Weakly Supervised Summarization of Web Videos

Rameswar Panda[1] Abir Das[2] Ziyan Wu[3] Jan Ernst[3] Amit K. Roy-Chowdhury[1]
[1] UC Riverside [2] Boston University [3] Siemens Corporate Technology
{rpand002@,amitrc@ece.}ucr.edu dasabir@bu.edu {ziyan.wu,jan.ernst}@siemens.com

Abstract

Most of the prior works summarize videos by either exploring different heuristically designed criteria in an unsupervised way or developing fully supervised algorithms by leveraging human-crafted training data in form of video-summary pairs or importance annotations. However, unsupervised methods are blind to the video category and often fail to produce semantically meaningful video summaries. On the other hand, acquisition of large amount of training data in supervised approaches is non-trivial and may lead to a biased model. Different from existing methods, we introduce a weakly supervised approach that requires only video-level annotation for summarizing web videos. Casting the problem as a weakly supervised learning problem, we propose a flexible deep 3D CNN architecture to learn the notion of importance using only video-level annotation, and without any human-crafted training data. Specifically, our main idea is to leverage multiple videos of a category to automatically learn a parametric model for categorizing videos and then adopt the model to find important segments from a given video as the ones which have maximum influence to the model output. Furthermore, to unleash the full potential of our 3D CNN architecture, we also explored a series of good practices to reduce the influence of limited training data while summarizing videos. Experiments on two challenging and diverse datasets well demonstrate that our approach produces superior quality video summaries compared to several recently proposed approaches.

1. Introduction

Video summarization, which automates the process of extracting a brief yet informative synopsis of a long video, has attracted intense attention in the recent years. Much progress has been made in developing a variety of ways to summarize videos, by either limiting the scope to a specific context (e.g., sports, news) [47, 62, 29] or exploring different design criteria (representativeness [10, 7, 6], interestingness [11, 45]) in an unsupervised manner. More recently, we see a shift of paradigm in video summarization. Supervision in terms of labeled summary [18, 15, 48, 72, 69] or importance annotation [69] is being used to train video summarization models which traditionally has been treated as an unsupervised learning problem.

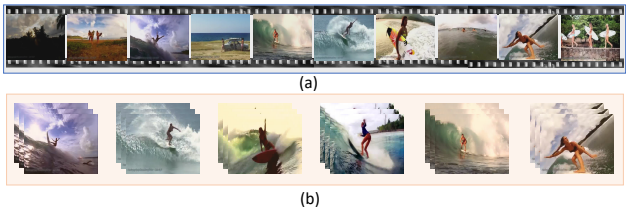

Figure 1. (a) Raw video, (b) Video summary. Given a set of videos with only video-level annotation (e.g., surfing), our method learns what aspects are important within a category, such as riding a wave with a surfboard, off the lip, and cutback in surfing.

Let us consider a video of *surfing* (see Fig. 1). An unsupervised approach, being blind to the video category, would fail to single out the short segments corresponding to riding a wave with a surfboard, off the lip, and cutback, etc, whereas a supervised method would require huge amount of human-labeled video-summary pairs which are difficult to collect especially for long and unconstrained web videos. Moreover, it is generally feasible to have only a limited number of users to annotate training videos, which may lead to a biased summarization model.

On the other hand, collecting videos with video-level annotation (e.g., surfing) is much easier, since many videos with attached tags are readily available on open video datasets such as YouTube-8M [1] as well as on web. Motivated by this observation, we pose an important question in this paper: *Can weakly supervised learning with only video-level annotation, be leveraged upon for summarzing web videos?* This is an extremely relevant problem to address due to the difficulty and non-scalability of obtaining large amount of human-annotated training data for web videos.

Recently, Convolutional Neural Networks (CNNs) have witnessed great success in many vision tasks such as image classification [26], object detection [14], localization [73, 63], and semantic segmentation [30]. Similarly, for videos, 3D CNNs have shown better performance in activity recognition, compared to 2D CNNs since they exploit the temporal aspects of activities typically shown in videos [60, 21].

The success of 3D CNNs also shed light on several video analysis tasks [33, 50, 68, 23, 66]. However, whether and how an end-to-end 3D CNN architecture could be exploited for video summarization still remains as a novel and rarely addressed problem. This motivates us to build upon 3D CNNs for weakly supervised summarization of web videos.

Moreover, deep 3D CNNs, in practice, require a large amount of training data to achieve optimal performance. However, publicly available datasets for video summarization remain limited, in size and diversity (e.g., CoSum [6], TVSum [54]). Thus, another important question that we address in this work is *how can we efficiently train the 3D CNN architecture given limited training data, with the goal of summarizing unconstrained web videos?*

1.1. Overview of Solution Strategy

A summary is a condensed synopsis that conveys the most important details of the original video. Since importance is a subjective notion, in this paper, we propose to identify and model important video segments as the most common activities among the videos of a category and remove uninteresting or idiosyncratic segments that occur relatively infrequently. Our method is motivated by the observation that *similar videos have similar summaries*. For instance, suppose we have a collection of videos of "surfing". It is quite likely good summaries for those videos would all contain segments corresponding to riding with surfboard, floating on water, and off the lip surfing, etc. Thus, we hypothesize that the notion of importance is intricately related to the video category and this relation can be learned. We accomplish this via a flexible 3D CNN architecture, namely Deep Summarization Network (`DeSumNet`), which can assign an importance score to each segment without requiring any human-annotated training data. Specifically, we have the access to only video-level annotation during training and our goal is to learn a parametric model, which could be applied to summarize new unconstrained web videos.

As an overview of our approach for summarizing a given video, (1) we perform a forward pass on the input video which generates a distribution of scores over the video categories; (2) calculate the CNN derivatives with respect to each video segment via back-propagation guided by the category with highest score in the forward pass; (3) compute spatio-temporal importance score, and then generate summaries of a given length based on the computed importance scores. An overview of our approach is illustrated in Fig. 2.

