

# For Your Eyes Only: Learning to Summarize First-Person Videos

Hsuan-I Ho  
National Taiwan University  
b01901029@ntu.edu.tw

Wei-Chen Chiu  
National Chiao Tung University  
walon@cs.nctu.edu.tw

Yu-Chiang Frank Wang  
National Taiwan University  
ycwang@ntu.edu.tw

## Abstract

With the increasing amount of video data, it is desirable to highlight or summarize the videos of interest for viewing, search, or storage purposes. However, existing summarization approaches are typically trained from third-person videos, which cannot generalize to highlight the first-person ones. By advancing deep learning techniques, we propose a unique network architecture for transferring spatiotemporal information across video domains, which jointly solves metric-learning based feature embedding and keyframe selection via Bidirectional Long Short-Term Memory (BiLSTM). A practical semi-supervised learning setting is considered, i.e., only fully annotated third-person videos, unlabeled first-person videos, and a small amount of annotated first-person ones are required to train our proposed model. Qualitative and quantitative evaluations are performed in our experiments, which confirm that our model performs favorably against baseline and state-of-the-art approaches on first-person video summarization.

## 1. Introduction

Wearable and head-mounted cameras have changed the way how people record and browse the videos. These devices enable users to capture life-logging videos without intentionally focus on particular subjects. Thus, compared to third-person videos, first-person videos (or egocentric videos) would exhibit very unique content and properties. As pointed out by Molino *et al.* [4], the lack of sufficient structural and repetitive information for first-person videos would limit viewing quality, and thus it is necessary to perform summarization of such videos for improved user satisfaction.

With the goal of encapsulating most descriptive segments from raw videos, video summarization aims at solving the problem of long video browsing. Existing approaches for video summarization either select the most representative video segments/keyframes [6, 3], or detect particular or pre-defined visual structures or objects [27, 33]. With the progress and development of deep learning,

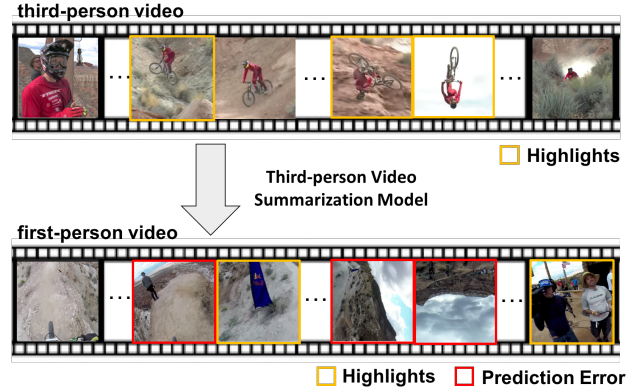


Figure 1. Example of models learned for summarizing a third-person video cannot be directly applied to summarize a first-person one due to significant changes in visual and temporal information, even if both videos describe the same type of activity.

a recent representative work [38] successfully applied deep neural networks for first-person video summarization. As expected, in order to achieve satisfactory performance, their approach requires a pre-collected and annotated first-person videos.

Since annotated first-person videos are difficult to obtain (the dataset in [38] is not publicly available), it thus becomes desirable to utilize annotated third-person videos to increase the training set size. However, as noted above, since significant visual appearance differences can be expected between third and first-person videos, existing third-person video summarization based approaches cannot be directly applied to first-person videos. Moreover, a satisfying first-person video summary should consist of segments important to both the viewer and the user (i.e., the recorder). Take Fig. 1 for example, a learned model for summarizing third-person videos might not generalize to first-person ones due to visual differences in view angle, appearance, or objects of interest.

The above problem is mainly due to the domain shift (or dataset bias) between first and third-person videos, which has been extensively studied in literature [28]. With the recent success of deep learning, architectures of deep neural

networks have also been utilized for alleviating the above problem [30]. Recent approaches like [39, 15] did not explicitly address this issue. Although the component of Long Short-Term Memory (LSTM) is introduced in their proposed models, their training and testing data are mixed with third and first-person videos. In other words, one cannot directly apply their methods for first-person video summarization.

Advancing the idea of domain adaptation for first-person video summarization, one could utilize fully annotated third-person videos plus a number of annotated first-person videos for designing the model. In order to increase the training set size for first-person videos without label information, one could further extend the above supervised domain adaptation setting to a semi-supervised one. That is, additional unlabeled first-person videos can be further utilized and included in the training set. As a result, we not only need to alleviate the dataset bias between first and third-person videos, we also need to perform semi-supervised learning in the above adaptation process.

To address first-person video summarization, we propose a unique deep learning framework integrating spatiotemporal feature embedding and recurrent neural networks (RNN). With the goal of transferring representative information across video domains while exploiting the video domain of interest, our proposed network architecture jointly performs domain adaptation (between first and third-person videos) and semi-supervised learning (within first-person videos). In particular, when performing semi-supervised learning within first-person videos, keyframe selection is introduced into our sequence-to-sequence learning formulation, which further exploits highlight information across video frames.

In summary, our contribution are threefold: 1) By reducing the semantics gap between third and first-person videos, our proposed network transfers informative spatiotemporal features to perform first-person video summarization; 2) By advancing a semi-supervised learning setting, we not only perform adaptation across video domains but also exploit the semantics within the data domain of interest; 3) we empirically verify that existing summarization approaches cannot generalize to highlight the first-person ones, while ours is able to produce satisfactory summarization results.

