.edu.cn).

J. Han is with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: junweihan2010@gmail.com).

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

effective 3D context modeling and precise temporal structure analysis. Specifically, we formulate 3D context modeling as a problem of sequentially performing 2D CNNs, 1D CNNs, and LSTM. In this way, the complicated task of 3D video summarization is reduced to a set of easy and flexible computational subtasks, where the main summarization job corresponds to 1D CNN+LSTM sequence learning for temporal structure modeling followed by standard fully-connected layers for final summarization determination. Technically, the 1D-CNNs capture the short-term temporal context in a local range around current frame, while the LSTMs capture the long-range context of the sequence. Both the long-range and local-range contexts are required in video summarization and can be well complementary to each other. In face of high-performance large-scale video processing, such a modeling architecture is advantageous in the following aspects: 1) easy implementation and high computational efficiency; 2) flexible CNN structure design and effective existing CNN model reuse instead of learning from scratch; 3) encoding the complicated and hierarchical semantic structure of video summarization along the temporal dimension.

In addition, in this paper we propose to address the labeling subjectivity problem of user-created video summarization. Usually, existing supervised video summarization methods directly adopt average user-annotated frame-level (or key-shots [1] preferred by majority users) importance scores [2] as ground truth. Since the annotations generated from different users are often inconsistent with each other, the aforementioned ground truth labeling strategy is likely to contaminate or inaccurately guide the above learning process. Motivated by this observation, we propose a simple but effective user-ranking method to evaluate the summary qualities for different users, resulting in more feasible and reliable ground truth for robust supervised learning.

In summary, the main contributions of this work are summarized as follows:

1) We present a novel video summarization scheme based on a hierarchical three-stage deep learning framework, which takes an effective divide-and-conquer strategy for 3D context modeling for feature representation as well as hierarchical temporal structure analysis for video summarization determination.
2) We propose a simple but effective user-ranking method to cope with the underlying subjectivity problems with multi-user ground truth annotations, leading to the labeling quality refinement for robust deep learning. To our knowledge, it is the first work to explore the multi-user annotation ambiguity in video summarization and refine the labeling quality for robust supervised learning.

II. Related Work

We first provide a brief survey of the video summarization methods proposed in recent literatures. Then, we discuss several related efforts on neural network based sequence modeling, especially the 1D convolution and LSTM, as they are the key components of our video summarization framework.

Video summarization: Researchers have proposed various unsupervised and supervised learning based video summarization algorithms [3–7]. Unsupervised approaches manually design intuitive criteria including representativeness [8–11], diversity [12–14] and importance [15, 16] to prioritize the frames. While supervised approaches [2, 17, 18] aim at learning subset selection models under supervision of human-created summaries such that they better align with how humans would summarize the video. In this paper we focus on the supervised approaches.

Prior work on supervised video summarization [18–21] formulates several hand-crafted criteria as submodulars and learns a combination function of them to approximate the human-created summaries. In the more recent literature, researchers completely discard the hand-crafted criteria such that they learn deep neural network models to directly pick out informative video frames. For instance, Yao *et al.* [1] propose a pairwise deep ranking model to learn the relations between high-light and non-highlight video segments. Zhang *et al.* [2] build deep regression model based on LSTM to infer frame-level importance scores annotated by humans.

From the perspective of context modeling, Zhang *et al.* [5] detect the local motion regions and build a dictionary of correlation feature graphs to model the interactions between motion regions. In contrast, our model is data-driven in which the spatial and temporal dependencies are learned by a hierarchical framework.

1D convolution and LSTM: In this work we address the problem of video summarization in view of temporal structure modeling based on 1D convolution and LSTM. The 1D convolutional architecture [22–24] is first proposed for sentence representation, and has achieved a great success on various computer vision tasks [25, 26]. With layer-by-layer composition and pooling, 1D convolutional architecture can hierarchically capture temporal context at different levels. The Long Short-Term Memory (LSTM) [27] is a specific recurrent neural network (RNN) architecture which is advantageous in modeling long-range temporal dependencies compared to conventional RNNs. Except for its prior success on various tasks [28–30], recently LSTM-based methods [2, 31–33] have achieved the state-of-the-art results on several benchmark video summarization datasets.

Nevertheless, video summarization still remains challenging as it needs to reveal complicated multi-level semantic structure along the temporal dimension. Although the combination of 1D-convs and LSTMs is rarely explored in computer vision community, it has been widely studied in other fields of sequential learning including text analysis [34–36], acoustics processing [37], and signal processing [38]. For instance, the C-LSTM model [34] utilizes CNNs to extract a sequence of higher-level phrase representations and feed them into LSTMs to obtain the sentence representation, capturing both local features of phrases as well as global sentence semantics. Similarly, from the perspective of sequence learning, we propose a novel three-stage deep learning scheme consisting of 2D CNNs, 1D fully convolutional networks (FCNs), and LSTM to jointly learn multi-level and multi-range spatio-temporal structural relations between videos and summaries, thus generating

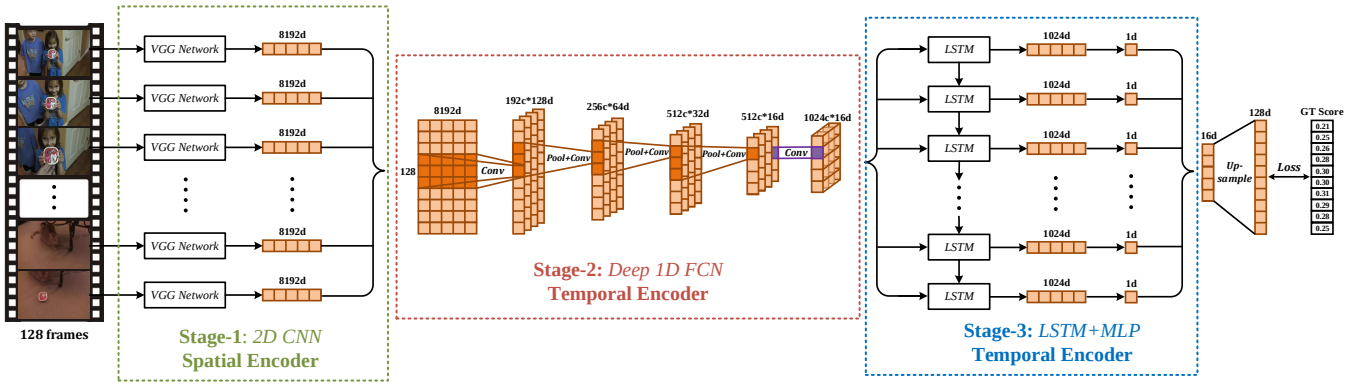


Fig. 2. **The architecture of our three-stage spatio-temporal network.** The input is a video clip consisting of 128 frames. Stage-1 extracts visual features of these frames using a two-stream VGG Network. Stage-2 employs deep 1D FCNs with a stack of 1D convolutional layers and 1D max-pooling layers to model multi-range temporal context. Stage-3 further employs an LSTM layer to capture the long-time dependencies and an MLP layer to output frame-level scores.

reasonable summaries which accord with human recognition.