Furthermore, to unleash the full potential of our 3D CNN architecture for video summarization, we explored a number of good practices to reduce the influence of limited training data, including (1) cross-dataset pre-training; (2) model adaptation with web data; (3) enhanced data augmentation. Experiments show that the above practices for training with limited data indeed improve the performance of our method when extracting summaries from web videos.

1.2. Contributions

We address a novel and practical problem in this paper—how to extract summaries from web videos without requiring large amount of human-crafted training data, but only video-level annotation. Towards solving this problem, we make the following contributions. (1) a weakly supervised approach based on 3D CNN that advances the frontier of learning for video summarization; (2) computing spatio-temporal importance scores based on CNN derivatives without resorting to additional training steps; (3) study on a series of good practices for learning 3D CNN with limited training data while extracting video summaries.

2. Related Work

Our work relates to three major research directions: video summarization, video highlight detection and CNNs for weakly supervised learning. Here, we focus on some representative methods closely related to our work.

Video Summarization has been studied from multiple perspectives (see reviews [34, 61]). Without supervision, summarization methods rely on low-level visual indices to determine the important parts of a video. Various strategies have been studied, including clustering [2, 8, 16, 17], sparse optimizations [10, 40, 38], and energy minimization [45, 11]. Leveraging crawled web images or videos is also another recent trend for video summarization [24, 54, 25, 39].

Departing from unsupervised methods, recent works formulate video summarization as a supervised learning problem. Representative methods along this direction learn how to select informative video subsets from human-created summaries [18, 15, 48, 71], or learn important facets, like faces, objects [28, 31, 5]. Similar in spirit, deep learning based methods have been applied for video summarization with the help of pair-wise deep ranking model [69] or RNNs [72]. However, these approaches assume the availability of large amount of human-created video-summary pairs or importance annotations, which are in practice difficult to obtain for unconstrained web videos. Our method, instead, learns the notion of importance from a set of videos belonging to a category (weak supervision), and hence provides much greater scalability in extracting summaries from web videos. Most relevant to our approach is the work in [44] which learns multiple SVM classifiers, one per each category for importance scoring. We differ from [44] in that we propose an end-to-end learning scheme for video summarization by modeling temporal aspects with a 3D CNN architecture instead of a computationally intensive feature representation that involves multi-scale SIFT feature extraction and fisher vector encoding with a Gaussian mixture model. Another distinctive feature of our approach is in computing the spatio-temporal importance scores via CNN derivatives without resorting to additional training steps.

Video Highlight Detection is highly related to summariza-

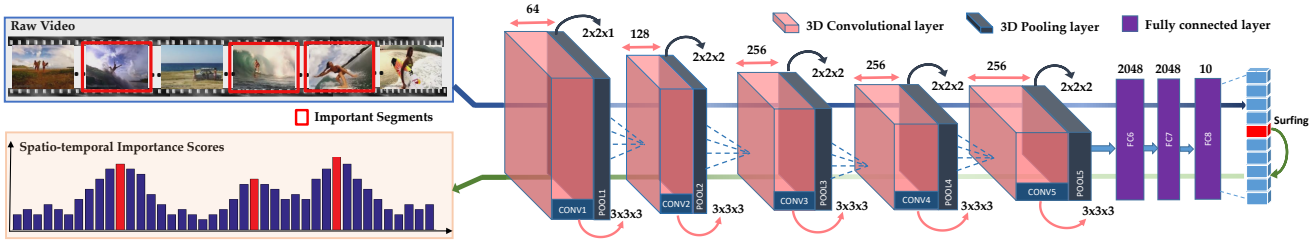

Figure 2. Illustration of DeSumNet for weakly supervised video summarization. During training, we simply have videos with video-level annotation and we train our network with these videos to learn what is important within a video category. During testing, we first perform a forward pass on the given video and then compute the spatio-temporal importance scores via back-propagation guided by the category with highest score in the forward pass. Convolutional and pooling kernel sizes are represented by arrows. Best viewed in color.

tion since both of them intend to extract a brief synopsis containing segments of special interest from a video [69]. Many earlier approaches have primarily been focused on highlighting sports videos [47, 65, 59]. A latent SVM model is employed to detect highlights by learning from pairs of raw and edited videos [58, 57]. Success of deep learning also imparted improved performance in highlight detection [67]. However, most of these techniques may not generalize well to web videos since they are either based on heuristic rules or require huge amount of human-crafted training data which are difficult to collect in many cases.

CNN-Based Weakly Supervised Learning have achieved promising performance in several vision tasks [4, 9, 35, 36, 70, 43, 42, 41, 49]. Most of these approaches employ feed-forward computation and/or back-propagation on a CNN to achieve segmentation with only image-level annotation. Gradient-based deep CNN visualizations have shown to be effective in localizing objects in images without relying on bounding box or pixel-level annotations [51, 49, 56, 37]. Although effective for images, there is relatively little work on applying CNNs for weakly-supervised learning on video data. While emphasizing the weak supervision principle, we extend [51] to the video domain and present the first end-to-end framework for weakly supervised extraction of summaries from web videos. A very recent work [46] generates spatio-temporal saliency maps using an encoder-decoder network with human annotated captions which are harder to obtain compared to video class labels.

3. Methodology

In this section, we first present our weakly supervised approach for computing importance scores (Section 3.1), followed by our study on good practices in learning a 3D CNN architecture (DeSumNet) given limited training data while summarizing unconstrained web videos (Section 3.2).