## 2. Related works

Video summarization has attracted the computer vision community over the past few years [4, 14, 26]. Most existing approaches follow a basic work flow consisting of (1) visual feature extraction and (2) keyframe selection or scene segmentation which is subject to several pre-defined criteria. For example, early works like [8, 7, 20] utilize measurements stemmed from attention [5, 24, 23] or representativeness [6, 3] heuristically upon low-level features, and

thus the resulting summarization is produced in an unsupervised manner.

In comparison to summarization approaches based on unsupervised learning, supervised video summarization methods utilize labeled training videos and aim to learn from how humans select key video segments. For instance, selection for summarization can be learned by extracting primary objects or leveraging context information [17, 18, 21], diversity [22, 40, 19], and data-driven representativeness [19, 11]. Other works like [38, 12] also use user-annotated importance scores to learn an associated feature embedding function via metric learning, followed by a classifier to score each video segment independently.

Some recent deep-learning based methods approach video summarization by solving a sequence-to-sequence problem, in which the video frames are encoded by Recurrent Neural Network (RNN) schemes. For example, Zhang *et al.* [39] propose a summarization model based on a bi-directional Long Short Term Memory (biLSTM) framework, which is trained on videos with annotated importance scores for keyframe selection. They additionally apply determinantal point process (DPP) to enhance the diversity of the chosen keyframes. Ji *et al.* [15] further extend such biLSTM models by integrating the attention mechanism. Their model considers temporal information in finer granularity when decoding the feature vectors of video segments generated by biLSTM.

Although supervised approaches exhibit promising video summarization results, existing datasets (with ground truth data) for video summarization typically are with smaller scales in comparison to those for object or face recognition [35, 9]. As a result, for each application domain of interest, it would be desirable to have a large amount of labeled videos for training purposes. Several research works thus attempt to utilize various techniques in order to address the issue coming from insufficient training data: Panda *et al.* [29] collect weakly annotated videos from YouTube 8M [2] and train their summarization model with auxiliary labels of activity classes; Sun *et al.* [36] train their highlight classifier with treating a collection of YouTube videos that have been edited as positive training data, in comparison to the negative ones from raw videos; Mahaseni *et al.* [25] predict video keyframes distribution with a sequential generative adversarial network. In the work of Li *et al.* [19], activity labels from the ADL dataset [31] are utilized to exploit a proper metric for evaluating the performance of summarization.

Nevertheless, the above approaches generally focus on summarizing third-person videos or a pre-collected dataset with mixed type of videos [9, 35]. They typically do not focus on highlighting the first-person ones, which are particularly challenging due to significant changes in visual content and appearances, plus the lack of sufficient amount of

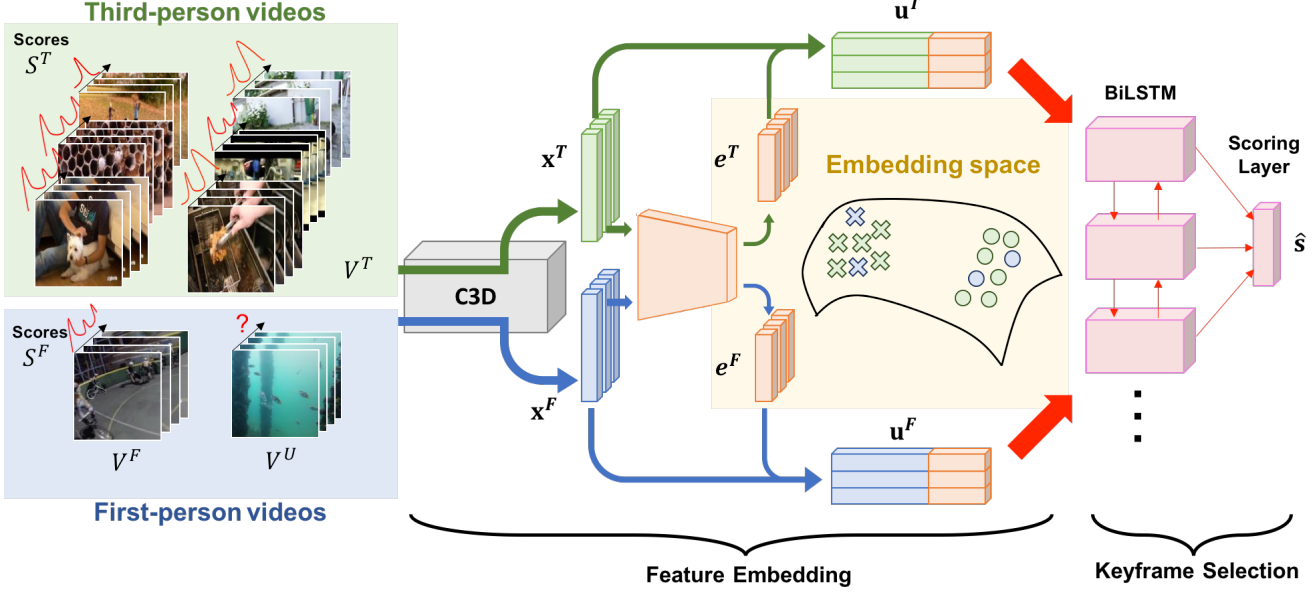


Figure 2. Our framework for first-person video summarization via solving semi-supervised domain adaptation. Note that fully annotated third-person videos  $V^T$  and unlabeled first-person videos  $V^U$  plus a number of annotated first-person ones  $V^F$  are presented during training. We have  $\mathbf{x}$ ,  $\mathbf{e}$ , and  $\mathbf{u}$  denote C3D, embedded and the resulting concatenated features in the feature embedding layers, respectively.

ground truth data. This is the reason why, as pointed out in Sect. 1, we choose to address first-person video summarization in a semi-supervised setting, in which most annotated data are from the third-person videos while only a small amount of them are first-person ones.

