# Temporal Cue Guided Video Highlight Detection with Low-Rank Audio-Visual Fusion

Qinghao Ye[1,2*†] Xiyue Shen[3*] Yuan Gao[4*] Zirui Wang[1*] Qi Bi[5] Ping Li[1†] Guang Yang[6]
[1] Hangzhou Dianzi University [2] University of California, San Diego [3] East China Normal University
[4] University of Oxford [5] Wuhan University [6] Imperial College London

## Abstract

*Video highlight detection plays an increasingly important role in social media content filtering, however, it remains highly challenging to develop automated video highlight detection methods because of the lack of temporal annotations (i.e., where the highlight moments are in long videos) for supervised learning. In this paper, we propose a novel weakly supervised method that can learn to detect highlights by mining video characteristics with video level annotations (topic tags) only. Particularly, we exploit audio-visual features to enhance video representation and take temporal cues into account for improving detection performance. Our contributions are threefold: 1) we propose an audio-visual tensor fusion mechanism that efficiently models the complex association between two modalities while reducing the gap of the heterogeneity between the two modalities; 2) we introduce a novel hierarchical temporal context encoder to embed local temporal clues in between neighboring segments; 3) finally, we alleviate the gradient vanishing problem theoretically during model optimization with attention-gated instance aggregation. Extensive experiments on two benchmark datasets (YouTube Highlights and TVSum) have demonstrated our method outperforms other state-of-the-art methods with remarkable improvements.*

## 1. Introduction

Recently, the rise of short-form video sharing applications (*e.g.*, TikTok and Reels) has attracted world-wide attention on the Internet. From a content producer's point of view, it is not a delightful experience for them to trim long videos and localize those highlight segments manually. Therefore, an automated method is needed desperately to identify highlight clips from untrimmed videos.

Video highlight detection has received many interests in the field of computer vision. Numerous methods [24, 31, 28] have been proposed to automatically tailor highlight from untrimmed videos tagged with a specific topic or keyword, which can be generally divided into two categories, *i.e.*, supervised learning and weakly supervised learning based methods. *Supervised* methods [22, 9, 11] generally trained a ranker detector with frame-level annotations to rank the highlight segments, of which scores were higher than those of non-highlight segments. However, annotating the frame-wise highlight manually is extremely labor-intensive and time-consuming. To overcome this problem, *weakly supervised* methods [20, 3, 30] use weakly labeled videos to train the model. For methods, videos are usually divided into two categories based on whether there are topic-specific segments presented: a positive video contains at least one highlighted snippet and a negative one should have no highlights. The trained highlight detector needs to learn from those topic-specific video snippets to identify highlights for unseen videos. However, these weakly supervised methods have limited capacity in terms of: (1) effectively capturing the complex interactions between audio and vision streams in videos while maintaining the efficiency; (2) enhancing the semantic continuity modeling by leveraging the so-far unexplored temporal evolution among video segments.

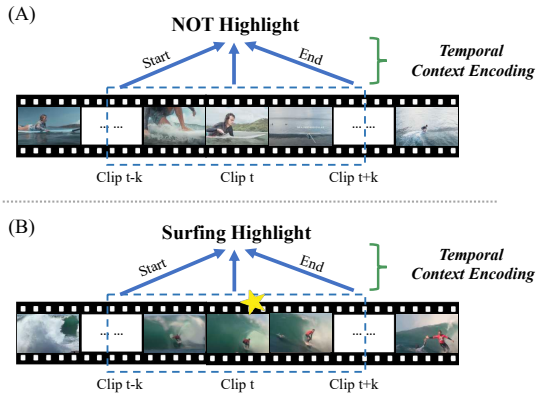



Figure 1. Temporal relationship reasoning for highlight detection. Part (A) contains different surfing tutorial sections with little temporal cues to infer highlight. However, in part (B), clips in the box with dashed line contain take off and wipe out footages, which indicate the occurrence of surfing, and they encoded contextual features among these clips that can be used to infer the highlight.

For the former bottleneck, previous approaches [12, 29] tended to adopt linear fusion schemes (*e.g.*, concatenation

*Equal contribution.
†Corresponding author: yuan.gao2@gmail.com

or summation between two modalities) to convey video representation. Nonetheless, such linear fusion methods were not able to fully capture the complex association between the two modalities due to the distinct feature distribution of each modality since audio does not always correspond to the visual frames. Subsequently, bilinear pooling [8] was designed to fuse two kinds of features through modeling the pairwise interaction. However, one drawback is that it requires a great number of parameters to train leading inefficient training for video highlight detection, which is notoriously resource-heavy while facing the risk of overfitting. As a high-rank tensor can be decomposed into several matrices and a core tensor [16], we introduce an audio-visual tensor fusion scheme and apply the low-rank constraint on the core tensor which can not only provide rich video representation by modeling the modalities interaction efficiently, but can also reduce the number of trainable parameters of the model which serves as the regularization.

In another aspect, most of existing methods [30, 12] tend to handle video segments individually while temporal evolution across consecutive segments is not adequately exploited. As depicted in Figure 1, due to the temporal characteristics of the video, those non-highlight segments may provide fruitful clues for inferring the highlight and foreshadow the occurrence of the highlight. For further illustration, the beginning of the highlight event or the ending would indicate the happening of the highlight segment, *e.g.*, taking off and wiping out footages suggest the occurrence of surfing (demonstrated in Figure 1 (B)). Inspired by this, we propose a hierarchical temporal context encoding scheme, which exploits the temporal context relationships among local adjacent segments via utilizing the temporal cues for the first time. In particular, a video segment encodes the contextual information from its neighbor segments in a hierarchical paradigm, which is beneficial for higher-order content interaction among segments. Therefore, the temporal contextual feature includes the representation of the original segment, and also models local dependency among adjacent segments. In conclusion, we propose a low-rank audio-visual tensor fusion mechanism and hierarchical temporal context encoding scheme to address the above limitations, which we believe are important signs of progress for weakly supervised video highlight detection.