III. Three-stage Video Summarization Framework

A. Problem formulation

Video summarization is a sequence subset selection problem, where the most informative subsets of a video are picked up for summary. To decide which subsets are valuable to keep, we formulate the problem as estimating the frame-level importance scores $\boldsymbol{y} = \{y_1, y_2, ..., y_T\}$ for all the corresponding video frames $\boldsymbol{x} = \{x_1, x_2, ..., x_T\}$, such that those subsets of the largest importance scores are picked up for summary.

The learning of mapping function $\boldsymbol{x} \rightarrow \boldsymbol{y}$ presents two key issues: 1) How to effectively model the complicated semantic structural relation between $\boldsymbol{x}$ and $\boldsymbol{y}$? 2) The summaries created by different users are largely inconsistent with each other due to the severe subjectivity of video summarization. Motivated by the first issue, we propose a three-stage video summarization scheme to learn the mapping function $\boldsymbol{x} \rightarrow \boldsymbol{y}$ based on hierarchical spatio-temporal modeling, as discussed in the following of Section III. Motivated by the second issue, we propose a user-ranking method to refine the label quality of $\boldsymbol{y}$ for supervised video summarization, as discussed in Section IV.

B. Three-stage video summarization scheme

In this paper, we propose a novel video summarization scheme to tackle the learning of $\boldsymbol{x} \rightarrow \boldsymbol{y}$ in three stages: (1) 2D CNNs, (2) 1D FCNs, and (3) LSTM, based on the observation that hierarchical spatial and temporal modeling is necessary for video analysis and summarization. More specifically, Stage-1 learns the visual representation $\boldsymbol{v}$ of input frames $\boldsymbol{x}$ based on 2D CNNs. Stage-2 uses 1D FCNs to model multi-range temporal context of $\boldsymbol{v}$ as $\boldsymbol{u}$. Stage-3 further uses LSTM to capture the long-time temporal dependencies in $\boldsymbol{u}$ to output the frame-level importance scores $\boldsymbol{y}$. The three-stage video summarization framework is formulated as

$$\boldsymbol{x} \xrightarrow{\text{S1: 2D CNNs}} \boldsymbol{v} \xrightarrow{\text{S2: 1D FCNs}} \boldsymbol{u} \xrightarrow{\text{S3: LSTM}} \boldsymbol{y}. \quad (1)$$

As illustrated in Fig. 2, the three stages are hierarchically learned in a unified framework. Based on the hierarchical spatio-temporal modeling, our framework is able to capture different types of temporal semantic structure, thus enabling good video summaries.

C. Hierarchical spatio-temporal modeling

2D CNNs: In Stage-1, we use 2D CNNs for visual representation extraction as $\boldsymbol{x} \xrightarrow{\text{Stage-1: 2D CNNs}} \boldsymbol{v}$, where $\boldsymbol{v} = \{v_1, v_2, ..., v_T\}$, T is the number of frames. The architecture of 2D CNNs is the two-stream VGG Network [39] proposed for video action recognition. The input to the two-stream network is respectively the RGB image and optical flow image of a frame x_t, and the output is the corresponding visual feature vector v_t, where $t = 1, 2, ..., T$.

1D FCNs: After representing the video frames as a sequence of visual feature vectors, video summarization is reformulated as a 1D sequence-to-sequence learning problem. It presents the challenge of modeling complicated semantic structure along the temporal dimension. Motivated by this, we address the problem under 1D scheme, and novelly leverage the 1D convolutional architecture into our neural network framework as a temporal encoder, in order to model deep temporal context of visual representation sequence $\boldsymbol{v}$, as $\boldsymbol{v} \xrightarrow{\text{Stage-2: 1D FCNs}} \boldsymbol{u}$.

Different from the commonly used 2D CNNs, 1D convolutional filters only slide on the temporal dimension of a vector sequence. Given sequence input $\boldsymbol{v}$, the feature map z of the l-th layer, f-th filter at location i is formulated as

$$z_i^{(l,f)} = \phi\left(\sum_{j=1}^{F_{l-1}} W^{(l,f,j)} z_i^{(l-1,j)} + b^{(l,f,j)}\right), \quad (2)$$

$z_i^{(l,f)} \in \mathbb{R}^{k \cdot D}$ denotes the segment at location i of the f-th feature map in layer l, where k is the convolutional filter width and D is the feature dimension. $W^{(l,f,j)}$ and $b^{(l,f,j)}$ are parameters of the convolutional filter connecting the j-th feature map in layer $l-1$ and the f-th feature map in layer l. ϕ is the activation function. F_l is the number of feature maps in layer l, such that $j = 1, 2, ..., F_{l-1}$ and

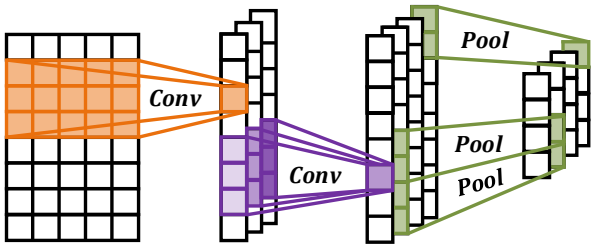


Fig. 3. 1D convolution and 1D max-pooling.

$f = 1, 2, ..., F_l$. The input to the 1D convolutional network is $\boldsymbol{v} = \{v_1, v_2, ..., v_t, ..., v_T\}$, forming the only feature map in layer 0, as

$$z_i^{(0,1)} = v_{i:i+k-1} \stackrel{\text{def}}{=} \begin{bmatrix} v_i \\ v_{i+1} \\ \vdots \\ v_{i+k-1} \end{bmatrix}. \tag{3}$$

An illustration of the 1D convolutional architecture is shown in Fig. 3. There is only one feature map in layer 0. Convolutional filters are operated along the first dimension of the feature map in a manner of sliding window. In deeper layers, there are multiple feature maps in a layer. The width of all the feature maps is equal to 1, and the filters work in a similar way. Our 1D FCNs are a stack of these 1D convolutional layers and max-pooling layers, such that the feature maps in deep layers can capture multi-level and multi-range semantic relations between units. The output of 1D FCNs is a sequence of vectors $\boldsymbol{u} = \{u_1, u_2, ..., u_t, ..., u_{T/d}\}$, where d is the downsampling ratio by max-pooling layers.