### 3. Proposed Method

To the best of our knowledge, we are the first to address the dataset bias between third and first-person videos, with the goal of realizing first-person video summarization in a cross-domain semi-supervised setting.

For the sake of completeness, we first introduce the notations in this paper. We have an annotated video collection including a set of third-person videos  $V^T = \{V_1^T, \dots, V_M^T\}$  and few first-person videos  $V^F = \{V_1^F, \dots, V_N^F\}$ . Their corresponding annotations (i.e., importance scores) at the frame-level are  $S^T = \{\mathbf{s}_1^T, \dots, \mathbf{s}_M^T\}$  and  $S^F = \{\mathbf{s}_1^F, \dots, \mathbf{s}_N^F\}$ , where  $M$  and  $N$  denote the numbers of third- and first-person videos respectively, with  $M \gg N$ . In addition, we have another set of first-person videos  $V^U = \{V_1^U, \dots, V_K^U\}$  without any annotation on importance score or keyframes, with the number of videos  $K$  larger than  $N$ . Our goal is to bridge the semantic gap between  $V^F$  and  $V^U$  for improving the capacity of our model for first-person video summarization.

The architecture of our proposed method is shown in Fig. 2, which consists of network components for feature embedding and sequential keyframe selection. Details of

our network will be described in the following subsections.

#### 3.1. Feature Embedding

A video may contain rich information by exploiting both of its temporal and spatial contents [38, 34]. While spatial contents such as objects and the context typically dominate videos summarized from third-person views, temporal information retrieved from camera motion and optical flow might reflect more representative information for first-person videos. Therefore, we utilize and extend the spatiotemporal feature extractor (i.e., C3D [37]) for improved representation of spatiotemporal video content. When using C3D, every video in  $V^T$  or  $V^F$  is split into fix-length segments to extract C3D visual features  $X_m^T = \{\mathbf{x}_{m1}^T, \dots, \mathbf{x}_{mt}^T\}$  or  $X_n^F = \{\mathbf{x}_{n1}^F, \dots, \mathbf{x}_{nt}^F\}$ , where  $X_m^T, X_n^F \in \mathbb{R}^{p \times t}$  ( $p$  represents the dimension of the C3D feature, and  $t$  denotes the number of segments in each video).

Unfortunately, there is no guarantee that C3D is able to capture video semantic information such as importance scores. By advancing the idea of metric learning, particularly the triplet network [13], we choose to learn an embedding representation of the C3D features not only for describing videos across third and first-person views but also reflect the summarization scores. More specifically, given C3D features  $\mathbf{x}$ , we aim to introduce network layers for deriving  $\mathbf{e}$  in a embedding space as  $f_E(\mathbf{x}) = \mathbf{e} \in \mathbb{R}^q$ , where  $f_E$  can be viewed as the embedding function and  $q$  is the size of the embedded feature.

To advancing metric learning, we aim to separate the

embedded features from each video into highlight and non-highlight subsets based on simple thresholding on their scores  $s_m^T, s_n^F$ . (Following [25, 39], we select the segments with top 15% scores as the highlight subsets, and the remaining as the non-highlight ones.) Thus, we have  $E_m^T = \{E_m^{T,high}, E_m^{T,non}\}$ , where  $E_m^{T,high} = \{e_{m,1}^{T,high}, \dots, e_{m,t}^{T,high}\}$ ,  $E_m^{T,non} = \{e_{m,1}^{T,non}, \dots, e_{m,t}^{T,non}\}$  where  $m$  denotes the video index for this third-person video, and  $t$  indicates the frame numbers. Similar, we have  $E_n^F = \{E_n^{F,high}, E_n^{F,non}\}$  for the  $n$  first-person video. To perform metric learning with the above definition, we can extract positive segment pairs from training data if both segments are from the same subset, and negatives pairs if they are from different subsets.

Since one of our goals is to eliminate the domain differences, our training process needs to include data across different video domains when calculating the triplet loss. Take Fig. 3 for example, if an input first-person segment  $e_{in}$  is picked from  $E_n^{F,high}$ , we then choose a positive segment  $e_{pos}$  from either  $E_m^{T,high}$  or  $E_n^{F,high}$  to form a positive pair, and choose  $e_{neg}$  from either  $E_m^{T,non}$  or  $E_n^{F,non}$  as the negative pair. Thus, the triplet loss is calculated as follows:

$$\mathcal{L}_{tri} = \max \{0, \mathcal{M} - \mathcal{D}(e_{in}, e_{neg}) + \mathcal{D}(e_{in}, e_{pos})\}, \quad (1)$$

where  $\mathcal{D}(e, e') = 1 - \frac{e \cdot e'}{\|e\|_2 \|e'\|_2}$  returns the cosine distance between the embedded features, and  $\mathcal{M}$  denotes the margin for metric learning. We note that, the cosine distance instead of the  $l_2$  distance is considered, since it would not be biased by the magnitude of the embedded features.