3.1. Gradient-based Importance Computation

Objective. As discussed in Section 1, fully supervised methods for video summarization, either require large amount of human-created summaries [18, 15, 48, 71] or segment-wise annotations [69], to train a model for selecting important segments from a video. Though effective for the task of extracting summary from videos, acquisition of such training data is non-trivial, since labeling video segments with importance scores is much more labor-intensive and often requires annotators with domain knowledge. Similarly, creating large number of video-summary pairs is also highly infeasible and not scalable in many cases. This is mainly due to the fact that an annotator may need to go through the entire video to extract a summary. To tackle this issue, we propose a video-level framework (DeSumNet) to compute importance scores without requiring large amount of human-crafted training data. Specifically, our core idea is to leverage multiple videos belonging to a specific category to automatically learn a parametric model for categorizing videos and then adopt the model to find important segments from a given video as the ones which have the maximum influence to the model output (i.e., the category of the video).

Approach Details. Let $\mathbf{X}$ be a video divided into n equal segments, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, $\mathbf{x}_i \in \mathbb{R}^{p \times q \times k}$, where p and q denote the height and width of each frame and k represents the number of frames in a segment. Inspired by the recent advances in image gradient activation [51, 73], we compute the importance score of each segment $\mathbf{x}$ in a weakly supervised way for summarizing videos. The main idea of our approach is to model the importance as an input sensitivity, i.e., which segments of a video are most responsible in characterizing the video to belong to a specific category. In other words, if a small change in a segment has a large effect on the network output then it is logical to assume that this segment is more important than others. Our proposed architecture manifests this notion of importance by the change of network output with respect to the video segments. This, in turn, is quantified by the relative strength of the gradient of the output class score with respect to the input segments. As an example, for a surfing video, the strength of the gradient will be large when it is computed with respect to a segment corresponding to riding a wave with a surfboard, compared to segments that are less significant and occur relatively infrequently in such videos, e.g.,

two people talking to each other near the car (see Fig. 1). We compute the gradient with respect to the input segments efficiently using backpropagation through the layers. Note that we train our network, DesumNet, with only video-level labels to learn what is important within a category.

Formally, given a video $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ with predicted class label c, and a well trained network, $\phi : \mathbf{x} \rightarrow \phi_1 \circ \phi_2 \circ \cdots \circ \phi_l$ where l is the number of layers, the spatio-temporal importance map, $\mathcal{S}(\phi, \mathbf{x}_i, \mathbf{h})$ for a segment $\mathbf{x}_i$ can be obtained by the derivative of the network output with respect to the video segment as follows:

$$\mathcal{S}(\phi, \mathbf{x}_i, \mathbf{h}) = \frac{\partial}{\partial \mathbf{x}} \langle \mathbf{h}, \phi(\mathbf{x}) \rangle \Big|_{\mathbf{x}=\mathbf{x}_i} \tag{1}$$

where $\mathbf{h}$ is an one-hot vector that selects c-th class in the output. The importance map $\mathcal{S}(\phi, \mathbf{x}_i, \mathbf{h})$ encodes how sensitive the output prediction is with respect to changes at the input video segment. Using chain rule, we can compute the importance map as follows:

$$\texttt{vec}[\mathcal{S}(\phi, \mathbf{x}_i, \mathbf{h})] = \mathbf{h}^\intercal \times \frac{\partial\, \texttt{vec}[\phi_l]}{\partial\, \texttt{vec}[\mathbf{x}_l]^\intercal} \times \cdots \times \frac{\partial\, \texttt{vec}[\phi_1]}{\partial\, \texttt{vec}[\mathbf{x}_i]^\intercal} \tag{2}$$

where the vec operator allows us to use matrix notations for the derivatives. Note that the computation of importance map is extremely fast, since it only requires a single back-propagation pass without any additional training steps.

Given a video that belongs to category c with n segments, each containing k frames of size $p \times q$, the spatio-temporal importance score of each segment are computed as follows: (1) we first compute $\mathcal{S}(\phi, \mathbf{x}_i, \mathbf{h})$ by back-propagation and rearrange it to a 3D map, $\mathcal{M}_i \in \mathbb{R}^{p \times q \times k}$ of same size as the input segment; (2) we then obtain the frame-level importance scores within the segment, as $\mathcal{I}_{i,j} = \frac{1}{p \times q} \sum_{x,y} \mathcal{M}_i(x, y, j),\ \forall [j]_{i=1}^{k}$, where $\mathcal{I}_{i,j}$ represents the importance score of j-th frame in i-th segment; (3) finally, the spatio-temporal importance score of a video segment is computed, as $\mathcal{I}_i = \mathcal{G}(\mathcal{I}_{i,1}, \cdots, \mathcal{I}_{i,k})$, where $\mathcal{G}$ is an aggregation function applied along the temporal dimension k. We evaluated two forms of the aggregation function $\mathcal{G}$, including maximum and averaging in our experiments and empirically choose averaging to report our final results.

3.2. Training DeSumNet for Summarization

Objective. In the previous section we have presented a weakly supervised video summarization approach for computing the spatio-temporal importance scores without requiring any human-crafted training data. However, to achieve optimal performance, a few practical concerns have to be taken care of, e.g., the limited number of training examples in standard benchmarking summarization datasets (e.g., CoSum [6], TVSum [54]). To handle such an important issue, we study a series of good practices in training DeSumNet for summarization, which are in general applicable while training 3D CNNs with limited data.

Approach Details. Our approach for training DeSumNet with limited examples involves the following steps: (1) cross-dataset pre-training; (2) progressive model adaptation with web data; (3) enhanced data augmentation.