Finally, considering that only video-level annotations are available, correctly classifying videos can provide useful inductive bias for topic-specific highlight detection since a video may contain highlights of various events. Therefore, we exploit this advantage by introducing a novel attention-gated instance aggregation module, which derives representative video score from individual segment scores. More importantly, the gradient vanishing issue occurs constantly when the score of a segment in the video is high, and traditional methods [2, 4] do not have theoretic insight for the solution. By contrast, our theoretical analysis has proved that the proposed instance aggregation module can effectively alleviate this problem.

The main contributions of our study can be highlighted as follows:

- We develop a low-rank audio-visual tensor fusion mechanism to capture the complex association between two modalities, which can efficiently generate informative audio-visual fused features.
- We propose a novel hierarchical scheme to encode temporal contextual features among video snippets with temporal cues for the first time in the video highlight detection. Experimental results demonstrate that our model outperforms competing approaches by a significant margin.
- We introduce an attention-gated instance aggregation module to formulate the video score and exploit inductive bias for topic-specific highlight detection. Moreover, theoretical analysis shows that it can ease the gradient vanishing problem during optimization effectively.

## 2. Related Work

**Video Highlight Detection** Video highlight detection has become increasingly popular in multimedia analysis with a great potential being deployed in practice, and many relevant studies have been explored in recent years. Previous approaches mainly focused on detecting highlights from sport videos [24, 31, 28]. Recently, various supervised methods were proposed to detect highlights of videos from the Internet [27] and first-person videos [34]. Gygli *et al.* [11] manually created Video-GIF pairs and utilized these pairs for ranking video segments in order to select the highlight segments. However, such supervised methods required manually labeled highlights which might not be easily collected on the Internet.

By contrast, weakly supervised and unsupervised methods can alleviate the problem of relying on highlight annotations. These methods can be divided into topic agnostic and topic specific methods. For topic agnostic methods, Yang *et al.* [32] employed category-aware reconstruction loss to narrow down the gaps between the highlight segment and the short-form video. Currently, topic specific methods [30, 12] trained on a set of videos sharing the same topic have achieved a noticeable gain in performance. Xiong *et al.* [30] mined the relation between video duration and highlight segments. Hong *et al.* [12] adopted rank loss between highlighted and non-highlight segments. However, these methods discarded the temporal dependencies in between the segments when making predictions so they often

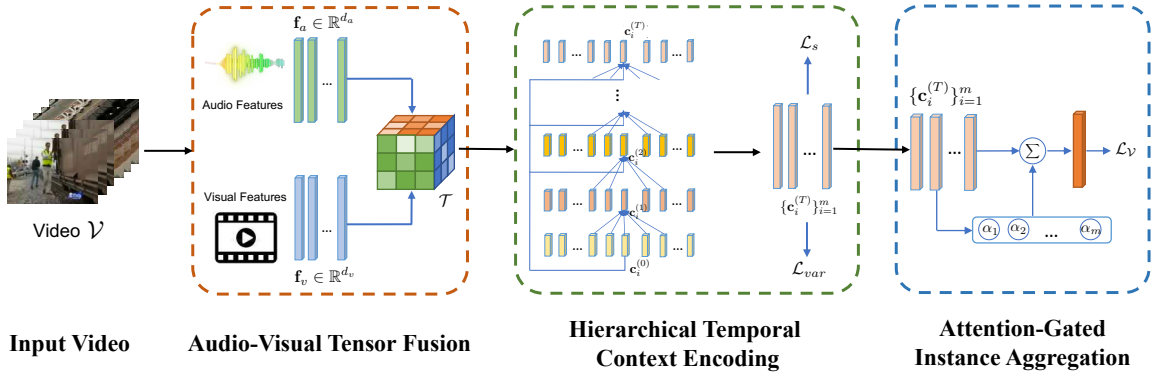



Figure 2. The overview of the proposed method. Note that $\mathbf{f}_a$ is the audio feature, $\mathbf{f}_v$ is the frame feature, $\mathcal{T}$ is the three-way projection tensor that fuses audio and frame features, and $\mathbf{c}_i^{(T)}$ is the temporal context feature for segment $v_i$.

led to sub-optimal performance for highlights detection. In order to explore the temporal cue of the video, we introduce hierarchical context modeling by incorporating contextual information in between adjacent segments to better localize the highlight segments.

**Video Summarization** Video summarization [7, 19, 35], a relevant task to video highlight detection, aims to select several representative and diverse segments from a video as the output summary. Yet highlight detection only selects the most interesting segment as the highlighted part, which does not require the integrity of the whole video. Recently, several deep learning based methods were proposed to generate high-quality video summary. For instance, Zhang *et al*. [37] utilized Determinantal Point Process (DPP) to enhance the diversity of the generated summary. Zhao *et al*. [39] applied a hierarchical structure of LSTM to encode the long-range temporal information among video segments. More recently, Li *et al*. [18] explored the global diversity for efficient video summarization. Besides, Rochan *et al*. [23] trained the model with unpaired samples from different sources under an unsupervised scenario.

## 3. The Proposed Approach

In this work, we explore topic-specific highlight detection with video-level labels only (*e.g*., surfing, playing guitar, etc.). We divide a candidate video uniformly into snippets (segments). In Section 3.2, we first introduce the audio-visual tensor fusion scheme with rank constraints in order to capture the association between video and audio signals and thus construct better feature representations of video segments. We then, in Section 3.3, build a contextual hierarchy for exploring temporal cues that are completely discarded by previous state-of-the-art methods. In addition, pairwise context information is considered for the sake of modeling the instance relation within positive videos or negative videos. Besides, we also propose an attention-gated instance aggregation module, in Section 3.4, which takes the encoded context features of all segments and corresponding highlight scores to estimate the highlight probability of the input video. In particular, the instance aggregation module alleviates the gradient vanishing issues leading to better convergence. The overview of our method is illustrated in Figure 2.