In addition to embedding highlight information via metric learning techniques, we further integrate a decoder for recovering the input C3D features, so that the learned encoder/embedding layers would not overfit the small amount of labeled data. Thus, with  $e_{in}, e_{pos}, e_{neg}$  as inputs, the decoder needs to reconstruct the corresponding C3D features  $\hat{x}_{in}, \hat{x}_{pos}, \hat{x}_{neg}$ . The standard  $l_2$  distance is applied for calculating this reconstruction loss  $\mathcal{L}_{rec}$ :

$$\mathcal{L}_{rec} = \sum_{d \in \{in, pos, neg\}} \|\hat{x}_d - x_d\|^2. \quad (2)$$

Finally, the feature embedding layers in our network are updated by minimizing  $(\mathcal{L}_{tri} + \mathcal{L}_{rec})$ , while the decoder part is only updated via  $\mathcal{L}_{rec}$ .

### 3.2. Keyframe Selection

With embedded short-term features (i.e., embedded C3D features from segments) extracted from the previous feature embedding network, it is now desirable to integrate longer dependencies across video segments for better exploiting the highlight information. Inspired by recent works like [39, 15, 25], we advance the network component of

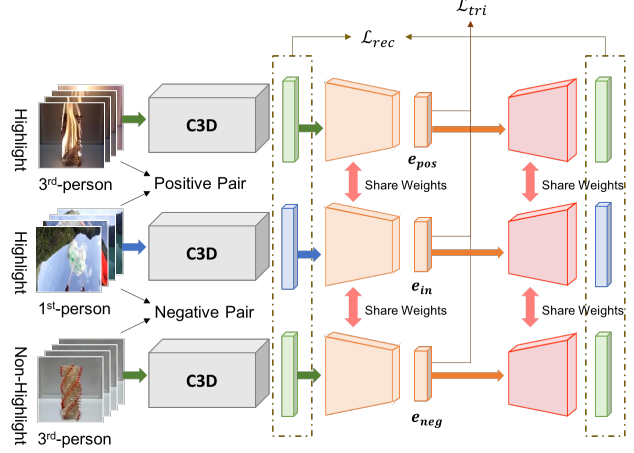


Figure 3. Training of feature embedding layers by observing cross-domain labeled data. The embedded features  $e$ ,  $e_{pos}$ , and  $e_{neg}$  calculate the triplet loss for separating highlight and non-highlight semantics, while their content information is preserved by the autoencoder architecture.

biLSTMs [39] in our proposed framework, aiming at modeling and identifying long-term dependencies between segments within a single video for summarization purpose.

As shown in Fig. 2, we have concatenated features  $U = \{u_1, \dots, u_t\}$  as the inputs for the biLSTMs. Note that

$$u_t = \text{concat}(x_t, e_t) u_t \in \mathbb{R}^{(p+q)}, \quad (3)$$

where  $X \in \mathbb{R}^{p \times t}$  and  $E \in \mathbb{R}^{q \times t}$  denoted the C3D and the associated embedded features, respectively.

As noted in Sect. 2, both forward and backward cells in biLSTM generate sequential outputs, i.e.,  $h^{forward}$  and  $h^{backward}$ , which better exploit and preserve semantic information across video segments. A single-hidden-layer scoring network is deployed for predicting the importance scores  $\hat{s} = \{\hat{s}_1, \dots, \hat{s}_t\}$  with inputs  $h^{forward}, h^{backward}$ . Note that each  $\hat{s}_i$  is a 2D softmax vector.

It can be seen that, using our proposed architecture, we approach the segment score prediction problem by solving a highlight classification task. With the introduced biLSTM and scoring layers, we convert ground truth scores in  $S^T$  and  $S^F$  into 2-dimensional one-hot vectors  $y_i = (0, 1) \vee (1, 0)$ , and  $y = \{y_1, \dots, y_t\}$  by thresholding the top 15% important scores in each video. The scoring loss  $\mathcal{L}_s$  is defined as,

$$\mathcal{L}_s = -\frac{1}{t} \sum_{i=1}^t y_i \cdot \log(\hat{s}_i), \quad (4)$$

where  $t$  denotes the number of segments in each video. During the training process of biLSTM and scoring layers, the gradients of  $\mathcal{L}_s$  would also be back-propagated to fine-tune the feature embedding layers.

Dataset	Description	Video type	Total length	# of videos	Usage	Annotations
SumMe [10]	User videos	1st-person	14 m	5	Testing	Frame-level importance scores
		3rd-person	50 m	20	Training( $V^T$ )	Frame-level importance scores
TvSum [35]	YouTube videos	3rd-person	3 h 30 m	50	Training( $V^T$ )	Shot-level importance scores
UnSum	GoPro sports raw videos	1st-person	30 m	5	Training( $V^F$ )	Selected keyshots
		1st-person	~10 h	65	Training( $V^U$ )	None

Table 1. Descriptions and properties of video datasets considered.

### 3.3. Learning from Unlabeled Data

As a major contribution of our proposed model, our network allows learning from unlabeled first-person videos  $V^U$  during the above adaptation and summarization process. The unlabeled segments first pass through the feature extractor and embedding network as embedded vectors  $\mathbf{e}_u$ . During the training stage, We sample a subset of features from  $X^T$  and  $X^F$  to perform the above embedding. By comparing the distances between  $\mathbf{e}_u$  and the sampled vectors in the embedding space, the farthest and nearest vector neighbors with respect to  $\mathbf{e}_u$  would be viewed as  $\mathbf{e}_{neg}$  and  $\mathbf{e}_{pos}$ , respectively. As a result, even with the presence of unlabeled first-person video data,  $\mathcal{L}_{tri}$  in (1) can be calculated and updated in each iteration.

