MINI-Net: Multiple Instance Ranking Network for Video Highlight Detection

Fa-Ting Hong[1,4,5], Xuanteng Huang[1], Wei-Hong Li[3], and Wei-Shi Zheng[1,2,5] ⋆

[1] School of Data and Computer Science, Sun Yat-sen University, China
[2] Peng Cheng Laboratory, Shenzhen 518005, China
[3] VICO Group, University of Edinburgh, United Kingdom
[4] Pazhou Lab
[5] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
{hongft3,huangxt57}@mail2.sysu.edu.cn, w.h.li@ed.ac.uk, wszheng@ieee.org

Abstract. We address the weakly supervised video highlight detection problem for learning to detect segments that are more attractive in training videos given their video event label but without expensive supervision of manually annotating highlight segments. While manually averting localizing highlight segments, weakly supervised modeling is challenging, as a video in our daily life could contain highlight segments with multiple event types, *e.g.*, skiing and surfing. In this work, we propose casting weakly supervised video highlight detection modeling for a given specific event as a multiple instance ranking network (MINI-Net) learning. We consider each video as a bag of segments, and therefore, the proposed MINI-Net learns to enforce a higher highlight score for a positive bag that contains highlight segments of a specific event than those for negative bags that are irrelevant. In particular, we form a max-max ranking loss to acquire a reliable relative comparison between the most likely positive segment instance and the hardest negative segment instance. With this max-max ranking loss, our MINI-Net effectively leverages all segment information to acquire a more distinct video feature representation for localizing the highlight segments of a specific event in a video. The extensive experimental results on three challenging public benchmarks clearly validate the efficacy of our multiple instance ranking approach for solving the problem.

1 Introduction

In our daily life, people like to share their shining moments by posting videos on social media platforms, such as *YouTube* and *Instagram*. These well-edited videos in the social media platforms can quickly attract audience and convey an owner's experience. However, behind a well-credited video, there is the owner's heavy workload, as producing highlight clips from a lengthy video by hand is a time-consuming and laborious task. Therefore, it would be highly demanded

⋆ Corresponding author

for developing an automated tool to cut out highlights from a lengthy video, automatically generating a highlight short-form video.

Recently, video highlight detection has attracted an increasing amount of attention. Existing methods are mainly divided into two strategies. The first category casts the video highlight detection into a supervised learning problem [9,31,12]. Given both unedited videos and their highlight annotations labelled manually, a ranking net is trained to score segments in videos such that the highlight segments have higher highlight scores than those non-highlight segments in the video. For example, in [9], they proposed a robust deep RankNet to generate a rank list of segments according to their suitability as graphic interchange format (GIF) and designed an adaptive Huber loss to resist the noise effect caused by the outlier data. However, these methods work in the supervised learning manner and requires massive annotation on highlights in the training videos, which is hard and costly to be collected.

The second strategy treats video highlight detection as a weakly supervised recognition task [30,27,25]. Given certain events' videos, they treat short-form videos as a collection of highlights, while long-form videos contain a high proportion of non-highlights. Specially, Xiong et al. [27] designed a model that learns to predict the relations between highlight segments and non-highlight segments of the same event such that the highlight segments would have higher scores than non-highlight segments in the same event. Additionally, the work [30] employs the auto-encoder structure to narrow the reconstruction error of segments in short-form videos, which are considered as highlights. However, video highlight detection remains as a challenging problem, as in real-world scenarios an unedited video in social media platforms may contain highlights of more than one event, and the above mentioned detectors that are trained on videos of target event cannot well filter out the highlights of the other events. Without such human annotation, it is hard and indeed challenging to locate the real highlight of a target event in a video and perform specific learning.

In this work, we provide a new and effective approach for solving the weakly supervised setting: even though the exact highlight annotations of a video are not available, the label whether a video has a type of highlight is provided. In such a weakly supervised setting, we know that there exists a segment of a video that corresponds to a target highlight, but we also understand that there exist other segments of the video that do not correspond to the target highlight. To cope with this setting, we consider each video as a bag, and each bag contains a set of segments of the video (*i.e.*, the video segments are treated as instances in each bag). Therefore, we cast the weakly supervised highlight detection as a multiple instance learning problem and proposed a Multiple InstaNce rankIng NETwork (MINI-Net) for video highlight detection. As shown in Figure 1, for each type of highlight event, we construct positive bags using the videos that that contain the target highlights (*e.g.*, surfing), and the videos that contain other irrelevant highlight events but not the target event (*e.g.*, dog show) are used to form the negative bags. For such bag-level classification, we introduce two objective functions, *i.e.*, max-max ranking loss and binary bag event classification