LSTM: In Stage-3, we further use LSTM to model the long-time dependencies within sequence $\boldsymbol{h}$ to better capture the summarizing criterion and estimate accurate frame-level importance scores $\boldsymbol{y}$, as $\boldsymbol{u} \xrightarrow{\text{Stage-3: LSTM}} \boldsymbol{y}$. LSTM [27] is a special kind of RNN which adopts memory cells to learn when to forget previous hidden states and when to update hidden states given new information. Specifically, in this work we employ LSTM architecture as described in [29] to map the spatio-temporal representation sequence $\boldsymbol{u} = \{u_1, u_2, ..., u_t, ..., u_{T/d}\}$ to a sequence of hidden states $\boldsymbol{h} = \{h_1, h_2, ..., h_t, ..., h_{T/d}\}$ as

$$\begin{aligned} i_t &= \sigma(W_{ui}u_t + W_{hi}h_{t-1} + b_i), \\ f_t &= \sigma(W_{uf}u_t + W_{hf}h_{t-1} + b_f), \\ o_t &= \sigma(W_{uo}u_t + W_{ho}h_{t-1} + b_o), \\ g_t &= \tanh(W_{ug}u_t + W_{hg}h_{t-1} + b_g), \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t, \\ h_t &= o_t \odot \tanh(c_t), \end{aligned} \tag{4}$$

where σ is the sigmoid nonlinearity function, and $\odot$ is the element-wise products. The core of LSTM is the memory cell $c_t \in \mathbb{R}^N$ which encodes the information of inputs up to time step t. c_t is a summation of the previous memory cell c_{t-1} modulated by forget gate $f_t \in \mathbb{R}^N$, and current input modulation gate $g_t \in \mathbb{R}^N$ modulated by input gate $i_t \in \mathbb{R}^N$. The output $h_t \in \mathbb{R}^N$ is a function of c_t modulated by output gate $o_t \in \mathbb{R}^N$. The three on/off knob gates i_t, f_t, o_t determine whether the LSTM keeps the values at the gates (1) or discard them (0), enabling the LSTM to learn complex and long-term temporal dynamics.

In theory, the LSTMs capture the long-term sequential information of the whole input sequence, e.g., the topic of a video. And, the 1D-convs capture the short-term temporal context in a local range around current frame, e.g., the motion or activity. Both these two types of temporal modeling are required in video summarization. The stack of 1D-convs and LSTMs enables a high temporal modeling performance for our video summarization model.

D. Summarization determination

We embed each h_t of LSTM's output $\boldsymbol{h}$ to a scalar value y_t using one layer of multi-layer perceptron (MLP). Due to the downsampling effect of max-pooling layers, we upsample $\boldsymbol{y} = \{y_1, y_2, ..., y_{T/d}\}$ to $\tilde{\boldsymbol{y}} = \{\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_T\}$ using bilinear filter, and $\tilde{\boldsymbol{y}}$ is the output frame-level importance scores. We use Euclidean loss L to measure the difference between estimated scores $\tilde{\boldsymbol{y}}$ and ground truth scores $\boldsymbol{y}_{\text{gt}}$:

$$L = \|\tilde{\boldsymbol{y}} - \boldsymbol{y}_{\text{gt}}\|_2^2. \tag{5}$$

Finally, we generate the summarization segments based on the frame-level importance scores. Since there is generally no ground-truth temporal segmentation provided by video summarization datasets, we follow [2] to split a video into a set of non-intersecting temporal segments by kernel temporal segmentation (KTS) [4]. Then, the segments with the largest importance scores are picked up for summary, while, their total duration is below a certain threshold l. Note that this is exactly the 0/1 knapsack problem and we solve it by dynamic programming [17]. The summary is then created by concatenating the selected segments in chronological order.

IV. User-Ranking

Due to the inherent subjectivity of video summarization, the video summaries created by different users are largely inconsistent with each other. To tackle the multi-user annotation ambiguity problem, we propose a simple but effective method, namely user-ranking, to refine the quality of multi-user ground truth annotations based on ranking the summary quality of users.

A. Analysis on user consistency

To measure the consistency of two summaries S_A and S_B for one video, we use harmonic mean F-score [2, 17] as

$$F(S_A, S_B) = \frac{2 \cdot P \cdot R}{P + R} \cdot 100\%, \tag{6}$$

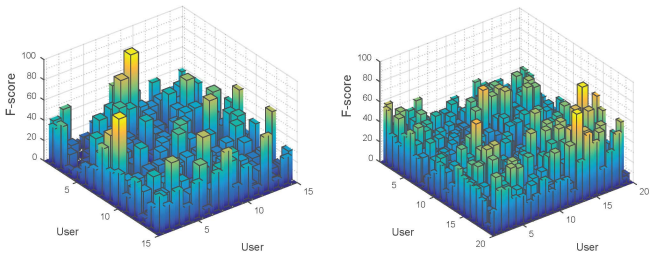
where

$$P = \frac{\text{length}(S_A \cap S_B)}{\text{length}(S_A)}, \quad R = \frac{\text{length}(S_A \cap S_B)}{\text{length}(S_B)}.$$

'length' denotes the time duration and $\cap$ denotes the temporal overlap. For a video summarization dataset of N videos,

TABLE I
USER CONSISTENCY AND USER QUALITY Q OF SUMME AND TVSUM DATASETS

	Baseline		Humans			User quality Q		
	Rand	Uni	Worst	Mean	Best	Min	Mean	Max
SumMe	18.7	15.4	17.9	31.1	40.9	0.53$\pm$0.21	1.0$\pm$0	1.33$\pm$0.14
TVSum	23.9	14.7	19.5	46.9	70.2	0.52$\pm$0.14	1.0$\pm$0	1.34$\pm$0.09



(a) (b)

Fig. 4. **Visualization of user-to-user consistency within two videos:** (a) 'Bearpark climbing' of SumMe dataset [17] with 15 user-created summaries, and (b) 'Will a cat eat dog food?' of TVSum dataset [10] with 20 user-created summaries. There is almost no user-to-user consistency larger than 80%, indicating the severe inconsistency among user-created summaries.

the user-to-user consistency C between two users i and j is estimated by

$$C(i,j) = \frac{1}{N}\sum_{n=1}^{N} F(S_i^n, S_j^n). \tag{7}$$

S_i^n is the summary of video n created by user $i \in 1, 2, ..., M$, where M is the number of users.

Fig. 4 illustrates the user-to-user consistency computed by C within two videos of two popular benchmark video summarization datasets SumMe [17] and TVSum [10]. since there is no explicit user-video pair given in these datasets, N is set to 1 for Eq. 7. Fig. 4 shows that almost no user-to-user consistency is larger than 0.8, and most of the consistencies are between 0 to 0.5.

The overall user consistency for SumMe and TVSum datasets is reported in the left part of Table I. Each evaluated summary is compared with the ground-truth summary averaged from summaries created by multiple users. For "**Baseline**", there is only one evaluated summary for a video, which is generated by a baseline method "Rand" or "Uni". The "Rand" baseline randomly creates summaries and the 'Uni' baseline creates summaries by uniformly sampling 2-second video clips along the timeline. For "**Humans**", there are M evaluated summaries created by M users on a video. They are compared with the ground-truth summary respectively. 'Worst', 'Mean', and 'Best' under "Humans" report the average of the worst, mean, and the best F-scores among user-created summaries. All the results shown in Table I are averaged over the videos.

Table I indicates that human annotators are largely inconsistent with each other. The performance of worst human-created summaries is close to that of the naive baseline methods. The F-scores of human's average level are respectively 31.1 and 46.9 on the two datasets, indicating that less than half of human-created summary fragments agree with the averaged preference of humans. All of the evidences reveal the challenge of severe subjectivity in video summarization task.

Algorithm 1: Estimation of summary qualities of users

Input: User-created summaries S
Output: User summary quality Q
1. Initialize Q with equivalent values;
2. **while** Q *is not in a steady state* **do**
   - Update Q using Eq. 8.