Network Architecture. Our proposed DeSumNet architecture is based on 3D CNNs since 3D convolution/pooling which operates in spatial and temporal dimensions simultaneously, can capture both appearance and motion for activities. Recent works have also shown that temporal aspects of activities play an important role in generating good video summaries [67, 69, 39]. We follow [60] and use a homogeneous setting with kernel size $3 \times 3 \times 3$ in all convolutional layers. We use max pooling for all 3D pooling layers with kernel size 2×2 in spatial with stride 2, while vary in temporal. Using the notations conv(number of filters) for 3D convolutional layer, pool(temporal kernel size, temporal stride) for the 3D pooling layer, and fc(number of filters) for the fully connected layer, the pattern of our DeSumNet architecture is as follows: conv1(64) – pool1(1,1) – conv2(128) – pool2(2,2) – conv3(256) – pool3(2,2) – conv4(256) – pool4(2,2) – conv5(256) – pool5(2,2) – fc6(2048) – fc7(2048) – fc8(C), where C is the number of categories.

Cross-Dataset Pre-training. Quantity of training data is crucial for training a deep neural network. However, our case is particularly difficult since standard video summarization datasets are limited in size (e.g., only 50 videos in TVSum [54]). Thus, we first use the large action recognition dataset, UCF101 [55] (101 action classes with ∼13k videos) to pre-train our DeSumNet architecture for parameter initialization. The goal of this cross-dataset pre-training is to learn generic video-level features and also to reduce the effect of over-fitting in experiments. With this initialization, we fine-tune the network by utilizing training data from summarization datasets (e.g., CoSum [6], TVSum [54]) to further adjust the parameters, specific to our target task.

Model Adaptation with Web Data. Cross-dataset pre-training provides a good initialization for training our DeSumNet architecture. However, to learn what is important within a category, we often need a large set of diverse examples. Given the maturity of commercial video search engines (e.g., YouTube), one obvious and cheap solution is to utilize top ranked videos that are highly correlated with the video category. However, there are two key difficulties which prevent us from using such videos directly in training. First, they are typically noisy containing lots of unrelated frames. Second, they are usually untrimmed and very lengthy, where some relevant activities are often hidden in between irrelevant ones. Inspired by the success of webly-supervised learning in computer vision [54, 12, 13, 24], we propose a simple, yet effective progressive model adaptation scheme for enhancing our DeSumNet architecture by harvesting noisy and untrimmed web videos. Our approach involves the following steps: (1) given the category name as

a keyword, we download a set of top-ranked videos at the best quality available from YouTube. In practice, we crawl about 30 videos on average for each category; (2) we then adopt the initial fine-tuned network to truncate the videos to relevant segments: we keep the segments whose probability of being in the category is more than a threshold; (3) having retained a set of relevant and trimmed videos, we finally update the network parameters to obtain an improved model for computing spatio-temporal importance scores. Adapting our model with the refined web videos not only reduces the effect of limited training data but also increases the diversity of training data, which are essential for learning the notion of importance in video summarization. Experiments show that this approach indeed improves the performance of our method in generating good video summaries.

Enhanced Data Augmentation. To further reduce the impact of limited data, we explore different data augmentation techniques to generate diverse training samples. In addition to the random cropping used in original 3D CNN architecture for action classification [60], we employ three new data augmentation techniques as follows: (1) horizontal flipping–we generate random crops of size $112\times112\times16$ from the input clips and then randomly flip all frames within a crop horizontally with 50% probability; (2) multi-scale jittering–we follow [52, 64] and use multi-scale jittering by using random crops with a size of $x \times y \times 16$, where x and y are randomly selected from $\{128, 112, 96, 84\}$; (3) corner cropping–we randomly pick $112 \times 112 \times 16$ crops from the center or four corners from the entire input segment. Augmentation with corner crops from the entire image has recently shown to be effective in object detection [20].

4. Experiments

Datasets. We conduct rigorous experiments on two different publicly available benchmark datasets to verify the effectiveness of our framework, namely CoSum [6] and TVSum [54] datasets. Both of the datasets are extremely diverse: while CoSum dataset consists of 51 videos covering 10 categories from the SumMe benchmark [17], the TVSum dataset contains 50 videos downloaded from YouTube in 10 categories defined in the TRECVid Multimedia Event Detection (MED) task [53]. Detailed description of these datasets are available in the supplementary material.

Experimental Settings.

• We implement our network using the Caffe [22] toolbox and conduct all our training on a NVIDIA Tesla K80 GPU.

• The input to the network is a segment of dimension $128 \times 171 \times 16$ (i.e., $p = 128, q = 171, k = 16$) and output is a category label which belongs to one of $\mathtt{C}$ video categories.

• For the parameter initialization, we train our network from scratch using stochastic gradient descent with a minibatch size of 50, momentum of 0.9, and weight decay of 0.005. The learning rate is initialized to 0.003 and is reduced to its $\frac{1}{10}$ after every 4 epochs (15 epochs in total).

• With this initialization, we fine-tune all the layers with an initial learning rate of 0.0003, except the last `fc8` layer which is changed to produce a 10-dimensional output on both datasets. We train the last layer from scratch with initial parameter values sampled from a zero-mean Gaussian distribution with $\sigma = 0.01$ and an initial learning rate of 0.003. We decrease the learning rate of all the layers to its $\frac{1}{10}$ after every 4 epochs (7 epochs in all). We use dropout with probabilities (= 0.5) in the first two fully connected layers and found it essential for training.

• During model adaptation, we take the refined web videos to further enhance the network. We run the model adaptation for 10 epochs on CoSum dataset and 6 epochs on TVSum dataset. For data augmentation, we use horizontal flipping, random cropping, multi-scale jittering and corner cropping, as described in Section 3.2.

• Following the literature [71, 72], we randomly choose 80% of the videos for training and use the remaining 20% for testing on both datasets. To produce predictions for an entire video, we follow [60] and average segment-level predictions of 10 segments which are randomly selected from the video. The average video classification accuracy over five such random sets, are 88% and 72% on CoSum and TVSum datasets, respectively.