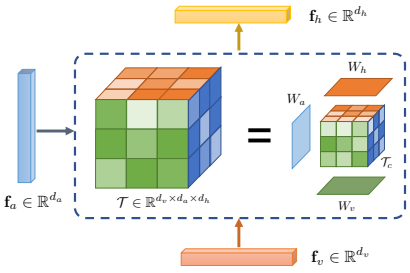
### 3.1. Preliminary

Given a set of $n$ videos $\mathcal{V} = \{\mathcal{V}^{(1)}, \mathcal{V}^{(2)}, \cdots, \mathcal{V}^{(n)}\}$, for each video $\mathcal{V}^{(i)} = \{v_s^{(i)}\}_{s=1}^{m}$ with $m$ segments, the observed label $y^{(i)} \in \{0, 1\}$ indicates whether this video contains the topic of interest. In this formulation, each segment $v_s^{(i)}$ is considered as an instance that we cannot assess whether it contains topic-specific highlight or not. If a video contains topic of interest segment, the video is treated as a positive video ($y^{(i)} = 1$), whereas a negative video ($y^{(i)} = 0$) solely consists of snippets that do not contain the specific topic. Formally, $\mathcal{V}_p = \{i \in [1, m] | y^{(i)} = 1\}$ is an index set of the positive videos, and $\mathcal{V}_n = \{i \in [1, n] | y^{(i)} = 0\}$ represents the indices of negative videos.

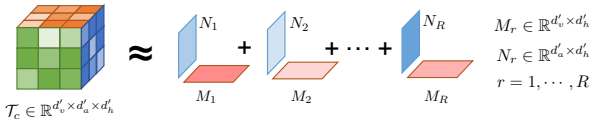
### 3.2. Audio-Visual Tensor Fusion Scheme

It has been widely proved that audio information can enrich the representation of video in various computer vision tasks [1, 13]. For highlight detection, the representation of video snippets should not only be learned from the appearance of frames but can also be learned from their corresponding audio tracks. Therefore, given a video segment, we develop an *audio-visual tensor fusion scheme* in order to exploit both audio and visual information for video representation learning.

Given the frame feature representation $\mathbf{f}_v \in \mathbb{R}^{d_v}$ and audio feature $\mathbf{f}_a \in \mathbb{R}^{d_a}$ for an arbitrary segment, to enhance the video feature representation, we model the high-level interactions between audio and visual features using bilinear pooling in a fully-parameterized way. The fused feature




(a) The mechanism of audio-visual tensor fusion procedure.




(b) The core tensor $\mathcal{T}_c$ is structured with a set of low-rank ($R$) matrices.

Figure 3. Overview of the audio-visual tensor fusion scheme.

$\mathbf{f}_h \in \mathbb{R}^{d_h}$ can be computed as follows

$$\mathbf{f}_h = (\mathcal{T} \times_1 \mathbf{f}_v) \times_2 \mathbf{f}_a, \tag{1}$$

where $\mathcal{T} \in \mathbb{R}^{d_v \times d_a \times d_h}$ is a three-way projection tensor that needs to be learned, and $\times_i$ is the mode-$i$ product operator. Although bilinear pooling can effectively model the pairwise interactions between the visual and the audio features, it also brings a substantial amount of trainable parameters giving the high cost of computation and memory [1].

To overcome this problem, inspired by tensor decomposition approaches [16], $\mathcal{T}$ can be factorized into a core tensor $\mathcal{T}_c$ and three factor matrices $W_v, W_a$ and $W_h$:

$$\mathcal{T} = ((\mathcal{T}_c \times_1 W_v) \times_2 W_a) \times_3 W_h \tag{2}$$

with $\mathcal{T}_c \in \mathbb{R}^{d'_v \times d'_a \times d'_h}, W_v \in \mathbb{R}^{d'_v \times d_v}, W_a \in \mathbb{R}^{d'_a \times d_a}$, and $W_h \in \mathbb{R}^{d'_h \times d_h}$. In particular, $d_a \ll d'_a, d_v \ll d'_v, d'_h \ll d_h$. Consequently, Eq. (1) can be reformed as:

$$\mathbf{f}_h = ((\mathcal{T}_c \times_1 (W_v \mathbf{f}_v)) \times_2 (W_a \mathbf{f}_a)) \times_3 W_h. \tag{3}$$

From the above equation, we can define $\tilde{\mathbf{f}}_v = W_v \mathbf{f}_v$ and $\tilde{\mathbf{f}}_a = W_a \mathbf{f}_a$. Meanwhile, we denote $\tilde{\mathbf{f}}_h = (\mathcal{T}_c \times_1 \tilde{\mathbf{f}}_v) \times_2 \tilde{\mathbf{f}}_a$ to be the pairwise interaction between the frame and the audio features in the embedding space. The fused feature $\mathbf{f}_h$ can thus be obtained by projecting $\tilde{\mathbf{f}}_h$ with the factor matrix $W_h$. The above decomposition procedure is depicted in Figure 3(a).

Besides, in order to balance the complexity and capacity of pairwise interactions, we incorporate a low-rank constraint to the procedure. According to the theory of matrix factorization [17], a high-rank matrix can be expressed by a set of rank-1 matrix. For each dimension $k \in [1, d'_h]$, the pairwise interaction between $\tilde{\mathbf{f}}_v$ and $\tilde{\mathbf{f}}_a$ can be formulated as

$$\tilde{\mathbf{f}}_h[k] = \tilde{\mathbf{f}}_v^\top \mathcal{T}_c[:, :, k] \tilde{\mathbf{f}}_a. \tag{4}$$

As shown in the Figure 3(b), we factorize each dimension of tensor $\mathcal{T}_c[:, :, k]$ as the combination of $R$ rank one matrices and plug it into Eq. (4), that is

$$\begin{aligned} \tilde{\mathbf{f}}_h[k] &= \tilde{\mathbf{f}}_v^\top \left( \sum_{r=1}^{R} M_r[:, k] N_r[:, k]^\top \right) \tilde{\mathbf{f}}_a \\ &= \sum_{r=1}^{R} \left( \tilde{\mathbf{f}}_v^\top M_r[:, k] \right) \left( \tilde{\mathbf{f}}_a^\top N_r[:, k] \right), \end{aligned} \tag{5}$$

where $M_r[:, k]$ and $N_r[:, k]$ are the $k$-th column vectors of $M_r \in \mathbb{R}^{d'_v \times d'_h}$ and $N_r \in \mathbb{R}^{d'_a \times d'_h}$. Therefore, we can rewrite $\tilde{\mathbf{f}}_h$ as the combination of $R$ rank one matrices as follows:

$$\tilde{\mathbf{f}}_h = \sum_{r=1}^{R} \left[ \left( \tilde{\mathbf{f}}_v^\top M_r \right) \circ \left( \tilde{\mathbf{f}}_a^\top N_r \right) \right]^\top, \tag{6}$$

where $\circ$ represents Hadamard product. In doing so, the fused multi-modal features can fully exploit audio and visual information for enhancing the video representation.