B. Ranking users

To tackle the problem of multi-user annotation ambiguity in video summarization, we propose a simple but effective user-ranking method. Specifically, we first rank the summary quality of individual users, and then refining the ground truth importance scores based on user quality.

To rank the user quality, we formulate an intuitive definition:

Definition 1 *A user owns high quality if he/she is consistent with most of the other users. A user owns higher quality if he/she is more consistent with the other high-quality users.*

On one hand, Def. 1 implies that the consistency between users is the fundamental metric for measuring user quality. The high-quality users share more similar summary preferences. On the other hand, Def. 1 implies that high-quality users have more influence in measuring other users' quality. It is inspired by PageRank[40] which lets a web-page have high rank if the sum of the ranks of its backlinks is high. In this case, PageRank propagates the influence of high-quality users and further encourages the high-quality summaries.

Based on Def. 1, we formulate the summarization quality Q of user i as

$$Q_i = \frac{1}{\|Z\|}\sum_{j=1, j\neq i}^{M} Q_j C(i,j). \tag{8}$$

Both the quality Q of the other users and the user-to-user consistency C jointly decide the quality of user i. The higher Q indicates the higher confidence level of video summaries created by the corresponding user. Z is the normalization factor which limits the mean value of Q to 1. As the elements in Q are calculated based on the other elements in it, we use a simple iterative method to update Q until the elements in Q form a dynamic equilibrium, namely, Q is in a steady state. The iterative update procedure is described in Algorithm 1.

The right part of Table I shows the minimum, mean, and the maximum user quality Q computed on the two datasets, where

TABLE II
VARIATIONS OF USER SUMMARY QUALITY Q

Iter.	Variations of Q	
	SumMe	TVSum
1	18.19%	17.89%
2	2.72%	2.24%
3	0.64%	0.42%

the quality of different users varies largely on both datasets. Table II shows the change of Q in the first three iterations. The iterative update method is efficient that Q could quickly converge to a steady state on both datasets.

C. Ground truth refinement

We create the weighted average frame-level importance scores $\hat{\boldsymbol{y}}_{\text{gt}}$ based on a set of scores $\boldsymbol{y}_{\text{user}}$ created by M users, as

$$\hat{\boldsymbol{y}}_{\text{gt}} = \sum_{i=1}^{M} \delta\left(\ln(Q_i)\right) \ln(Q_i) \boldsymbol{y}^i_{\text{user}}. \quad (9)$$

δ is the indicator function where $\delta(\theta)$=1 if θ>0 else $\delta(\theta)$=0. The generation of $\hat{\boldsymbol{y}}_{\text{gt}}$ only refers to the importance scores annotated by users of above-average level, for the goal of further reducing the noise in $\hat{\boldsymbol{y}}_{\text{gt}}$. The logarithm is used to further encourage the highest-quality users, for instance transforming user quality $Q \in [1.01, 1.34]$ into $\ln(Q) \in [0.01, 0.29]$ on TVSum dataset.

Our user-ranking approach refines the ground truth summary by paying more attention to high-quality user-created summaries and discarding low-quality summaries. Thus the model is learned under the guidance of preference of good users, leading to performance improvement by generating summaries closer to the most common preference. In experiments, $\hat{\boldsymbol{y}}_{\text{gt}}$ is demonstrated to improve the performance of video summarization model compared to its conventional version $\boldsymbol{y}_{\text{gt}}$.

V. EXPERIMENTS

A. Experimental Setup

Network architecture: Our three-stage spatio-temporal network is successively built upon 2D CNNs, 1D FCNs, and LSTM. The input to the network is 128 video frames and the output is 128 corresponding frame-level scores. The network architecture is $two\_stream\_vgg\_fc6$-$conv_1$(11,192)-$pool_1$(2)-*LRN*-$conv_{2\_1}$(5,256)-$conv_{2\_2}$(5,256)-$pool_2$(2)-$conv_{3\_1}$(3,512)-$conv_{3\_2}$(3,512)-$conv_{3\_3}$(3,512)-$pool_3$(2)-$conv_4$(1,1024)-$lstm$(1024,16)-mlp(1,16)-up(8). Numbers in the parentheses of $conv$ are respectively kernel size and number of channels. Each convolutional layer is followed by a ReLU activation function and is padded to keep the same as its last layer. The input and output of $lstm$ is 16 1024d vectors.

Implementation details: We use the Caffe toolbox [41] to implement the proposed framework. The network is trained for 10K iterations with a batch size of 40 and learning rate of 10^{-4}. Due to the effect of gradient vanishing for deep neural networks, we fix the parameters of two-stream VGG Network in the training phase, as the network is already pretrained on large-scale video datasets [39] thus providing good and robust visual features of video scenes. For kernel temporal segmentation, the minimal length of a shot segment is set as two seconds. The summary length ratio is set as 15%.

Dataset: We evaluate our method on two benchmark datasets of video summarization, including SumMe dataset [17] and TVSum dataset [10]. The SumMe dataset consists of 25 user videos covering holidays, events and sports. The TVSum dataset consists of 50 videos from YouTube in various genres, including news, documentaries, and user-generated content. Most of the videos are 1 to 5 minutes in length. Both datasets provide frame-level importance scores annotated by multiple users. Following [2, 17], we perform 5-fold cross validation on each dataset.

Evaluation metric: By following the convention of the existing literature [2, 17] for video summarization, we use the F-score as described in Eq. 6 to evaluate the performance of machine generated summary compared with human summary. As multiple summaries are possible for one video, we follow [2, 20] to compare the machine generated summary with all the human-annotated summaries and take the largest F-score as its performance.

TABLE III
COMPARING TO STATE-OF-THE-ART METHODS, INCLUDING PERFORMANCES ON TWO DATASETS AND NUMBERS OF LEARNABLE PARAMETERS (MILLION).

Method	SumMe	TVSum	Params
Video MMR, 2010 [3]	26.6	-	-
Super-frame, 2014 [17]	39.4	-	-
Submodular, 2015 [20]	39.7	-	-
DPP, 2016 [18]	40.9	-	-
Web-image prior, 2013 [8]	-	36.0	-
LiveLight, 2014 [14]	-	46.0	-
TVSum, 2015 [10]	-	50.0	-
ERSUM, 2017 [21]	43.1	59.4	-
MSDS-CC, 2018 [42]	40.6	52.3	-
vsLSTM, 2016 [2]	37.6	54.2	2.63
dppLSTM, 2016 [2]	38.6	54.7	2.63
SUM-GAN$_{\text{dpp}}$, 2017 [31]	39.1	51.7	295.86
SUM-GAN$_{\text{sup}}$, 2017 [31]	41.7	56.3	295.86
A-AVS, 2017 [33]	43.9	59.4	4.40
M-AVS, 2017 [33]	44.4	61.0	4.40
SASUM, 2018 [43]	40.6	53.9	44.07
SASUM$_{\text{sup}}$, 2018 [43]	45.3	58.2	44.07
DR-DSN, 2018 [44]	41.4	57.6	2.63
DR-DSN$_{\text{sup}}$, 2018 [44]	42.1	58.1	2.63
Ours	46.1	60.0	16.18
Ours+user ranking	**48.0**	**62.0**	16.18