As for updating the keyframe selection layers by observing the unlabeled data, we also view the segments with top 15% scores as pseudo (highlight) labels  $\hat{\mathbf{y}} \in \{0, 1\}^{2 \times t}$  from  $\hat{\mathbf{s}}$ , and the remaining as pseudo non-highlight segments. Thus, the scoring loss  $\mathcal{L}_s$  measured by  $\hat{\mathbf{y}}$  would update our network accordingly, which completes the joint adaptation and summarization process via observing such unlabeled first-person video data.

## 4. Experiments

**Datasets** We now describe the datasets considered in our experiments. The two publicly available datasets, **SumMe** [10] and **TVSum** [35], are recently used to evaluate the performance of video summarization task. **SumMe** consists of 25 user videos with length varying from 1 to 6 minutes, in which the annotations of frame-level importance scores are provided. Within this dataset, there are five first-person videos, "Base jumping, Bike Polo, Scuba, Valparaiso Downhill, Uncut Evening Flight", which are applied as test data for quantitative evaluation and comparison. **TVSum** consists of 50 third-person videos collected from YouTube, and each of them are annotated with shot-level importance scores via crowdsourcing (20 annotations per video). The videos in this dataset is viewed as the third-person labeled data for training purposes.

In addition to SumMe and TVSum, we collect a new first-person dataset **UnSum** with total number of 70 videos, including GoPro sport videos from YouTube as well as videos from existing HUJI [32] video-indexing dataset. La-

bels at the shot-level are manually annotated for 5 videos by several users, and the remaining ones are viewed as unlabeled data (for the purpose of semi-supervised learning).

In summary, all the labeled third-person video data in TVSum and SumMe are viewed as  $V^T$ , while the annotated and unannotated first-person ones in UnSum are served as  $V^F$  and  $V^U$ , respectively. For fair comparisons, we have all annotated first-person videos in SumMe as test data for evaluation. Detailed properties of each dataset are listed in Table 1. The baseline and state-of-the-art approaches are implemented using the same experimental setting.

**Implementation details** The input to C3D feature extractor is a  $16 \times 112 \times 112 \times 3$  tensor, therefore we split videos into 2-second segments composed of 16 frames. We then crop each frame into a overlapping left-right image pair, both with size of  $112 \times 112$ , in order to fit the input size for C3D feature extractor. The C3D model was pre-trained on Sport1M [16] and kept fixed in our training procedure. We concatenate the features (8192-d) extracted from pool-5 layer of C3D model for both images, as an intact feature for the video segment. Our feature embedding network is a two-layer fully connected network (2048 and 512 SELU units) followed by a linear layer (128 units), whereas the feature decoder has a symmetric structure to the feature embedding network. The keyframe selection network consists of a biLSTM with 1024 hidden units in each cell and a one-layer fully connected scoring layer followed by softmax output.

To train the proposed model, we perform a two-stage optimization process. In the first stage, We train our feature embedding network based the 60K triplets generated from both first- and third-person videos with importance annotations. The margin parameter  $\mathcal{M}$  of cosine distance is set as 1.5 and the size of embedding features is 128. We train our network using Adam optimizer with a batch size of 8, first- and second-momentum of 0.9 and 0.99, weight decay of 0.001, and dropout probability of 0.6. The learning rate of the feature embedding network is set to  $10^{-5}$  while the decoder is set to  $10^{-4}$ . The network is trained in total 40K iterations.

In the second stage, a training video of full-length is used for learning the parameters of the keyframe selection network. We optimize our network by using Adam optimizer

with a batch size of 2, first- and second-momentum of 0.9 and 0.99, dropout probability of 0.6. The learning rate of the keyframe selection networks is set as  $10^{-5}$  whereas the feature embedding network is fine-tuned with a learning rate of  $10^{-6}$ . The overall network is trained in total 3K iterations.

We also fine-tune our network with unlabeled data after each stage. The learning rate of all parameters are set to  $10^{-6}$ . We fuse 50% labeled data and 50% unlabeled data during the fine-tuning process. We implement all our framework with TensorFlow [1] toolkit and conduct training and evaluation using NVIDIA GTX 1080Ti.

**Evaluation metrics** Following the keyshot-based evaluation criteria used in [25, 39], the length of video summaries should be less than 15% of the total length of original video. Let  $\mathcal{A}$  be the set of extracted keyshots, and  $\mathcal{B}$  is the set of keyshots selected by user-annotated importance scores. The precision  $\mathcal{P}$  and recall  $\mathcal{R}$  according to the overlap of  $\mathcal{A}$  and  $\mathcal{B}$  can be define as:

$$\mathcal{P} = \frac{\text{total overlap duration of } \mathcal{A} \text{ and } \mathcal{B}}{\text{total duration of } \mathcal{A}}, \quad (5)$$

$$\mathcal{R} = \frac{\text{total overlap duration of } \mathcal{A} \text{ and } \mathcal{B}}{\text{total duration of } \mathcal{B}}, \quad (6)$$

thus F-measure is computed as,

$$\mathcal{F}_{total} = \frac{2 \times \mathcal{P} \times \mathcal{R}}{\mathcal{P} + \mathcal{R}}. \quad (7)$$

We find that several testing videos dominate the total F-measure values owing to their long video durations, therefore we also report the F-measure averaged over videos,  $\mathcal{F}_{mean}$ , in our experiments.

#### 4.1. Methods to be compared

We compare our work with respect to three baselines (noted as **Random**, **Uniform**, and **SumMe**) and two state-of-the-art video summarization algorithms: **TDCNN** [38] and **vsLSTM** [39]. We first describe sequentially how the three baselines are obtained.