### 3.3. Hierarchical Temporal Context Encoding

Previous methods [30, 12] estimated individual highlight segments from input features directly, without taking the temporal cue into consideration. We argue that modeling temporal context is essential for the highlight detection. Intuitively, a highlight often leads to noticable changes in temporal context, *i.e.*, in adjacent segments, and therefore a useful indicator. We design a hierarchical temporal context encoding mechanism as a strategy to generate temporal contextual features by aggregating the consecutive segments. The proposed hierarchical temporal context encoding mechanism is able to locate highlight segments with the assistance of temporal cues that estimate individual highlight scores and variation among contextual features.

Initially, the fused features $\{\mathbf{f}_h^s\}_{s=1}^m$ are extracted from the video $\mathcal{V}^{(i)} = \{v_s^{(i)}\}_{s=1}^m$, where $m$ denotes the number of segments. Then, each anchored segment $\mathbf{f}_h^s$ and its adjacent segments with neighbor size $k$ are regressed as:

$$\mathbf{c}_s^{(t)} = \sum_{j=-k, \cdots, k,\ j \neq 0} W_j \mathbf{c}_{s+j}^{(t-1)} + W_0 \tilde{\mathbf{f}}_h^s + \mathbf{b}_\mathbf{c}, \tag{7}$$

where $\mathbf{c}_s^{(t)}$ is the contextual feature for segment $v_s^{(i)}$ at the $t$-th iteration, $k$ is the neighbor size, $W_j \in \mathbb{R}^{d_h \times d_h}$ is the learnable projection matrix for the adjacent $j$-th segment,

[1] A float32 tensor shaped in $1024 \times 1024 \times 2048$ takes up to 8GBytes, which is extremely memory consuming and therefore requires large GPU memories to compute.

and $\mathbf{b_c} \in \mathbb{R}^{d_h}$ is the bias term. Initially, we set $\mathbf{c}_s^{(0)} = \tilde{\mathbf{f}}_h^s$, and the above procedure executes $T$ times to encode the contextual information of segment $v_s^{(i)}$ and its $2k$ neighbor segments $(v_{s-k}^{(i)}, \cdots, v_{s-1}^{(i)}, v_{s+1}^{(i)}, \cdots, v_{s+k}^{(i)})$. Finally, we have the temporal contextual feature, denoted by $\mathbf{c}_s^{(T)}$ for segment $v_s^{(i)}$. Next, the temporal contextual feature is employed to predict the highlight confidence score, denoted as $p_s$, for the segment $v_s^{(i)}$, which is formulated as

$$p_s = \sigma(W_p \mathbf{c}_s^{(T)} + b_p), \tag{8}$$

where $\sigma(\cdot)$ is the Sigmoid activation function, $W_p \in \mathbb{R}^{1\times d_h}$ and $b_p \in \mathbb{R}$ are the linear transformation parameters to be learned. Furthermore, we model the variation between two adjacent segments with cosine similarity, represented as $(1 - \cos(\mathbf{c}_{s-1}^{(T)}, \mathbf{c}_s^{(T)}))/2$. Intuitively, the higher variation in between the segments, the larger cosine distance would be. In practice, we use the second order variation to measure the discrepancy between the anchored segment $\mathbf{c}_s^{(T)}$ and its adjacent segments as

$$\varphi_s = \left(2 - \cos(\mathbf{c}_{s-1}^{(T)}, \mathbf{c}_s^{(T)}) - \cos(\mathbf{c}_s^{(T)}, \mathbf{c}_{s+1}^{(T)})\right)/4. \tag{9}$$

To learn a highlight detection model, we anticipate that scores of highlight segments would be largely greater than those of non-highlight segments in topic-specific videos. However, it is unlikely to use temporal annotations to decide which segments would be the highlight under weakly supervised scenario. Likewise, scores of the negative videos are expected to be smaller than scores of the positive video, and scores of segments should be distributed uniformly and all close to zero in negative videos. Thus, for comparing the highlight confidence, we adapt the maximum score margin in between paired segments as follows:

$$\mathcal{X}_s = \max_{i,j=1,\cdots,m} |p_i - p_j|, \tag{10}$$

We use a hinge loss to formulate this as

$$\mathcal{L}_s = \max\{0, 1 - \frac{1}{|\mathcal{V}_p|}\sum_{i\in\mathcal{V}_p}\mathcal{X}_s^{(i)} + \frac{1}{|\mathcal{V}_n|}\sum_{j\in\mathcal{V}_n}\mathcal{X}_s^{(j)}\}. \tag{11}$$

By minimizing the score loss $\mathcal{L}_s$, model is encouraged to discriminate highlight segments from non-highlight ones. Similarly, we would also expect to maximize the discrepancy between the positive and negative videos. To do this, we introduce a variation loss $\mathcal{L}_{var}$ to enlarge the maximum variation margin between positive videos and negative videos with the second order variation $\varphi$ defined in Eq. (9). We define the maximum variation margin $\mathcal{X}_{var}$ and the variation loss $\mathcal{L}_{var}$ as:

$$\mathcal{X}_{var} = \max_{i,j=1,\cdots,m} |\varphi_i - \varphi_j|, \tag{12}$$

$$\mathcal{L}_{var} = \max\{0, 1 - \frac{1}{|\mathcal{V}_p|}\sum_{i\in\mathcal{V}_p}\mathcal{X}_{var}^{(i)} + \frac{1}{|\mathcal{V}_n|}\sum_{j\in\mathcal{V}_n}\mathcal{X}_{var}^{(j)}\}. \tag{13}$$