B. Evaluation on video summarization

Comparing to state-of-the-art methods: Table III compares the performances of state-of-the-art methods and our method on SumMe dataset [17] and TVSum dataset [10]. The first section and the second section show results of conventional methods and deep neural network based methods, respectively. The third section show results of our method where 'Ours' is the result of our proposed three-stage video summarization

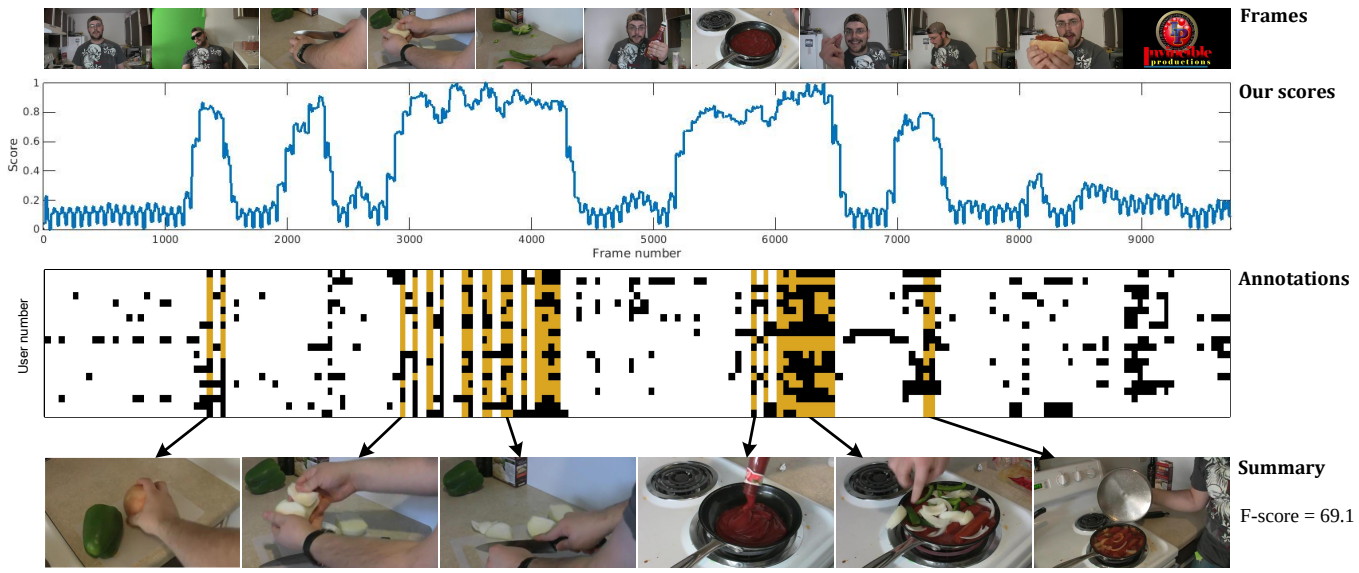


Fig. 5. Qualitative results of video 'Poor Man's Meals: Spicy Sausage Sandwich'. Four rows respectively show input video frames, frame-level scores estimated by our method, summaries created by 20 different users, and key frames of our summary.

scheme and 'Ours+user ranking' is the result of our model learned by refined ground truth labels. We collect the implementations of the state-of-the-art methods, replicate the models using PyTorch [45], and estimate their numbers of learnable parameters, as shown in the last column of Table III.

For the conventional methods, Video MMR [3] is the Video Maximal Marginal Relevance method which aims at rewarding relevant frames and penalizing redundant frames. Super-frame [17] is the super-frame interestingness estimation. Submodular is the submodular maximization model [20] which focuses on interestingness and representativeness of frames. DPP [18] is the determinantal point process which is a non-parametric summary transfer method. Web-image prior [8] is the unsupervised learning method using web-image based prior information. LiveLight model [14] focuses on importance and interestingness of frames. TVSum framework [10] uses video titles to find visually important shots. MSDS-CC [42] optimizes the frame selection in each individual view and regularizes the view-specific selections towards a consensus selection.

For the deep learning based methods, vsLSTM [2] uses LSTMs for sequence modeling and dppLSTM [2] combines LSTMs and DPP for unsupervised learning. SUM-GAN [31] is the LSTM-based generative adversarial networks (GAN). A-AVS and M-AVS [33] apply the attention mechanism to deep models for video summarization. ERSUM [21] learns a weighted combination of four properties including importance, representativeness, diversity, and storyness. SASUM [43] minimizes the distance between the generated text description of the video summary and the ground-truth text description. DR-DSN [44] builds a reinforcement learning framework in which the reward function accounts for diversity and representativeness.

Our method shows a good performance compared to the other state-of-the-art methods on both datasets. It demonstrates

TABLE IV
COMPARING TO HUMAN ANNOTATORS.

No.	Humans			Ours
	Worst	Mean	Best	
#1	17.9	38.2	60.9	**41.2**
#2	15.5	45.6	73.8	**60.2**
#3	18.8	59.7	85.5	**66.0**
#4	4.8	51.1	79.1	**53.4**
#5	22.9	44.2	61.9	**47.1**
#6	15.9	43.4	67.6	**53.5**
#7	27.6	44.5	62.8	**58.6**
#8	15.3	45.1	78.2	**47.0**
#9	19.3	50.6	76.4	**71.6**
#10	22.6	**47.2**	75.5	32.0
Ave	18.1	47.0	72.2	**53.1**

that the three-stage spatio-temporal network built with 2D CNNs, 1D CNNs, and LSTM can successfully generate reasonable video summaries. Technically, our method uses more types of temporal neural network models including 1D CNNs and LSTM to model different types of temporal semantic information, better revealing the complicated structural relations between videos and summaries. In addition, the proposed user-ranking method improves the performance on both datasets by 1.9 and 2.0. We will discuss it later in Section V-D.

Qualitative results: Fig. 5 shows some qualitative results of our method on video 'Poor Man's Meals: Spicy Sausage Sandwich' of TVSum dataset. The first row is several input video frames along the timeline. The second row shows the frame-level importance scores inferred by our method, where the larger score means more important a frame is. The third row shows the ground truth summaries created by 20 users, where black parts denote the frames selected by users and

TABLE V
PERFORMANCES OF DIFFERENT NETWORK ARCHITECTURES

Method	Network Architecture	F-score	Params (million)	Speed (ms)
MLP	3*MLP*	54.7±2.4	1.05	3.3
LSTM	*LSTM*+2*MLP*	54.3±3.2	38.81	30.0
	MLP+*LSTM*+2*MLP*	56.1±2.7	9.45	27.1
CNN	7*CNN*+2*MLP*	53.2±1.8	8.84	16.9
	MLP+7*CNN*+2*MLP*	57.7±2.5	12.05	8.7
combination	*LSTM*+7*CNN*+2*MLP*	57.5±1.4	42.46	33.4
	7*CNN*+*LSTM*+1*MLP*	**60.0**±3.1	16.18	17.5