4.1. Generating Video Skims

Goal. The objective of this experiment is to validate the effectiveness of our approach in extracting video skims of user-defined length, which can convey the most important details of the original video. Specifically, a video skim is composed of several shots that represent most important portions of the input video within a short duration.

Solution. A common practice towards generating video skims is to first perform a video shot boundary detection [19] as it maintains visual coherence of the output summary. We follow [6, 39] and divide videos into multiple non-uniform shots. After this, we perform a mean pooling over the segment-wise importance scores within a video shot. The pooled result serves as the final importance score of a shot to be used in generating skims. To generate a video skim, we first sort the shots by decreasing importance score (resolving ties by favoring shorter video shots), and then construct the optimal video skim from the top-ranked shots that fit in the user defined length constraint.

Evaluation. Following [6, 24, 39], we assess the quality of an automatically generated video skim by comparing it to human judgment. In particular, given a proposed summary (i.e., video skim) and a set of human-created summaries, we compute the pairwise average precision (AP) and then report the mean value motivated by the fact that there exists not a single ground truth summary, but multiple such summaries are possible. We finally average over the number of videos to obtain the overall performance on a dataset.

For evaluation, both datasets provide multiple user-

Table 1. Experimental results on CoSum dataset.

Mean Average Precision	Humans			Unsupervised Methods				Supervised Methods			Proposed
	Worst	Mean	Best	SMRS	Quasi	MBF	CVS	KVS	seqDPP	SubMod	DeSumNet
Top-5	0.668	0.814	0.887	0.491	0.507	0.588	0.676	0.684	0.692	0.735	0.721
Relative to average human	82.1%	100%	109.1%	60.4%	62.6%	72.3%	83.2%	84.1%	85.2%	90.3%	88.5%
Top-15	0.682	0.821	0.916	0.506	0.527	0.579	0.677	0.686	0.709	0.745	0.736
Relative to average human	83.0%	100%	111.5%	61.7%	64.3%	70.6%	82.5%	83.6%	86.5%	90.8%	89.7%

Table 2. Experimental results on TVSum dataset.

Mean Average Precision	Humans			Unsupervised Methods				Supervised Methods			Proposed
	Worst	Mean	Best	SMRS	Quasi	MBF	CVS	KVS	seqDPP	SubMod	DeSumNet
Top-5	0.382	0.516	0.608	0.322	0.334	0.353	0.388	0.398	0.447	0.461	0.424
Relative to average human	74.2%	100%	117.8%	62.5%	64.8%	68.5%	75.3%	77.3%	86.7%	89.6%	82.2%
Top-15	0.372	0.507	0.589	0.320	0.325	0.342	0.371	0.387	0.435	0.443	0.415
Relative to average human	73.5%	100%	116.3%	63.2%	64.1%	67.4%	73.2%	76.5%	85.8%	87.4%	81.8%

annotated summaries for each video. For CoSum dataset, we follow [6, 39] and compare each video skim with five human-created summaries. For TVSum dataset, we first average the frame-level importance scores, created via crowdsourcing [54] to compute shot-level scores and then select top 50% shots for each video as human-created summary, as in [6, 39]. Finally, we compare each system-generated video skim with twenty shot-based human-created summaries to obtain the performance metric in TVSum dataset.

Compared Methods. We compare our approach with several methods that fall into three main categories:

(1) Unsupervised approaches including SMRS [10] (CVPR'12), Quasi [10] (CVPR'14), MBF [6] (CVPR'15), and CVS [39] (CVPR'17); First two baselines (SMRS, Quasi) use sparse coding for selecting important shots, whereas MBF leverage visual co-occurrence across videos of a category to generate a summary. The recent method CVS, is based on collaborative sparse representative selection to extract a video skim by exploiting visual context from additional videos within a category.

(2) Supervised methods including KVS [44] (ECCV'14), seqDPP [15] (NIPS'14), and SubMod [18] (CVPR'15); KVS learns multiple SVM classifiers for importance scoring, whereas seqDPP, and SubMod use video-summary pairs to train a model for extracting video summaries.

(3) Human performance comparison including Worst Human, Mean Human, and Best Human. The worst human score is computed using the summary which is least similar to the rest of the human-created groundtruth summaries whereas the best score represents the most similar summary containing most shots selected by many humans. The purpose of comparing with human performance is to provide a pseudo-upper bound for the summarization task, and thus we also report normalized average precision scores by rescaling the mean AP of human selections to 100%.

Following [39], we use C3D feature [60] (4096 dimensional) to represent the shots and tune the parameters in each method to have the best performance. We follow the procedure described in [15, 71] to generate training groundtruths (i.e. oracle summaries) from multiple human-created summaries in both datasets.

Comparsion with Unsupervised Methods. Table 1 shows the mean AP on both top 5 and 15 shots included in the summaries for CoSum dataset, whereas Table 2 shows the results on TVSum dataset. From both tables, the following observations can be made: (1) The proposed weakly supervised approach consistently outperforms all compared unsupervised methods on both datasets by a significant margin. (2) Among the alternatives, the recent CVS method is the most competitive. However, the gap is still significant due to the two introduced components working in concert: exploiting temporal aspects of activities via an end-to-end 3D CNN architecture and learning the notion of importance from similar category-related videos. The top-5 mAP performance improvements over CVS are 5.3% and 6.9% on CoSum and TVSum datasets respectively. (3) Our approach performed particularly well on CoSum dataset since it contain videos that have their visual concepts described well by the other category-related videos, e.g., all the videos of the surfing category contain visually similar shots depicting different aspects of surfing such as riding with surfboard, off the lip and cutback surfing, etc. (4) Summarization on TVSum dataset, however, is more challenging because of unconstrained categories, e.g, grooming an animal. Our approach still outperforms all the unsupervised alternatives to achieve a top-5 mAP of 82.2%, showing that the notion of importance can still be learned from similar videos without any heuristically designed criteria in summarizing videos.