Moreover, considering the scores are sparse in both positive and negative videos, we add a sparsity constraint on above loss functions with the weighting factor $\beta$, that is

$$\mathcal{L}_{ins} = \mathcal{L}_s + \mathcal{L}_{var} + \frac{\beta}{n}\sum_{i=1}^{n}\left(\|\mathcal{X}_s^{(i)}\|_1 + \|\mathcal{X}_{var}^{(i)}\|_1\right). \tag{14}$$

### 3.4. Attention-Gated Instance Aggregation

Early MIL related works [2, 4] pointed out that simply applying the video label to segment label might be inaccurate since labels in positive video could be noisy. Instead of assigning video label to each individual segment directly, we aggregate instance scores to estimate the probability that a video belongs to specific topic. Following the conventional Noisy-OR MIL [36], we can express the probability for the video $\mathcal{V}^{(i)}$ as:

$$\hat{p}_{\mathcal{V}}^{(i)} = 1 - \prod_{s=1}^{m}(1 - p_s^{(i)}), \tag{15}$$

where $p_s^{(i)}$ is the confidence score for segment $v_s^{(i)}$ of video $\mathcal{V}^{(i)}$. Then binary cross entropy loss is applied for video-level supervision, and it can be defined as:

$$\mathcal{L}_{\mathcal{V}} = -\frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)}\log\hat{p}_{\mathcal{V}}^{(i)} + (1 - y^{(i)})\log(1 - \hat{p}_{\mathcal{V}}^{(i)})\right]. \tag{16}$$

However, optimization of Eq. (15) would suffer from the gradient vanishing issue. For a positive video, the gradient of $\mathcal{L}_{\mathcal{V}}$ is computed as:

$$\frac{\partial\mathcal{L}}{\partial p_s^{(i)}} = \frac{\partial\mathcal{L}}{\partial\hat{p}_{\mathcal{V}}^{(i)}}\left(\prod_{k=1,k\neq s}^{m}(1 - p_k^{(i)})\right). \tag{17}$$

From above derivation, we know when there exists one segment of which confidence score is close to 1, gradients from other segments will be suppressed. This is divergent from the assumption that multiple highlight segments are tailored from the original videos.

To alleviate this problem , we aggregate the temporal context features of all segments in the video to generate the

| Topic | Supervised Methods | | Weakly Supervised Methods | | | |
|---|---|---|---|---|---|---|
| | GIFs | LSVM | RRAE | LIM-s | MINI-Net* | Ours |
| dog | 0.308 | **0.60** | 0.49 | 0.579 | 0.5768 | 0.5538 |
| gymnastics | 0.335 | 0.41 | 0.35 | 0.417 | 0.5737 | **0.6266** |
| parkour | 0.540 | 0.61 | 0.50 | 0.670 | 0.6975 | **0.7088** |
| skating | 0.554 | 0.62 | 0.25 | 0.578 | 0.5219 | **0.6906** |
| skiing | 0.328 | 0.36 | 0.22 | 0.486 | 0.5390 | **0.6005** |
| surfing | 0.541 | 0.61 | 0.49 | **0.651** | 0.5931 | 0.5976 |
| Average | 0.464 | 0.54 | 0.38 | 0.564 | 0.5837 | **0.6297** |

Table 1. Performance comparison (mAP score) on YouTube Highlights dataset. * indicates our implementation trained on self-collected dataset. See Supplementary Material for details.

video score as:

$$\hat{p}_{\mathcal{V}}^{(i)} = \sigma\left(W_p \sum_{j=1}^{m} \alpha_j \mathbf{c}_j^{(T)} + b_p\right), \tag{18}$$

$$\alpha_j = \frac{\exp(W_c \mathbf{c}_j^{(T)} + b_c)}{\sum_{q=1}^{m} \exp(W_c \mathbf{c}_q^{(T)} + b_c)}, \tag{19}$$

where $\{\alpha_j\}_{j=1}^{m}$ are weighted factors, $W_c$ and $b_c$ are parameters to be learned, and $W_p$ and $b_p$ share the same parameters in Eq. (8). It can be theoretically proved that proposed instance aggregation method with Equation (18) can ease the gradient vanishing problem, which is provided in the Appendix. Besides, the visualization of the gradient $\partial\mathcal{L}/\partial p_s^{(i)}$ in the Appendix demonstrate that the area of non-zero gradient has significantly enlarged via our method.

Finally, we apply Eq. (16) to compute video classification loss for videos. By combining Eq. (14) and Eq. (16), we can obtain the total loss for the proposed model:

$$\mathcal{L} = \mathcal{L}_{ins} + \mathcal{L}_{\mathcal{V}}. \tag{20}$$

# 4. Experiments

In this section, we evaluate highlight detection performance for the proposed model extensively on two public datasets and compare with other state-of-the-art methods. More experimental results and implementation details are reported and analyzed in the Supplementary Material.

## 4.1. Datasets and Metrics

We have evaluated different highlight detection approaches on two benchmark datasets, *i.e.*, YouTube Highlights [27] and TVSum [25]. YouTube Highlights dataset includes six topic-specific categories: *dog, gymnastics, parkour, skating, skiing*, and *surfing*, where each topic contains about 100 videos and the total accumulated length is 1,430 minutes. TVSum has 50 user videos collected from *YouTube* with 10 topic-specific queried tag including: *changing vehicle tires, grooming an animal, parade, flash mob gathering*, and others. Because only frame-level importance scores are provided in TVSum, following the current evaluation protocol [30, 12], we average the frame-level importance scores to obtain segment-level scores and select the top 50% segments for each video as the human-created highlights. We compare our model predicted highlight segments with the human-created summaries and report mean average precision (mAP) for both datasets.

## 4.2. Quantitative Results

We compare our proposed model with numerous state-of-the-art video highlight detection methods including supervised and weakly supervised approaches. For supervised methods, we compare with Video2GIF approach [11], Latent SVM [27], KVS [22], DPP [9], sLSTM [37], and SM [10]. Note that these methods require detailed temporal annotations to be trained while our method does not. Additionally, a number of weakly supervised methods are also compared, including RRAE [33], MBF [5], SMRS [6], Quasi [14], CVS [21], SG [19], LIM-s [30], VESD [3], DSN [20], and MINI-Net [12]. All of these approaches are evaluated using the same metrics mentioned above.