- **Random:** 15% of segments from each test video are randomly sampled as the highlight.
- **Uniform:** Instead of random sampling, now the 15% of segments from each test video are equidistantly selected as the highlight.
- **SumMe:** The baseline method proposed together with the SumMe dataset [10] reports their performance numbers on each video, which enables the direct comparison in our experiments.

Next, for the two state-of-the-art video summarization algorithms, since their objectives and experimental settings

are distinct from the ones we are targeting here, the performance numbers shown in their papers can not be directly referred here for quantitative comparison. Hence, we re-implement their works which are further trained by using the identical setting as our framework.

- **TDCNN:** We train a highlight classifier as proposed in [38], which is built upon a 5-layers fully connected Siamese network and outputs shot-level importance scores. The loss function for the TDCNN classifier is defined as  $\mathcal{L}_{pair} = \max\{0, 1 - s(\mathbf{x}^{high}) + s(\mathbf{x}^{non})\}$ , where  $s(\mathbf{x}^{high})$  and  $s(\mathbf{x}^{non})$  are the scores of highlight and non-highlight segments. The positive and negative pairs of training data for learning the Siamese network of TDCNN classifier is produced by following the same criteria described in Sect. 3.
- **vsLSTM:** As shown in [39], it is implemented as an architecture of stacking a video feature extraction and a biLSTM with 1024 hidden units, where the parameters of biLSTM are learned by using the scoring loss  $\mathcal{L}_s$  as defined in Eq. 4. We experiments on two variants of vsLSTM different in their feature extractors: vsLSTM\* with GoogleNet features and vsLSTM<sup>†</sup> with C3D features.

Both TDCNN and vsLSTM models are trained with identical training data (i.e.  $V^T$  and  $V^F$ ), and the resulting summaries are selected from segments with top 15% scores. Also, all the aforementioned methods split videos in the same way if the segmentation is necessary for their deployment.

#### 4.2. Quantitative results and analysis

**Results** Table. 2 summarizes the quantitative results of our framework, baselines, and the state-of-the-art video summarization algorithms. Our full model achieves the best performance in comparison to the others, particularly we observe a 8% boost by leveraging the additional unlabeled training data  $V^U$ , as shown in the last two column of Table. 2. These results demonstrate that our framework is capable of exploring the knowledge from limited labeled data as well as large amount of unlabeled data jointly.

**Feature embedding** Among the baselines, vsLSTM is the most similar one to our proposed method from the perspective of model architecture, while the main difference comes from the feature embedding network of our framework. As observable from the performance of our work and vsLSTM in Table. 2, with both being trained on the same data  $V^T + V^F$ , our feature embedding network successfully encodes the semantic related to importance scores in the embedding representation, which therefore serves as an extra guideline for estimating the keyframes.



	Baseline			Non-sequence-to-sequence			Sequence-to-sequence			
Method	Random	Uniform	SumMe [10]	vsLSTM w/oLSTM	Ours w/o LSTM	TDCNN [38]	vsLSTM [39]		Ours	
Video feature				C3D			GoogleNet	C3D	C3D	
Training data				$V^T + V^F$			$V^T + V^F$		$V^T + V^F$	$V^T + V^F + V^U$
Base jumping	0.100	0.121	0.121	0.100	0.218	0.236	0.209	0.273	0.263	0.273
Bike Polo	0.229	0.081	0.356	0.386	0.414	0.257	0.229	0.286	0.429	0.286
Scuba	0.140	0.189	0.184	0.060	0.260	0.540	0.320	0.300	0.400	0.560
Valparaiso Downhill	0.177	0.190	0.242	0.300	0.323	0.331	0.154	0.408	0.246	0.384
Uncut Evening Flight	0.171	0.170	0.271	0.263	0.077	0.098	0.170	0.430	0.545	0.553
Total frame $\mathcal{F}_{total}$	0.1630	0.1567	-	0.2386	0.215	0.2302	0.1933	0.3563	0.4017	<b>0.4336</b>
Video mean $\mathcal{F}_{mean}$	0.1632	0.1501	0.234	0.2217	0.258	0.2924	0.2163	0.3392	0.3766	<b>0.4112</b>

Table 2. Performance evaluation and comparisons on first-person video summarization in terms of F-measures.

Moreover, in comparison to TDCNN, which encapsulates a C3D feature directly into a single importance score, our feature embedding not only reflects the importance but also preserves the spatiotemporal content of a video segment. Yet the feature embedding learned by our proposed model also aids in assigning pseudo-labels to unlabeled videos thus expanding the scale of training data.

In addition, the original vsLSTM framework utilizes GoogleNet to independently encode the video frames of a segment into feature vectors and further take the concatenation over them as a proxy to obtain the spatiotemporal feature representation (cf. 8-th column in Table. 2). In order to have fair comparison with our proposed method, their GoogleNet-based spatiotemporal feature extraction is replaced by the C3D framework (cf. 9-th column in Table. 2). The better performance stemmed from C3D features shows that simply concatenating feature representation of frames can not fully capture the important content within the spatiotemporal volume of a video segment, in which it provides evidence (implicitly) to support the usage of C3D features in our proposed model.

**Sequence-to-sequence modeling** In order to demonstrate the contribution of using recurrent neural network in video summarization task, we experiment on some model variants which remove biLSTM and select keyframes by a single fully-connected network. The main difference can be seen between "Non-sequence-to-sequence" and "Sequence-to-sequence" groups in Table. 2. Sequence-to-sequence models surpass their opponents by a significant margin in performance. The results indicate that although C3D features are capable of describing the short-term content in a video, the long-term dependencies captured by LSTM benefit much to video summarization.