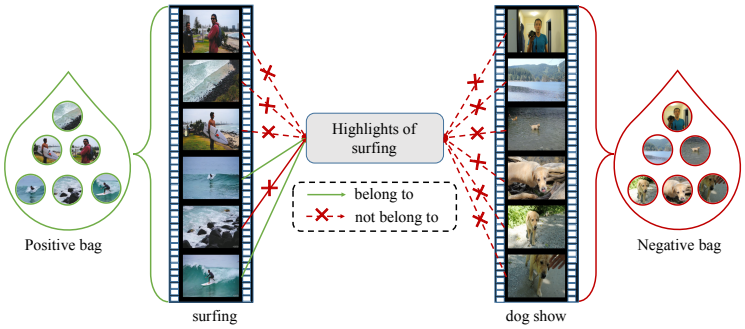

Fig. 1: To train a model to detect surfing's highlights, we can collect unannotated videos of various events from the internet using query tags. Although the highlight annotations (*i.e.*, labels telling which segments are highlights) are not available, we know that the videos with the "surfing" tag (*e.g.*, the left video) potentially contain "surfing" highlights, while the videos of other event tags would not have highlights about surfing (*e.g.*, videos of "dog show" shown on the right side would not have highlights of surfing). We cast video highlight detection as a multiple instance learning problem, where we can treat videos of "surfing" as positive bags as they contain highlights of "surfing", while videos of other events are negative bags as they do not have highlights of "surfing".

loss, to effectively train the MINI-Net. In particular, the max-max ranking loss is designed to acquire a reliable relative comparison between the most likely positive segment instance and the most hard negative segment instance. And meanwhile, minimizing the binary bag event classification loss enforces model to produce more informative bag representation for the specific event. To our best knowledge, this is the first to develop a multiple instance learning approach for weakly supervised video highlight detection.

In addition to the bag classification module, our MINI-Net also consists of two other modules: vision-audio fusion module and highlight estimation module. The vision-audio fusion module leverages both vision features and audio features, which is beneficial as, inspired by [1] learning about video segments both visually and aurally can produce more informative features. The highlight estimation module utilizes these features to estimate the highlight score for them. We aggregate all instance features weighted by their immediate highlight scores to generate the bag feature for the bag classification module.

In our experiments, we compared the proposed model with other related methods for three challenging public video highlight detection benchmarks. *i.e.*, YouTube Highlights dataset [21], TVSum dataset [19] and CoSum dataset [4]. Additionally, we have conducted an ablation study to investigate the effect of the proposed max-max ranking loss and bag classification module and validate the use of audio features and vision features. The experimental results show that our

proposed model achieves a state-of-the-art performance for three public datasets and verify its efficacy for video highlight detection.

2 Related Work

- Video Highlight Detection. In recent years, video highlight detection has attracted increasing attention. Researchers have mainly developed approaches to detect highlights of sport videos [24,29,22] in the early stage. Recently, supervised video highlight detection has been proposed for general videos from social media platforms [21] and first-person videos [31]. These methods require massive annotations for training videos which is a time-consuming and laborious task. The Video2GIF [9] method, learns from manually created GIF-video pairs, proposed a robust deep RankNet to generate a ranked list of segments according to their suitability as a GIF, and used an adaptive Huber loss to suppress the noise effect caused by outlier data. Weakly supervised methods on video highlight detection can effectively reduce the pressure of manual labeling. More recently, methods that trained on a collection of videos of the same topic [30,27] gain a satisfactory performance. They leverage category-aware reconstruction loss [30] to identify the highlights or exploit the video duration as an implicit supervision [27].

Like these weakly supervised video highlight detection methods, our approach also tailors highlights to the topic event. However, existing methods cannot filter the highlights of irrelevant events as they are trained on specific event videos. Unlike existing methods, our approach formulates a multiple instance learning framework to tackle the video highlight detection problem. Treating videos of other events as negative bags in our framework and using proposed max-max ranking loss to enlarge the gap between instances of target event and those of other events in terms of highlight scores can help to filter the segments of irrelevant events and detect the highlights of the target event in a general video.

- Video Summarization. Video summarization [19,2,28,14], which is highly related to video highlight detection, outputs a video summary by the estimated importance of segments. Different from video highlight detection, video summarization focuses on the integrity of the video story. Mahasseni et al. [14] proposed an adversarial long short-term memory (LSTM) network, consisting of a summarizer and a discriminator, to regularize the consistency between the story of the summary and the original video. In addition, by using deep reinforcement learning, [33] formulated video summarization as a sequential decision-making process, rewarded by the diversity and representativeness of the generated video summaries. Recently, [2] presented a generative modeling framework, which contains two important components: a variational auto-encoder for learning the latent semantics from web videos and an encoder-attention-decoder for saliency estimation of the raw video and the summary generation, to learn the latent semantic video representations to bridge the benchmark data and web data. Different from video summarization, our approach selects the highlight segments by comparing the instances in the training pair, which consists of one most likely highlight an instance from the positive bag and one hard non-highlight instance