**Results on YouTube Highlights dataset** Table 1 summarizes the experimental results of different state-of-the-art methods. We can find that our method achieves the best performance with respect to the average mAP over all six topics. In particular, our method yields a better performance by 4.6% higher than the multi-modal method MINI-Net [12]. This verifies the benefit of adding contextual information into video representation learning via the hierarchical temporal encoding. Also, with our proposed low-rank decomposition technique, our model is capable of exploiting audio-visual structures efficiently and thus learn more discriminative video representations, which we believe this is also vital to the improvement. Besides, it can be observed that our approach outperforms the supervised methods, *i.e.*, GIFs [11] and LSVM [27], by a large margin that further proves event-specific temporal annotations are trivial and our weakly supervised model is able to leverage the unlabeled video segments for capturing the highlights precisely. Since our method does not require any human-created annotations, our method is more adaptable for real world scenarios for videos with hashtags on social media.

**Results on TVSum dataset** Table 2 presents the experimental results on TVSum dataset [25]. Our method outperforms all of the compared methods by a large margin. In particular, we found MINI-Net [12] is the most competitive multi-modal weakly supervised method, which also regards individual video-audio pairs as instances to handle video structure. Our method achieves a relative gain of 6.62% on average Top-5 mAP than MINI-Net. This result further reinforces the advantages of the temporal cue in video and audio clips contributing to better highlight detection. We demonstrate that modeling temporal context is useful and essential for highlight detection. For example, for parkour

| Topic | Supervised Methods | | | | | Weakly Supervised Methods | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | KVS | DPP | sLSTM | SM | SMRS | Quasi | MBF | CVS | SG | LIM-s | DSN | VESD | MINI-Net* | Ours |
| Vehicle Tire | 0.353 | 0.399 | 0.411 | 0.415 | 0.272 | 0.336 | 0.295 | 0.328 | 0.423 | 0.559 | 0.373 | 0.447 | 0.7854 | **0.8501** |
| Vehicle Unstuck | 0.441 | 0.453 | 0.462 | 0.467 | 0.324 | 0.369 | 0.357 | 0.413 | 0.472 | 0.429 | 0.441 | 0.493 | 0.5659 | **0.7144** |
| Grooming Animal | 0.402 | 0.457 | 0.463 | 0.469 | 0.331 | 0.342 | 0.325 | 0.379 | 0.475 | 0.612 | 0.428 | 0.496 | 0.7360 | **0.8187** |
| Making Sandwich | 0.417 | 0.462 | 0.477 | 0.478 | 0.362 | 0.375 | 0.412 | 0.398 | 0.489 | 0.54 | 0.436 | 0.503 | 0.7529 | **0.7859** |
| Parkour | 0.382 | 0.437 | 0.448 | 0.445 | 0.289 | 0.324 | 0.318 | 0.354 | 0.456 | 0.604 | 0.411 | 0.478 | 0.7687 | **0.8021** |
| Parade | 0.403 | 0.446 | 0.461 | 0.458 | 0.276 | 0.301 | 0.334 | 0.381 | 0.473 | 0.475 | 0.417 | 0.485 | 0.6325 | **0.7552** |
| Flash Mob | 0.397 | 0.442 | 0.452 | 0.451 | 0.302 | 0.318 | 0.365 | 0.365 | 0.464 | 0.432 | 0.412 | 0.487 | 0.6115 | **0.7155** |
| Beekeeping | 0.342 | 0.395 | 0.406 | 0.407 | 0.297 | 0.295 | 0.313 | 0.326 | 0.417 | 0.663 | 0.368 | 0.441 | 0.7560 | **0.7727** |
| Bike Tricks | 0.419 | 0.464 | 0.471 | 0.473 | 0.314 | 0.327 | 0.365 | 0.402 | 0.483 | 0.691 | 0.435 | 0.492 | 0.7556 | **0.7860** |
| Dog Show | 0.394 | 0.449 | 0.455 | 0.453 | 0.295 | 0.309 | 0.357 | 0.378 | 0.466 | 0.626 | 0.416 | 0.488 | 0.6555 | **0.6812** |
| Average | 0.398 | 0.447 | 0.451 | 0.461 | 0.306 | 0.329 | 0.345 | 0.372 | 0.462 | 0.563 | 0.424 | 0.481 | 0.7020 | **0.7682** |

Table 2. Performance comparison (Top-5 mAP score) on TVSum dataset. Our method surpasses all of the compared methods significantly. * indicates our implementation trained on self-collected dataset.

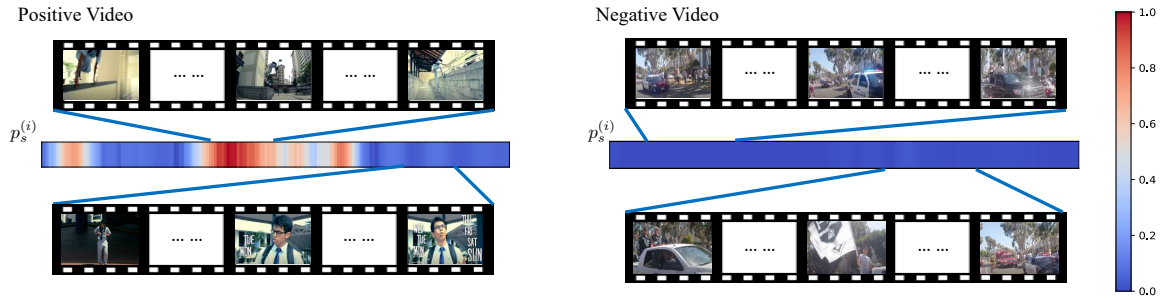



Figure 4. The example highlight prediction of our method. The highlight detection model is aimed to detect "Parkour" highlight. The positive video is tagged with "Parkour", and the negative one is marked as "Parade".