Comparsion with Supervised Methods. While comparing with supervised alternatives, we have the following key findings from Table 1, 2: (1) DeSumNet outperforms KVS on both datasets by a big margin (maximum improvement of 6.1% in top-15 mAP on CoSum), showing the advantage of our gradient-based spatio-temporal importance computation and more powerful representation learning with 3D CNNs. (2) On Cosum dataset, DeSumNet outperforms seqDPP by a margin of 3.3% in top-5 mAP, and 3.2% in top-15 mAP, respectively. SubMod, however, overcomes DeSumNet but the difference is moderate ($\sim$1%). This results suggest that although being a weakly supervised approach, our method is still competitive with the fully supervised methods in extracting important shots from

Table 3. Exploration study on training strategies. Numbers show top-5 mAP scores, relative to the average human score (in %).

Methods	CoSum	TVSum
Scratch	71.4	66.7
Scratch+NoisyWebData	76.3	69.5
Pre-train	83.5	75.2
Pre-train+NoisyWebData	84.4	77.3
Pre-train+ModelAdaptationwithRefinedWebData	87.7	80.8
Pre-train+ModelAdaptation+EnhancedDataAugmentation	88.5	82.2

videos. (3) On TVSum dataset, the performance gap between our method and fully supervised methods (`seqDPP`, `SubMod`) begins to appear. This is expected as with a challenging dataset involving uncontrolled and very diverse web videos, a weakly supervised approach can not compete with a fully supervised one, especially when the later one is using large amount of human-annotated video-summary pairs. However, we would like to point out once more that in practice collecting human-labeled summaries are very difficult and unrealistic in actual scenarios.

Comparsion with Human Performance. As can be seen from Table 1, 2, our method outperforms the `Worst Human` score in both dataset (Top-5 mAP: 88.5% vs 82.1% on CoSum and 89.7% vs 83.0% on TVSum dataset). This indicates that our method produces informative summaries comparable to the groundtruth human-created summaries.

Exploration Study. To better understand the contributions of various training strategies described in Section 3.2, we analyzed the performance in following six different settings: (1) training from scratch; (2) training from scratch but adding the downloaded videos from YouTube; (3) with cross-data pre-training; (4) combination of pre-training and directly mixing the downloaded videos; (5) combination of pre-training and model adaptation with refined web videos; and (6) combination of pre-training, model adaptation and enhanced data augmentation. We have the following key observations from Table 3: (1) Performance of training from scratch is much worse than that of pre-training, which conforms the common perception about 3D CNNs: while they are powerful, they often desire a larger amount of annotated data in order to perform well. (2) Model adaptation with refined web data achieves better performance compared to only pre-training which shows that the proposed adaptation scheme is effective in learning discriminative information within a category. We further utilize different data augmentation techniques to regularize the training, which improves the top-5 mAP to 88.5% on CoSum and 82.2% on TVSum dataset respectively. (3) Directly adding the noisy and untrimmed web videos without any refinement performs worst in both datasets. This is not surprising, since irrelevant content about a category will lead the training to the wrong direction, and in turn, the fine-tuned model has a hard time to find what is important within a category, thus even hurting the final summarization performance.

Diversity. Following [18], we performed an experiment where we clustered video segments beforehand and used an uncorrelated subset of segments in generating summaries to explicitly enforce diversity in the extraxted summary. This however led to no significant improvement ($\sim$0.8% top-5 on CoSum), suggesting that our method produces both interesting and diverse summaries. We also observe that both of the summarization datasets mostly contain user videos which rarely contain multiple interesting but redundant events.

Qualitative Results. Fig. 3 shows the exemplar summaries produced by our approach, `DeSumNet` and the recent `CVS` method in summarizing a video of the *base jumping* category from the CoSum dataset. The scores below our result indicate the predicted spatio-temporal importance scores of each segment in the summary. As can be seen from Fig. 3, our approach can efficiently learn what aspects are important within a base jumping category and thus identify the most important shots from the video, i.e., the 2nd shot depicting jumping from a cliff and the 4th one indicating jumping with hand-together. On the other hand, `CVS`, completely misses such shots and rather selects shots that are irrelevant to the category of base jumping—we believe this is because `CVS` focuses on selecting shots that can well reconstruct the original video with low reconstruction error and hence does not capture the notion of importance properly while summarizing videos.

Fig. 4 shows a failure case of our method. This video records very diverse contents and the scenes change frequently among the indoor house and the outdoor field. In particular, this video appears to be completely different from the other videos belonging to the category of *Grooming an Animal* in the TVSum dataset. For these reasons, we found the returned summaries of our method and `CVS` to be largely similar. From the summarization results, we see that `DeSumNet` still selects diverse contents, but fails to capture the fine details on grooming the dog, e.g., cutting nails. While our current approach has been designed to be weakly supervised, we believe it could be made more robust to handle such videos by explicitly using semantic analysis [32] and could also benefit from domain adaptation techniques [27] for more challenging datasets.

4.2. Generating Video Time-lapse

Goal. The goal of this experiment is to analyze the performance of our approach in generating time-lapse videos to enable a more efficient and engaging viewing experience. Video time-lapse is a condensed summary which is normally created by adjusting the playing speed of segments based on the importance score. Specifically, segments with high importance score are played at a smaller rate and segments with lower importance are played at a higher rate [3].