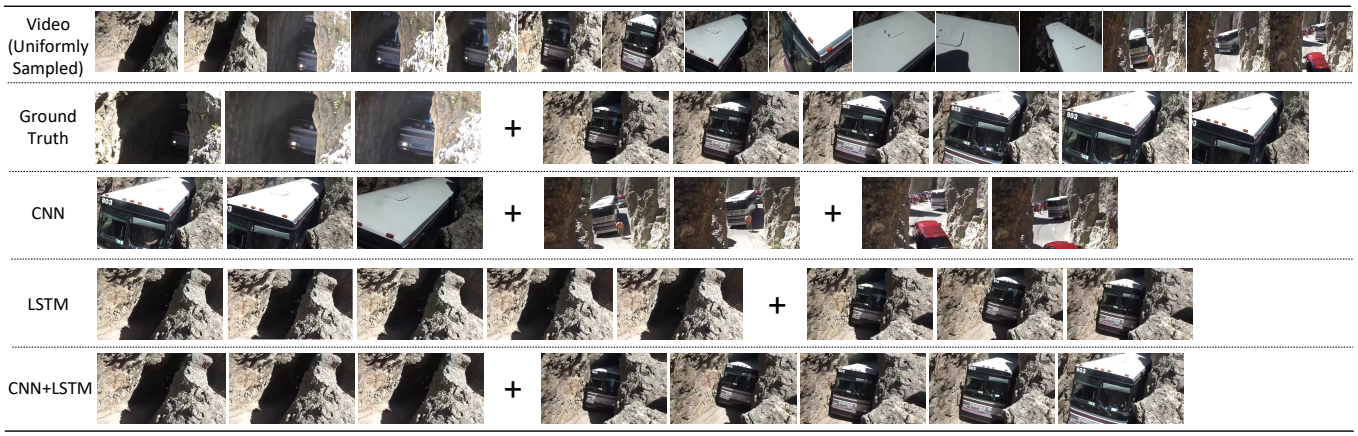


Fig. 6. Summaries generated by different networks on video 'Bus in Rock Tunnel'. The first row is the frames uniformly sampled on the original video. The second row is the ground truth summaries. The third, the forth, and the fifth rows show the summaries generated by CNN, LSTM, and CNN+LSTM, respectively.

yellow parts denote the frames selected by our method. Note that the summaries created by different users are largely inconsistent with each other, and the summary of our method is in accordance with most of the user summaries. The last row is several key frames of our summary clearly showing how to make a spicy sausage sandwich step by step, where the redundant frames are not included in the summary.

Comparing to Human Annotators: We further compare the performance of our method to human annotators on TVSum datasets. For each dataset, we randomly test 20% of the videos and the rest ones are used for training. Table IV reports the performance of summaries which are respectively created by human annotators and our method. In columns of 'Humans', 'Worst', 'Mean', and 'Best' respectively denote the worst, mean, and the best F-score of human-created summaries on a video. The column of 'Ours' is the performance of summary generated by our method. The bold numbers denote the best summaries except the 'Best' human annotations. Different from the evaluation setting used in Table III, in Table IV we compare a summary against the summary computed with averaged user scores.

Table IV shows the inconsistency among human annotators. The mean F-scores of human's average level is 47.0, indicating that less than half of human-created summary fragments accord with the averaged preference of humans. It reveals the severe subjectivity of video summarization task. Compared to human's average level, our method performs better on most videos. It indicates that our method gets close to the human recognition on video summarization to a certain degree.

C. Analysis on network architectures

In this work, different types of temporal models including CNN and LSTM are used for building a unified video summarization framework. We further analyze the network architectures in this section. We compare the 5-fold performance and the efficiency of several good try network architectures on TVSum dataset, as shown in Table V. The '3*MLP*' is the network of 3 MLP layers. The '*MLP*+*LSTM*+2*MLP*' and '*MLP*+7*CNN*+2*MLP*' respectively add an MLP layer before LSTM and CNN to reduce the input dimension of them. The rows of 'combination' are the results of combining LSTM and CNN together. **Params** is the size of models, and **Speed** is the time cost of testing 128 input frames. All the networks are implemented in a server with a Tesla K40c GPU, 8-core Intel i7-4790K CPU, and 32GB memory. The inputs of all the networks are frame-level visual features extracted by the fc6 layer of two-stream VGG network.

In Table V, CNN (53.2) and LSTM (54.3) show similar performance. Adding an MLP layer before CNN-based network and LSTM-based network significantly improves the performance and testing speed. It indicates that the first layer of 1D CNN and LSTM may better accept relatively low-dimensional input vectors. CNN+LSTM (60.0±3.1) outperforms CNN-only method (53.2±1.8) and LSTM-only method

(54.3±3.2) significantly considering the variance, demonstrating the temporal representation capability of stacked convolutions in video summarization. CNN+LSTM (60.0) outperforms CNN-only (53.2), showing the significance of LSTM in such an architecture combination. Conceptually, the CNN focuses on capturing the local temporal context, while, the LSTM is good at capturing the long-range information and the order of sequence, such that they are well complementary to each other in sequence encoding. Compared to the combination of LSTM+CNN (57.5), the combination of CNN+LSTM (60.0) performs much better, indicating that CNN is appropriate to be adopted before LSTM for video summarization. This is in line with the common perspective of speech recognition [37] and natural language processing [34].

Accordingly, in Fig. 6 we show the summarization results on the video 'Bus in Rock Tunnel' from SumMe dataset. The video shows a simple story of driving a bus out of the rock tunnel. The ground-truth summary consists of two video clips, including a short clip depicting the bus is in a tunnel and a longer clip depicting the procedure of bus's moving out. The summary of 1D CNN does not present the video background and only shows the video clips with large movements, due to the local temporal modeling by CNNs. The summary of LSTM is more similar to the ground-truth summary, while having a longer clip of the background which is redundant for users. It is because the LSTM focuses more on the global modeling. The summary of CNN+LSTM is more reasonable, which contains a short clip of the background and a longer clip of the bus's moving out. This example indicates that the CNN+LSTM could integrate the global and the local modeling.

Fig. 7 shows the key frames computed by different neural networks, where the F-score of each summary is shown at the end of each row. The frames denoted with purple bounding boxes are parts of ground truth summaries. The first example is the video 'When to Replace Your Tires GMC', describing how to check the tires. Although the three networks provide similar F-scores, their generated key frames are much different from each other. Most key frames of 1D CNN+LSTM are critical steps of checking a tire, while, other networks generate some redundant frames like the headline. The second video is 'Singapore Parkour Free Running', telling the story of how a boy runs to school with parkour skills. The 1D CNN+LSTM performs better than other networks by large margins, meanwhile, its key frames are more consistent with ground truth summary. Its second key frame shows an important storyline of the video, telling that the boy misses metro. Its third and forth key frame show dynamic and impressive scenes during parkour. These cases show that the combination of deep temporal models generates better summaries for videos of different themes.