**Domain adaptation** The contribution of introducing domain adaption into our proposed method is clearly exhibited by the experimental results. We conduct a comparison between the performance of TDCNN, vsLSM, and our work as shown in Table. 2. With all being trained on the same

data  $V^T + V^F$ , TDCNN and vsLSTM suffer from learning first-person video summarization as only little amount of annotated first-person videos is available. On the other hand, our framework overcomes this issue by discovering a joint feature embedding across first- and third-person videos and mitigating the domain shift, in which the knowledge from both first- and third person videos can be explored jointly for learning better summarization.

**Unlabeled videos** The benefits of introducing unlabeled videos during training are obvious to our model. Unlabeled videos expand the diversity and the richness of training data, for both feature embedding and biLSTM networks, in which our framework obtain significant growth of performance after including unlabeled data into the learning framework. By contrast, other state-of-the-art methods demand fully supervised training data, thus not able to handle unannotated video in their training procedures. In brief, our proposed framework yields better results without taking any effort on extra data annotation.

### 4.3. Qualitative Results

Fig. 4 shows example summaries of a challenging testing video "Uncut Evening Flight" in SumMe dataset along with its ground truth scores. The testing video is a 6-minutes-long uncut video recorded by a camera mounted on an airplane. It is a typical first-person video since the video content reflects the motion of the recorder (i.e. the airplane) and no specific objects are intentionally focused. The blue bars in Fig. 4 indicate frame-level ground truth scores annotated by users. The interval colored in green, red and yellow are corresponded to summaries generated by our work, vsLSTM, and TDCNN respectively. The red horizontal lines represent the threshold for splitting highlight and non-highlight parts, as described in Sect. 3.

In this video, the takeoff and landing actions of the plane cause huge vibration on the camera, resulting shaky and blurry scenes which are usually rarely found in third-person videos. TDCNN fails to predict takeoff and landing and meanwhile generate many erroneous summaries. vsLSTM