Solution. A simple option is to select frames from a segment based on the importance score while generating a time-lapse video. We first select the sampling rate s_i as, $r_i = 1 - \mathcal{I}_i / \sum_i^n \mathcal{I}_i$, where $\mathcal{I}_i$ represent the spatio-temporal importance score of the i-th segment and then uniformly re-

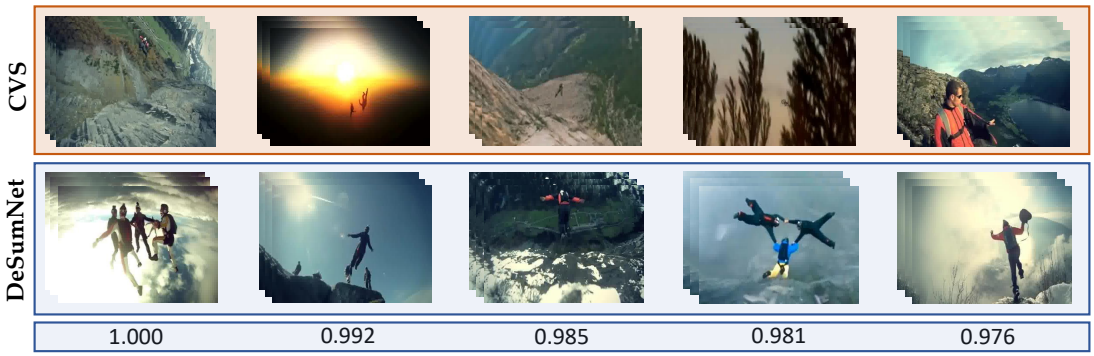

Figure 3. Exemplar video summaries generated by CVS and DeSumNet, along with our predicted spatio-temporal importance score $\in$ [0,1]. As can be seen, CVS often selects some shots that are irrelevant and not truly related to the *base jumping* category. Our method, DeSumNet, on the other hand, automatically selects the maximally informative shots by leaning what aspects are important in base jumping from a set of similar videos. We show the top-5 results represented by three central frames from each shot. Best viewed in color.

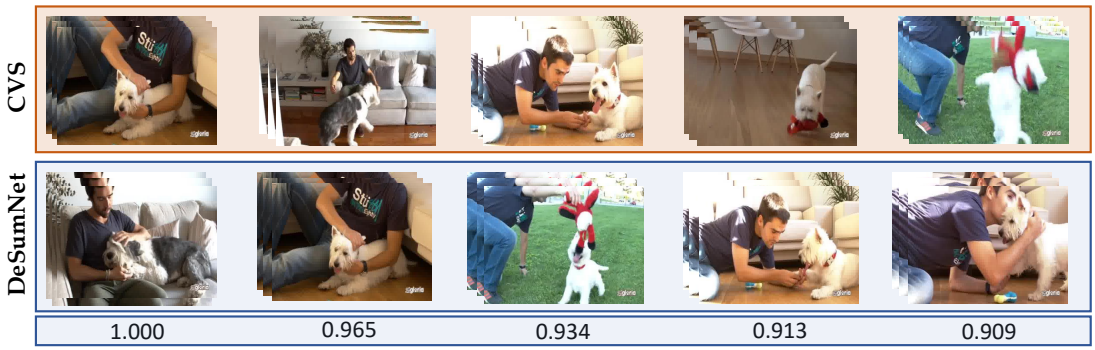

Figure 4. Exemplar summaries generated while summarizing a video of the category *Grooming an Animal* from the TVSum dataset. This video records very diverse contents and the scenes change frequently among the indoor house and the outdoor field. For these reasons, we found the returned summaries of our method and CVS to be largely similar. See text for more details. Best viewed in color.

Table 4. Average human ratings in evaluating video time-lapse.

Datasets	CVS	KVS	DeSumNet
CoSum	3.23	3.15	4.03
TVSum	2.34	2.56	3.18

move $r_i \times k$ number of frames from i-th segment to produce a video time-lapse. Note that since a time-lapse consists of all video segments, there is no need of explicitly dividing videos into non-uniform shots, as in skimming.

Evaluation. Since there exists no publicly available ground-truth to evaluate the quality of video time-lapse, we performed subjective evaluation using ten users. Given a video, the study experts were first required to watch the original video and then shown the time-lapse videos constructed using different methods. They were asked to rate the overall quality of each system-generated video time-lapse by assigning a rating from 1 (worst) to 5 (best).

Compared Methods. We compare our approach with two methods, CVS [39] and KVS [44] that use sparse coding and SVM classifiers, respectively for importance scoring. We follow the same procedure in all methods to extract a time-lapse summary from a given video.

Results. Table 4 shows average human ratings for both datasets. Similar to the results of video skimming, our approach outperforms both of the methods in creating an informative time-lapse video. This again corroborates the fact that using category-level supervision for extracting important segments from web videos captures what humans consider important within a video.

Additional results and discussions along with qualitative summaries are included in the supplementary material.

5. Conclusion

We presented a weakly supervised approach to summarize videos with only video-level annotation. Motivated by the fact that importance is related to the network input sensitivity, we introduced an effective method for computing spatio-temporal importance scores without resorting to additional training steps. In addition, we explored a series of good practices for efficiently training our network architecture with limited training data while summarizing web videos. Extensive experiments on two standard datasets well demonstrate the efficacy of our method over several competing methods.

Acknowledgements: This work is partially supported by NSF grants IIS-1316934 and CPS-1544969.