D. Evaluation on user-ranking

In this section, we evaluate our proposed user-ranking method for video sumarization. For qualitative comparison, Fig. 8 illustrates the importance scores computed with different weighting methods, including constant-weighting, linear-weighting, and log-weighting (described in Eq. 9), on two

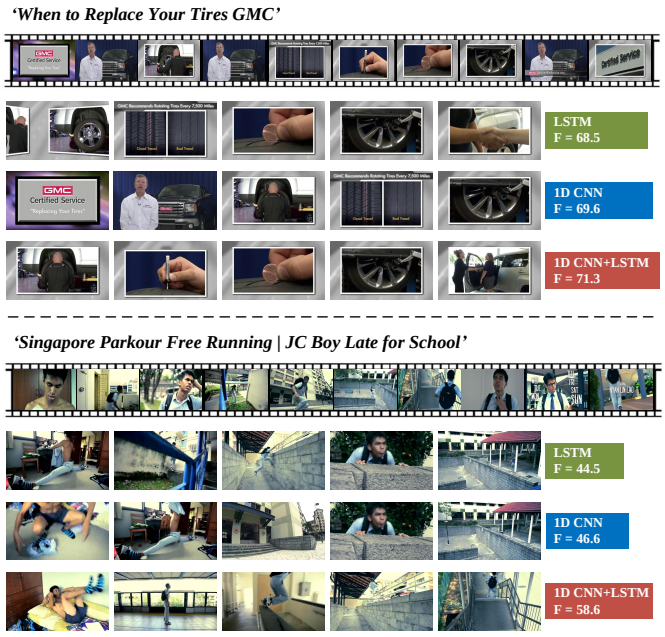


Fig. 7. Key frames extracted by LSTM, 1D CNN, and 1D CNN+LSTM, respectively. The F-scores are denoted at the right.

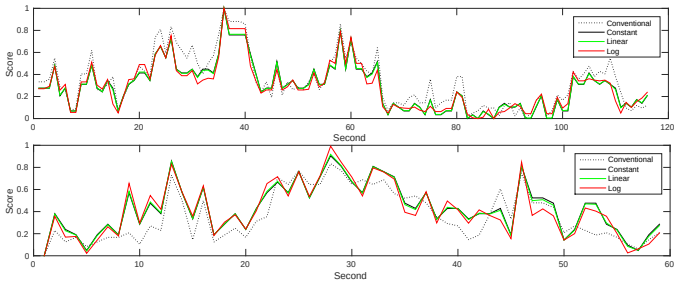


Fig. 8. Frame-level importance scores computed with different weighting methods on two videos, including conventional scores (black dotted), constant-weighted average scores (black solid), linear-weighted average scores (green), and log-weighted average scores (red).

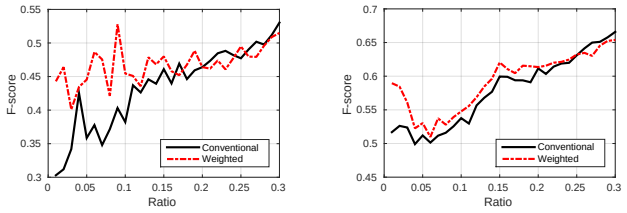
different videos. All of the weighted scores show significant difference from the conventional scores.

Table VI quantitatively compares the 5-fold performance of three versions of importance scores, including the conventional version, the linear-weighted version, and the log-weighted version on SumMe and TVSum datasets. The weighted average methods show better performances and smaller variance than the conventional method. It demonstrates that the refinement of ground truth labels is able to improve the performance of supervised video summarization models. It is also in line with our motivation of using user-ranking to refine the supervision labels by alleviating the inconsistency of multi-user annotations. The log-weighted average score shows the best mean performance, and, performing well on most of the folds, indicating that the log-weighting is probably a better weighting method in this case.

Fig. 9 compares the performance of conventional score and log-weighted average score with respect to different summary length ratios from 0 to 0.3. The log-weighted average score performs better for low length ratios (<0.2). It indicates

TABLE VI
EVALUATION ON WEIGHTING METHODS

Dataset	Weighting	#1	#2	#3	#4	#5	**Mean**
SumMe	conventional	45.6	44.4	50.1	42.7	47.6	46.1±2.9
	linear	45.4	44.5	50.7	49.0	44.7	46.9±2.8
	log	46.2	47.8	46.2	48.8	51.0	**48.0**±2.0
TVSum	conventional	59.6	60.3	58.1	56.9	64.9	60.0±3.1
	linear	60.6	61.3	58.7	57.2	57.7	59.1±1.8
	log	60.3	64.6	60.4	60.5	64.5	**62.0**±2.3
	log(-PR)	60.6	59.3	57.4	59.5	57.6	58.9±1.4



(a) SumMe (b) TVSum

Fig. 9. Comparison of conventional score and log-weighted score with respect to different summary length ratios on (a) SumMe dataset and (b) TVSum dataset. The refinement of ground truth labels makes models generate better short summaries.

that models based on user-ranking can generate better short summaries containing the highlights of videos, mainly because the weighted average score better remains the most important parts of summarization preference. The log-weighted average score performs slightly worse for long length ratios, possibly because the weighted averaging operation discards some useful information underlying in summaries of low quality users thus leading to less diverse contents.

VI. CONCLUSION

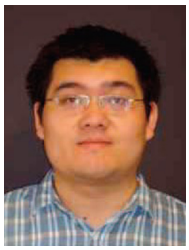
In this paper, we present a novel video summarization scheme based on three-stage deep neural networks. The scheme takes an effective divide-and-conquer strategy for spatio-temporal modeling and video summarization determination by sequentially performing 2D CNNs, 1D CNNs, and LSTM. In addition, we propose a simple yet effective user-ranking method to tackle the subjectivity problem of multi-user annotation in video summarization, resulting in more feasible and reliable ground truth for robust supervised learning. In experiments, our approach significantly outperforms the state-of-the-art video summarization methods.