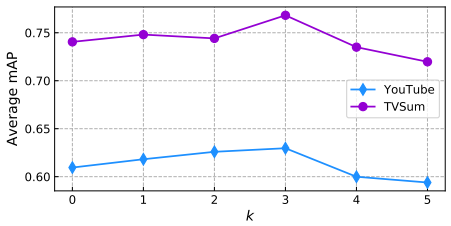



Figure 5. Variations in performance by changing the neighbor size $k$. It shows that our model achieves the best performance with $k = 3$.

scenarios, the highlight would occur when someone jumps in and moves fast from one point to another leading to a drastic change of foreground context. We believe the transition of activities is an important temporal cue indicating the appearance of the highlight moments. To further verify our intuition, we visualize, in Figure 4, the segment scores in a positive video and a negative video for parkour. We can see the segment scores in the negative video are almost uniform and close to 0. Meanwhile, we observe that in the positive video, the most discriminative highlight segments are captured with high confidence even if they are distributed sparsely and fragmentally in time. And also, we can see the segments that are irrelevant to the topic *i.e.* "parkour" tend to have very low probabilities. All these evidences verify the effectiveness of our contrastive learning design with the loss function $\mathcal{L}_{ins}$.

| Exp. | $\mathcal{L}_s$ | $\mathcal{L}_{var}$ | $\mathcal{L}_{\mathcal{V}}$ | YouTube | TVSum |
|---|---|---|---|---|---|
| 1 | √ | | | 0.5625 | 0.6805 |
| 2 | | √ | | 0.5749 | 0.7161 |
| 3 | | | √ | 0.5944 | 0.7326 |
| 4 | √ | √ | | 0.6011 | 0.7470 |
| 5 | √ | √ | √ | **0.6297** | **0.7682** |

Table 3. Average mAP comparison of different components on YouTube and TVSum datasets.

### 4.3. Ablation Studies

In this section, we study the behavior of each component of our model under different conditions.

**Impact of neighbor size $k$.** Firstly, we investigate the effect of tuning the neighbouring size $k$ for temporal context encoding, as depicted in Figure 5 . It can be observed that with the increment of $k$, the average mAP increases as the result of aggregating more contextual features, until the point *i.e.* $k = 3$ (where the best performance is attained), the model no longer benefits from adding excessive contextual information. This suggests incorporating localised temporal context does boost the highlight detection effectively, while modeling long-term decencies is trivial and adding excessive context may hinder the model from learning the discriminative highlight features and consequentially, leads to degenerated detection performance.

**Impact of loss terms.** Furthermore, we have examined the contribution of individual loss terms and the results are shown in Table 3. As can be seen from the table, the combination of variation loss and score loss (Exp.4) improves the performance by at least 1.1% and 2.2%, on YouTube and TVSum respectively, in comparison with those that use them individually (Exp.1 and Exp.2). In addition, we found the video level supervision (Exp.3) is also vital to the detection, which boosts the ultimate detection results (Exp.5) by 2.8% and 2.1% on YouTube and TVSum respectively, since it not only estimates the overall probability that a video contains highlights but also utilize inductive bias for topic-specific highlight detection.

**Impact of audio-visual fusion schemes.** We compare various audio-visual fusion schemes to explain the performance variation in Table 4. The most common practice of fusing two modalities is a summation of two kinds of feature vectors or concatenating them directly. As shown in the first two rows (*i.e*. summation and concatenation), they achieve similar performance due to the limited expressive power of linear models. To alleviate this problem, [12] proposed a submodule multi-layer perceptron to model the local complex feature interaction, and it gains improvement by 1.5% and 1.4% on YouTube and TVSum respectively. The compact bilinear pooling method [8] pushes up the performance even further as can be seen in the fourth row of Table 4 that it achieves the best performance by far. However, one drawback of the method is it introduces significantly more trainable parameters comparing to the above methods, which makes it prone to be overfitted in practice, while the low-rank constraint in our method serves as the regularization that mitigates the risk of overfitting [15]. In comparison, our model surpasses the compact bilinear pooling method [8] while saving 85% of the trainable parameters. This strongly indicates that regularizing a bilinear model through our low-rank decomposition approach provides an effective trade-off between model capacity and the number of parameters. In addition, factorized bilinear pooling method [38] constrains the whole tensor $\mathcal{T}$ to be of low rank, which achieves comparable results as compact bilinear pooling. However, it merely projects the audio and visual features into a shared $R$-dimensional space by computing element-wise production, which restricts the interaction between the intra-dimension of two modalities. By contrast, instead of constraining the whole tensor $\mathcal{T}$, our method makes the low-rank constraint on the core tensor $\mathcal{T}_c$ leading to better performance than [38], which allows the visual and audio features to be modeled into distinctive projection spaces, leading to better audio-visual representation.

**Impact of video score modeling.** To further prove that our proposed instance aggregation module is more suitable for highlight detection. We compare several most popular video score modeling methods in the context of highlight detection. DMIL-RM [26] introduced ranking methods to rank the scores of video segments without direct video-level supervision. However, we have demonstrated in earlier sessions that adding video-level supervision during training helps improve the detection performance by a lot. It is clearly shown that our method outperforms both the Noisy-OR video score aggregating method and MINI-Net approach [12] by a large margin (4.3% and 6.6%) on two datasets respectively. The results in Table 5 demonstrate that our proposed instance aggregation method is superior to the state-of-the-art video classification methods.

| Methods | $\lvert\Theta\rvert$ | YouTube | TVSum |
|---|---|---|---|
| Summation | 0.3288M | 0.5809 | 0.7259 |
| Concatenation [29] | 0.3271M | 0.5953 | 0.7187 |
| SubModule MLP [12] | 0.7886M | 0.6103 | 0.7328 |
| Compact Bilinear Pooling [8] | 8.3887M | 0.6256 | 0.7509 |
| Factorized Bilinear Pooling [38] | 0.9197M | 0.6147 | 0.7511 |
| Ours | 1.2499M | **0.6297** | **0.7682** |

Table 4. Average mAP comparison of different audio-visual fusion methods on YouTube and TVSum datasets. MLP denotes multi-layer perceptron, and $\lvert\Theta\rvert$ represents the number of parameters.

| | DMIL-RM [26] | Noisy-OR [36] | MINI-Net [12] | Ours |
|---|---|---|---|---|
| YouTube | 0.5235 | 0.5868 | 0.5836 | **0.6297** |
| TVSum | 0.6731 | 0.6838 | 0.7020 | **0.7682** |

Table 5. Average mAP comparison of different video score modeling methods on YouTube and TVSum datasets.