References

[1] 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 preprint arXiv:1609.08675*, 2016. 1

[2] J. Almeida, N. J. Leite, and R. da S. Torres. VISON: VIdeo Summarization for ONline applications. *PRL*, 2012. 2

[3] E. P. Bennett and L. McMillan. Computational time-lapse video. In *TOG*, 2007. 7

[4] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with posterior regularization. In *BMVC*, 2014. 3

[5] G. K. Bo Xiong and L. Sigal. Storyline representation of egocentric videos with an application to story-based search. In *ICCV*, 2015. 2

[6] W.-S. Chu, Y. Song, and A. Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *CVPR*, 2015. 1, 2, 4, 5, 6

[7] Y. Cong, J. Yuan, and J. Luo. Towards Scalable Summarization of Consumer Videos Via Sparse Dictionary Selection. *TMM*, 2012. 1

[8] S. E. F. de Avila, A. P. B. Lopes, A. da Luz Jr., and A. de Albuquerque Arajo. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *PRL*, 2011. 2

[9] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool. Weakly supervised cascaded convolutional networks. *arXiv preprint arXiv:1611.08258*, 2016. 3

[10] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *CVPR*, 2012. 1, 2, 6

[11] S. Feng, Z. Lei, D. Yi, and S. Li. Online content-aware video condensation. In *CVPR*, 2012. 1, 2

[12] C. Gan, C. Sun, L. Duan, and B. Gong. Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In *ECCV*, 2016. 4

[13] C. Gan, T. Yao, K. Yang, Y. Yang, and T. Mei. You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. In *CVPR*, 2016. 4

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1

[15] B. Gong, W.-L. Chao, K. Grauman, and F. Sha. Diverse sequential subset selection for supervised video summarization. In *NIPS*, 2014. 1, 2, 3, 6

[16] G. Guan, Z. Wang, S. Mei, M. Ott, M. He, and D. D. Feng. A Top-Down Approach for Video Summarization. *TOMCCAP*, 2014. 2

[17] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool. Creating summaries from user videos. In *ECCV*, 2014. 2, 5

[18] M. Gygli, H. Grabner, and L. Van Gool. Video summarization by learning submodular mixtures of objectives. In *CVPR*, 2015. 1, 2, 3, 6, 7

[19] A. Hanjalic. Shot-boundary detection: Unraveled and resolved? *TCSVT*, 2002. 5

[20] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014. 5

[21] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *TPAMI*, 2013. 1

[22] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 5

[23] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 2

[24] A. Khosla, R. Hamid, C. J. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *CVPR*, 2013. 2, 4, 5

[25] G. Kim, L. Sigal, and E. P. Xing. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *CVPR*, 2014. 2

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[27] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 2011. 7

[28] Y. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 2

[29] B. Li, H. Pan, and I. Sezan. A general framework for sports video summarization with its application to soccer. In *ICASSP*, 2003. 1

[30] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[31] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013. 2

[32] T. Mei, L. X. Tang, J. Tang, and X. S. Hua. Near-lossless semantic video summarization and its applications to video analysis. *TOMCAP*, 2013. 7

[33] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *CVPR*, 2016. 2

[34] A. G. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *JVCIR*, 2008. 2

[35] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014. 3

[36] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015. 3

[37] H. Pan and H. Jiang. A deep learning based fast image saliency detection algorithm. *arXiv preprint arXiv:1602.00577*, 2016. 3

[38] R. Panda, N. C. Mithun, and A. Roy-Chowdhury. Diversity-aware multi-video summarization. *TIP*, 2017. 2

[39] R. Panda and A. K. Roy-Chowdhury. Collaborative summarization of topic-related videos. In *CVPR*, 2017. 2, 4, 5, 6, 8

[40] R. Panda and A. K. Roy-Chowdhury. Sparse modeling for topic-oriented video summarization. In *ICASSP*, 2017. 2

[41] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015. 3

[42] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015. 3

[43] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015. 3

[44] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *ECCV*, 2014. 2, 6, 8

[45] Y. Pritch, A. R. Acha, A. Gutman, and S. Peleg. Webcam synopsis: Peeking around the world. In *ICCV*, 2007. 1, 2

[46] V. Ramanishka, A. Das, J. Zhang, and K. Saenko. Top-down visual saliency guided by captions. In *CVPR*, 2017. 3

[47] Y. Rui, A. Gupta, and A. Acero. Automatically extracting highlights for tv baseball programs. In *MM*, 2000. 1, 3

[48] A. Sharghi, B. Gong, and M. Shah. Query-focused extractive video summarization. In *ECCV*, 2016. 1, 2, 3

[49] W. Shimoda and K. Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *ECCV*, 2016. 3

[50] Z. Shou, D. Wang, and S.-F. Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*, 2016. 2

[51] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 3

[52] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[53] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR*, 2006. 5

[54] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. Tvsum: Summarizing web videos using titles. In *CVPR*, 2015. 2, 4, 5, 6

[55] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 4

[56] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *ICLR*, 2014. 3

[57] M. Sun, A. Farhadi, T.-H. Chen, and S. Seitz. Ranking highlights in personal videos by analyzing edited videos. *TIP*, 2016. 3

[58] M. Sun, A. Farhadi, and S. Seitz. Ranking domain-specific highlights by analyzing edited videos. In *ECCV*, 2014. 3

[59] H. Tang, V. Kwatra, M. E. Sargin, and U. Gargi. Detecting highlights in sports videos: Cricket as a test case. In *ICME*, 2011. 3

[60] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. 2015. 1, 4, 5, 6

[61] B. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *TOMCCAP*, 2007. 2

[62] J. Wang, C. Xu, E. Chng, and Q. Tian. Sports highlight detection from keyword sequences using hmm. In *ICME*, 2004. 1

[63] L. Wang, Y. Xiong, D. Lin, and L. V. Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, 2017. 1

[64] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015. 5

[65] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang. Highlights extraction from sports video based on an audio-visual marker detection framework. In *ICME*, 2005. 3

[66] H. Xu, A. Das, and K. Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *ICCV*, 2017. 2

[67] H. Yang, B. Wang, S. Lin, D. Wipf, M. Guo, and B. Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *ICCV*, 2015. 3, 4

[68] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *ICCV*, 2015. 2

[69] T. Yao, T. Mei, and Y. Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *CVPR*, 2016. 1, 2, 3, 4

[70] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, 2016. 3

[71] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Summary transfer: Exemplar-based subset selection for video summarization. In *CVPR*, 2016. 2, 3, 5, 6

[72] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory. In *ECCV*, 2016. 1, 2, 5

[73] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 1, 3