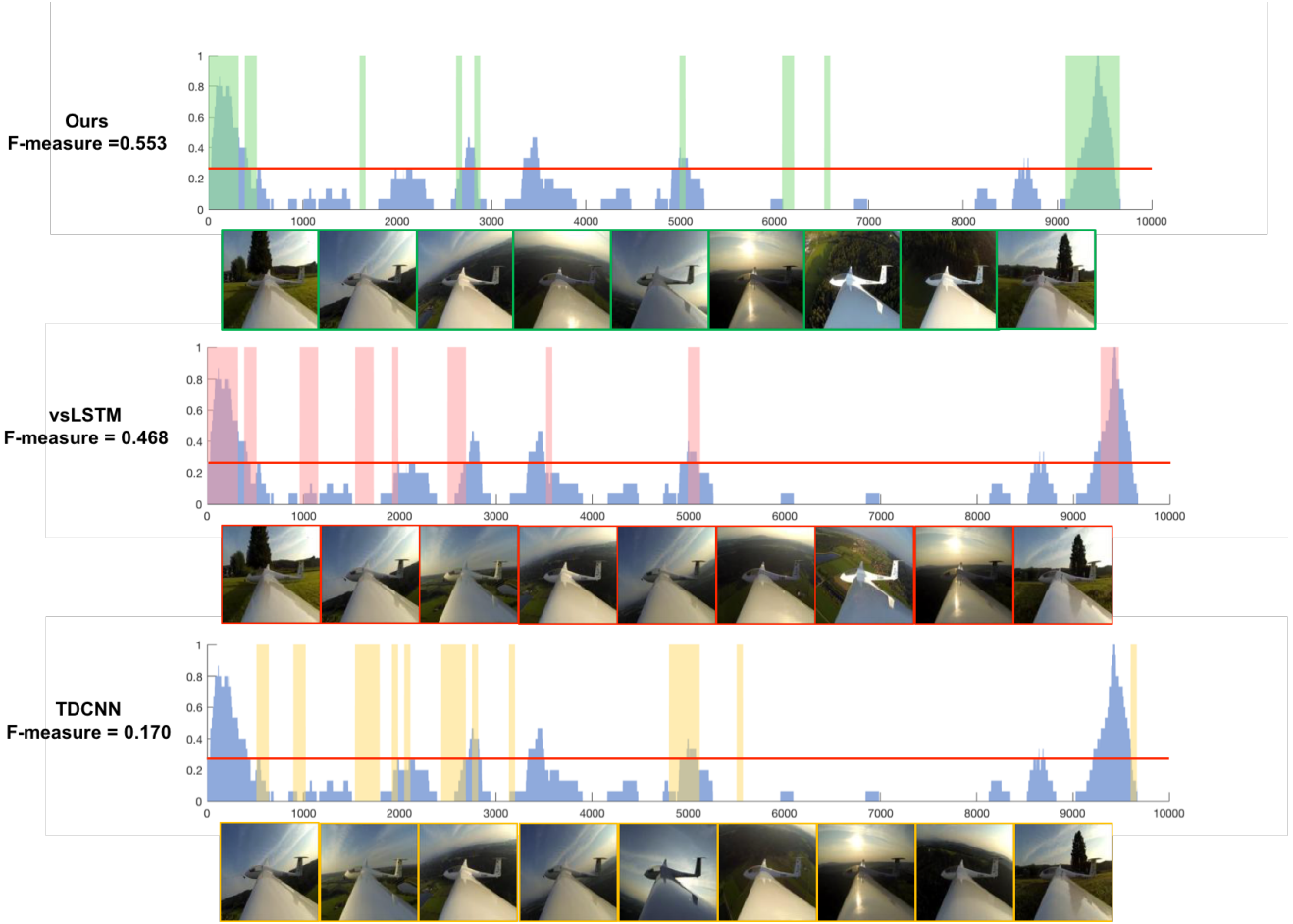


Figure 4. Example results of first-person video summarization from testing video "Uncut Evening Flight". Note that the user-annotated scores (ground truth) are shown in blue, while the predicted summaries from our works, vsLSTM, and TDCNN are shown in green, red and yellow, respectively. The red horizontal line split the scores into highlight (i.e., top 15%) and non-highlight ones (see Sect. 3). Note that our method captures moments like take-off, landing, or particular sunset scenes in the summarization output.

manages to predict takeoff and landing moments in the summary, while it also suffers from selecting video segments of less importance. In contrast, the summarization estimated by our work not only successfully captures takeoff and landing, but also locates more accurately than others the particular sunset scenes of interest. The superior performance of our proposed method in this challenging first-person video can also be seen qualitatively in Table. 2.

## 5. Conclusion

In this paper, we proposed a novel framework of first-person video summarization. Our network uniquely integrates spatiotemporal feature embedding and keyframe selection, which are realized by the architectures of triplet, auto-encoder, and recurrent neural networks, and allows end-to-end model learning and refinement. A practical semi-supervised setting is considered, so that only a small amount of annotated first-person video data is required, while the

remaining labeled third-person and unlabeled first-person videos are observed during training. Our experiments verified the use of our model for summarizing first-person videos, which exhibited improved performance over state-of-the-art video summarization approaches.

## References

- [1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning.
- [2] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv:1609.08675*, 2016.
- [3] Y. Cong, J. Yuan, and J. Luo. Towards scalable summarization of consumer videos via sparse dictionary selection. *IEEE Transactions on Multimedia (TMM)*, 14(1):66–75, 2012.



- [4] A. G. del Molino, C. Tan, J.-H. Lim, and A.-H. Tan. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems*, 47(1):65–76, 2017.
- [5] N. Ejaz, I. Mehmood, and S. W. Baik. Efficient visual attention based framework for extracting key frames from videos. *Signal Processing: Image Communication*, 28(1):34–44, 2013.
- [6] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [7] A. M. Ferman and A. M. Tekalp. Two-stage hierarchical video summary extraction to match low-level user browsing preferences. *IEEE Transactions on Multimedia (TMM)*, 5(2):244–256, 2003.
- [8] C. Gianluigi and S. Raimondo. An innovative algorithm for key frame extraction in video summarization. *Journal of Real-Time Image Processing*, 1(1):69–88, 2006.
- [9] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.
- [10] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.
- [11] M. Gygli, H. Grabner, and L. Van Gool. Video summarization by learning submodular mixtures of objectives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [12] M. Gygli, Y. Song, and L. Cao. Video2gif: Automatic generation of animated gifs from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [13] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, 2015.
- [14] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(6):797–819, 2011.
- [15] Z. Ji, K. Xiong, Y. Pang, and X. Li. Video summarization with attention-based encoder-decoder networks. *arXiv:1708.09545*, 2017.
- [16] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [17] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [18] Y. J. Lee and K. Grauman. Predicting important objects for egocentric video summarization. *International Journal of Computer Vision (IJCV)*, 114(1):38–55, 2015.
- [19] X. Li, B. Zhao, and X. Lu. A general framework for edited video and raw video summarization. *IEEE Transactions on Image Processing (TIP)*, 2017.
- [20] Z. Li, G. M. Schuster, and A. K. Katsaggelos. Minmax optimal video summarization. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 15(10):1245–1256, 2005.
- [21] Y.-L. Lin, V. I. Morariu, and W. Hsu. Summarizing while recording: Context-based highlight detection for egocentric videos. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015.
- [22] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [23] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang. A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia (TMM)*, 7(5):907–919, 2005.
- [24] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li. A user attention model for video summarization. In *Proceedings of the ACM Conference on Multimedia (MM)*, 2002.
- [25] B. Mahasseni, M. Lam, and S. Todorovic. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [26] A. G. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation*, 19(2):121–143, 2008.
- [27] S. Nepal, U. Srinivasan, and G. Reynolds. Automatic detection of ‘goal’ segments in basketball videos. In *Proceedings of the ACM Conference on Multimedia (MM)*, 2001.
- [28] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(10):1345–1359, 2010.
- [29] R. Panda, A. Das, Z. Wu, J. Ernst, and A. K. Roy-Chowdhury. Weakly supervised summarization of web videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [30] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015.
- [31] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [32] Y. Poley, A. Ephrat, S. Peleg, and C. Arora. Compact cnn for indexing egocentric videos. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.
- [33] Y. Rui, A. Gupta, and A. Acero. Automatically extracting highlights for tv baseball programs. In *Proceedings of the ACM Conference on Multimedia (MM)*, 2000.
- [34] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.
- [35] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

- [36] M. Sun, A. Farhadi, and S. Seitz. Ranking domain-specific highlights by analyzing edited videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.
- [37] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.
- [38] T. Yao, T. Mei, and Y. Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [39] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
- [40] B. Zhao and E. P. Xing. Quasi real-time summarization for consumer videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.