REFERENCES

[1] T. Yao, T. Mei, and Y. Rui, "Highlight detection with pairwise deep ranking for first-person video summarization," in *Proc. IEEE Conf. CVPR*, 2016, pp. 982–990.
[2] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. ECCV*, 2016, pp. 766–782.
[3] Y. Li and B. Merialdo, "Multi-video summarization based on video-mmr," in *WIAMIS Workshop*, 2010.
[4] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *Proc. ECCV*, 2014, pp. 540–555.
[5] S. Zhang, Y. Zhu, and A. K. Roy-Chowdhury, "Context-aware surveillance video summarization." *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5469–5478, 2016.
[6] W.-S. Chu, Y. Song, and A. Jaimes, "Video co-summarization: Video summarization by visual co-occurrence," in *Proc. IEEE Conf. CVPR*, 2015, pp. 3584–3592.
[7] A. G. del Molino, C. Tan, J.-H. Lim, and A.-H. Tan, "Summarization of egocentric videos: A comprehensive survey," *IEEE Trans. on Human-Machine Systems*, vol. 47, no. 1, pp. 65–76, 2017.
[8] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan, "Large-scale video summarization using web-image priors," in *Proc. IEEE Conf. CVPR*, 2013, pp. 2698–2705.
[9] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Proc. IEEE Conf. CVPR*, 2013, pp. 2714–2721.
[10] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "Tvsum: Summarizing web videos using titles," in *Proc. IEEE Conf. CVPR*, 2015, pp. 5179–5187.
[11] X. Li, B. Zhao, and X. Lu, "Key frame extraction in the summary space," *IEEE Trans. Cybernetics*, vol. 48, no. 6, pp. 1923–1934, 2018.
[12] T. Liu and J. Kender, "Optimization algorithms for the selection of key frame sequences of variable length," in *Proc. ECCV*, 2006, pp. 301–305.
[13] S. E. F. De Avila, A. P. B. Lopes, A. da Luz, and A. de Albuquerque Araújo, "Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recogn. Lett.*, vol. 32, no. 1, pp. 56–68, 2011.
[14] B. Zhao and E. P. Xing, "Quasi real-time summarization for consumer videos," in *Proc. IEEE Conf. CVPR*, 2014, pp. 2513–2520.
[15] R. Hong, J. Tang, H.-K. Tan, S. Yan, C. Ngo, and T.-S. Chua, "Event driven summarization for web videos," in *SIGMM Workshop on Social Media*, 2009, pp. 43–48.
[16] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Proc. IEEE Conf. CVPR*, 2012, pp. 1346–1353.
[17] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Proc. ECCV*, 2014, pp. 505–520.
[18] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Summary transfer: Exemplar-based subset selection for video summarization," in *Proc. IEEE Conf. CVPR*, 2016, pp. 1059–1067.
[19] B. Gong, W.-L. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," in *Proc. Adv. NIPS*, 2014, pp. 2069–2077.
[20] M. Gygli, H. Grabner, and L. Van Gool, "Video summarization by learning submodular mixtures of objectives," in *Proc. IEEE Conf. CVPR*, 2015, pp. 3090–3098.
[21] X. Li, B. Zhao, and X. Lu, "A general framework for edited video and raw video summarization," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3652–3664, 2017.
[22] P. Blunsom, E. Grefenstette, and N. Kalchbrenner, "A convolutional neural network for modelling sentences," in *Proc. ACL*, 2014, pp. 655–665.
[23] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. EMNLP*, 2014, pp. 1746–1751.
[24] B. Hu, Z. Lu, H. Li, and Q. Chen, "Convolutional neural network architectures for matching natural language sentences," in *Proc. Adv. NIPS*, 2014, pp. 2042–2050.
[25] L. Ma, Z. Lu, and H. Li, "Learning to answer questions from image using convolutional neural network," in *Proc. AAAI*, 2016, pp. 3567–3573.
[26] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation

with pseudo-3d residual networks," in *Proc. IEEE ICCV*, 2017, pp. 5533–5541.
[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[28] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. Grue Simonsen, and J.-Y. Nie, "A hierarchical recurrent encoder-decoder for generative context-aware query suggestion," in *Proc. ACM CIKM*, 2015, pp. 553–562.
[29] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. CVPR*, 2015, pp. 2625–2634.
[30] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proc. IEEE Conf. CVPR*, 2015, pp. 4534–4542.
[31] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *Proc. IEEE Conf. CVPR*, 2017, pp. 202–211.
[32] B. Zhao, X. Li, and X. Lu, "Hierarchical recurrent neural network for video summarization," in *Proc. ACM Multimedia*, 2017, pp. 863–871.
[33] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder-decoder networks," *arXiv preprint arXiv:1708.09545*, 2017.
[34] C. Zhou, C. Sun, Z. Liu, and F. Lau, "A c-lstm neural network for text classification," *arXiv preprint arXiv:1511.08630*, 2015.
[35] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proc. Conf. EMNLP*, 2015, pp. 1422–1432.
[36] Y. Xiao and K. Cho, "Efficient character-level document classification by combining convolution and recurrent layers," *arXiv preprint arXiv:1602.00367*, 2016.
[37] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. IEEE ICASSP*, 2015, pp. 4580–4584.
[38] F. J. Ordez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.
[39] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream convnets," *arXiv preprint arXiv:1507.02159*, 2015.
[40] L. Page, "The pagerank citation ranking : Bringing order to the web," *Stanford Digital Libraries Working Paper*, vol. 9, no. 1, pp. 1–14, 1998.
[41] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Multimedia*, 2014, pp. 675–678.
[42] J. Meng, S. Wang, H. Wang, J. Yuan, and Y.-P. Tan, "Video summarization via multi-view representative selection," *IEEE Trans. Image Process.*, pp. 2134–2145, 2018.
[43] H. Wei, B. Ni, Y. Yan, H. Yu, X. Yang, and C. Yao, "Video summarization via semantic attended networks," in *Proc. AAAI*, 2018, pp. 216–223.
[44] K. Zhou and Y. Qiao, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proc. AAAI*, 2018, pp. 7582–7589.
[45] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

Siyu Huang is currently a fifth year PhD student in College of Information Science and Electronic Engineering at Zhejiang University, Hangzhou, China. His advisors are Prof. Zhongfei Zhang and Prof. Xi li. Earlier, he received his bachelor's degree in information and communication engineering from Zhejiang University, China, in 2014. His current research interests are primarily in computer vision and deep learning.



Xi Li is currently a full professor at the Zhejiang University, China. Prior to that, he was a senior researcher at the University of Adelaide, Australia. From 2009 to 2010, he worked as a postdoctoral researcher at CNRS Telecomd ParisTech, France. In 2009, he got the doctoral degree from National Laboratory of Pattern Recognition, Chinsese Academy of Sciences, Beijing, China. His research interests include visual tracking, motion analysis, face recognition, web data mining, image and video retrieval.

Zhongfei Zhang is a QiuShi Chaired Professor at Zhejiang University, China, and directs the Data Science and Engineering Research Center at the university while he is on leave from State University of New York (SUNY) at Binghamton, USA, where he is a professor at the Computer Science Department and directs the Multimedia Research Laboratory in the Department. He received a B.S. in Electronics Engineering (with Honors), an M.S. in Information Sciences, both from Zhejiang University, China, and a PhD in Computer Science from the University of Massachusetts at Amherst, USA. His research interests include knowledge discovery from multimedia data and relational data, multimedia information indexing and retrieval, and computer vision and pattern recognition.

Fei Wu received the B.S. degree from Lanzhou University, Lanzhou, Gansu, China, the M.S. degree from Macao University, Taipa, Macau, and the Ph.D. degree from Zhejiang University, Hangzhou, China. He is currently a Full Professor with the College of Computer Science and Technology, Zhejiang University. He was a Visiting Scholar with Prof. B. Yu's Group, University of California, Berkeley, from 2009 to 2010. His current research interests include multimedia retrieval, sparse representation, and machine learning.

Junwei Han received the Ph.D. degree in pattern recognition and intelligent systems from the School of Automation, Northwestern Polytechnical University, Xian, China, in 2003. He is currently a Professor with Northwestern Polytechnical University. His current research interests include multimedia processing and brain imaging analysis.

Prof. Han is an Associate Editor of the IEEE Transactions on Human-Machine Systems, Neurocomputing, and Multidimensional Systems and Signal Processing.