## 5. Conclusion

This paper has proposed a novel video highlight detection model, which integrates audio and visual features with an efficient low-rank tensor fusion mechanism. To exploit the temporal cue in the video, the model encodes the adjacent segments to generate temporal context features in a hierarchical way that we believe variation in the contextual features is considered to be key to characterise the highlight moments in topic-specific videos. During the optimization stage, the video score is reformulated to alleviate the gradient vanishing problem. Furthermore, we conducted extensive experiments on two publicly available datasets and the results have verified the effectiveness and the superiority of our approach compared with other state-of-the-art methods.

## Acknowledgements

Thanks to Guodun Li, and Hongxiang Chen for valuable comments and advice for this work. This work was supported in part by the National Natural Science Foundation of China under Grants 61872122, 61502131.

## References

[1] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference*

*on Computer Vision*, pages 609–617, 2017. 3

[2] Avrim Blum and Adam Kalai. A note on learning from multiple-instance examples. *Machine learning*, 30(1):23–29, 1998. 2, 5

[3] Sijia Cai, Wangmeng Zuo, Larry S Davis, and Lei Zhang. Weakly-supervised video summarization using variational encoder-decoder and web prior. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 184–200, 2018. 1, 6

[4] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018. 2, 5

[5] Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3584–3592, 2015. 6

[6] Ehsan Elhamifar, Guillermo Sapiro, and Rene Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1600–1607. IEEE, 2012. 6

[7] Hui Fang, Jianmin Jiang, and Yue Feng. A fuzzy logic approach for detection of video shot boundaries. *Pattern Recognition*, 39(11):2092–2100, 2006. 3

[8] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016. 2, 8

[9] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *Advances in neural information processing systems*, pages 2069–2077, 2014. 1, 6

[10] Michael Gygli, Helmut Grabner, and Luc Van Gool. Video summarization by learning submodular mixtures of objectives. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3090–3098, 2015. 6

[11] Michael Gygli, Yale Song, and Liangliang Cao. Video2gif: Automatic generation of animated gifs from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1001–1009, 2016. 1, 2, 6

[12] Fa-Ting Hong, Xuanteng Huang, Wei-Hong Li, and Wei-Shi Zheng. Mini-net: Multiple instance ranking network for video highlight detec- tion. In *European Conference on Computer Vision*, 2020. 1, 2, 4, 6, 8

[13] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pages 4193–4202, 2017. 3

[14] Gunhee Kim, Leonid Sigal, and Eric P Xing. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4225–4232, 2014. 6

[15] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 8

[16] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009. 2, 4

[17] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. 4

[18] Ping Li, Qinghao Ye, Luming Zhang, Li Yuan, Xianghua Xu, and Ling Shao. Exploring global diverse attention via pairwise temporal relation for video summarization. *Pattern Recognition*, 111:107677, 2021. 3

[19] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 202–211, 2017. 3, 6

[20] Rameswar Panda, Abir Das, Ziyan Wu, Jan Ernst, and Amit K Roy-Chowdhury. Weakly supervised summarization of web videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3657–3666, 2017. 1, 6

[21] Rameswar Panda and Amit K Roy-Chowdhury. Collaborative summarization of topic-related videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7083–7092, 2017. 6

[22] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. Category-specific video summarization. In *European conference on computer vision*, pages 540–555. Springer, 2014. 1, 6

[23] Mrigank Rochan and Yang Wang. Video summarization by learning from unpaired data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7902–7911, 2019. 3

[24] Yong Rui, Anoop Gupta, and Alex Acero. Automatically extracting highlights for tv baseball programs. In *Proceedings of the eighth ACM international conference on Multimedia*, pages 105–115, 2000. 1, 2

[25] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015. 6

[26] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018. 8

[27] Min Sun, Ali Farhadi, and Steve Seitz. Ranking domain-specific highlights by analyzing edited videos. In *European conference on computer vision*, pages 787–802. Springer, 2014. 2, 6

[28] Hao Tang, Vivek Kwatra, Mehmet Emre Sargin, and Ullas Gargi. Detecting highlights in sports videos: Cricket as a test case. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2011. 1, 2

[29] Peng Wu, jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also

listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 8

[30] Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. Less is more: Learning highlight detection from video duration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1258–1267, 2019. 1, 2, 4, 6

[31] Ziyou Xiong, Regunathan Radhakrishnan, Ajay Divakaran, and Thomas S Huang. Highlights extraction from sports video based on an audio-visual marker detection framework. In *2005 IEEE International Conference on Multimedia and Expo*, pages 4–10. IEEE, 2005. 1, 2

[32] Huan Yang, Baoyuan Wang, Stephen Lin, David Wipf, Minyi Guo, and Baining Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *Proceedings of the IEEE international conference on computer vision*, pages 4633–4641, 2015. 2

[33] Huan Yang, Baoyuan Wang, Stephen Lin, David Wipf, Minyi Guo, and Baining Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *Proceedings of the IEEE international conference on computer vision*, pages 4633–4641, 2015. 6

[34] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 982–990, 2016. 2

[35] Li Yuan, Francis EH Tay, Ping Li, Li Zhou, and Jiashi Feng. Cycle-sum: Cycle-consistent adversarial lstm networks for unsupervised video summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9143–9150, 2019. 3

[36] Cha Zhang, John C Platt, and Paul A Viola. Multiple instance boosting for object detection. In *Advances in neural information processing systems*, pages 1417–1424, 2006. 5, 8

[37] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *European conference on computer vision*, pages 766–782. Springer, 2016. 3, 6

[38] Yuanyuan Zhang, Zi-Rui Wang, and Jun Du. Deep fusion: An attention guided factorized bilinear pooling for audio-video emotion recognition. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 8

[39] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. Hierarchical recurrent neural network for video summarization. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 863–871, 2017. 3