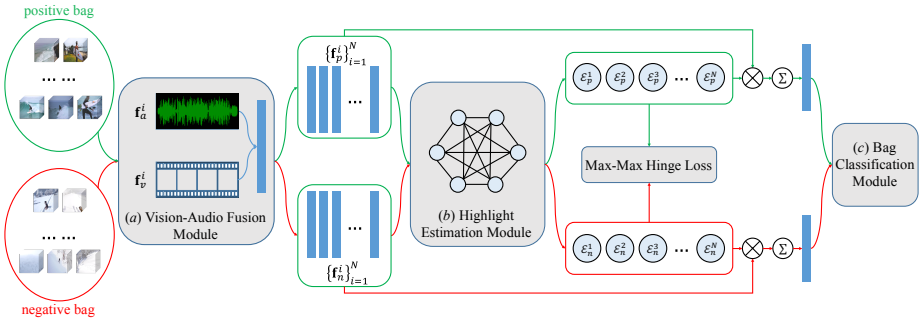

Fig. 2: Illustration of our proposed MINI-Net. We feed two bags, positive bag and negative bag, into vision-audio Fusion Module (Figure (a)) to encode the vision-audio fusion feature. The highlight estimation module (Figure (b)) takes as input these features to estimate the highlight scores. Beyond this, the immediate highlight scores and vision-audio fusion features are fed into the bag classification module (Figure (c)) for bags' event category classification. The max-max ranking loss is designed to ensure that the score of the segment in the positive bag with highest score is higher than the score of the segment in the negative bag with the highest score with a margin. Beyond this, the binary cross entropy loss is adopted for bags' event classification.

from the negative bag. The inherent characteristics that there is at least one positive instance in the positive bag and instances are all negative in negative bag improve our MINI-Net's distinguishing power for detecting highlights.

- Multiple Instance Learning. The multiple instance learning (MIL) is a form of weakly supervised learning in which the training instances are arranged in sets, called bags, and a label is provided for the entire bag. The field of MIL has generated a large amount of interest and is still growing [26,5,20,11,3,23,15]. Ilse et al. [11] proposed a neural-network-based permutation-invariant aggregation operator, a gated attention mechanism that provides insight into the contribution of each instance to the bag label, to produce bag features. Considering normal and anomalous videos as bags and video segments as instances in multiple instance learning framework, the work in [20] develops a deep multiple instance ranking framework to predict high anomaly scores for anomalous video segments.

In this work, the objective of multiple instance learning is different from the above, and ours is for solving weakly supervised video highlight detection, which has not been attempted before, and some of the above MIL methods may not be applicable or effective for our problem. In addition, unlike the above MIL methods that only explore the relations among instances of a bag to encode informative bag representation and the bag classification for learning, we introduce a max-max ranking loss to acquire a reliable relative comparison between the most likely positive segment instance and the hardest negative segment instance. This enables our method for more effectively distinguishing highlight from videos, which is verified in our experiments.

3 Approach

In this work, we explore event-specific[6] video highlight detection under weakly supervised setting; that is we trained on unannotated data samples, in each of which the event-specific highlight exists but the annotation on its location is not specified. In such a weakly supervised setting, we know there exists a segment of a video corresponding to an event-specific highlight, but we also understand that there exist other segments of the video not corresponding to the event-specific highlight but probably others. Therefore, we cast the weakly supervised highlight detection as a multiple instance learning problem, and develop a Multiple InstaNce rankIng NETwork (MINI-Net) for video highlight detection. We consider each video as a bag, and each bag contains a set of segments of the video (*i.e.*, the video segments are treated as instances in each bag). We denote the event of interest as *interest event* and the other as *non-interest events*, and therefore a video contains the event of interest is called a *positive video* and a video that does not is called a *negative video*.

More specifically, we represent a positive video as a bag $\mathcal{B}_p = \{\mathcal{I}_p^i\}_{i=1}^N$, namely a positive bag. The positive bag contains N individual instances $\{\mathcal{I}_p^i\}_{i=1}^N$ (*i.e.*, segments of the positive video). Similarly, the negative bag $\mathcal{B}_n$ contains N different segments $\{\mathcal{I}_n^i\}_{i=1}^N$ from a negative video. Our model learns the highlights of interest event through positive bag; and through the learning of negative bag, the segments of the videos in non-interest events are treated as non-highlights for the specific event.

Given a pair of bags (*i.e.*, a positive bag $\mathcal{B}_p$ and a negative bag $\mathcal{B}_n$), we first pre-extract the vision features $\{\mathbf{f}_v^i\}_{i=1}^N$ and audio features $\{\mathbf{f}_a^i\}_{i=1}^N$ using pretrained models. We then feed the pre-extracted features of both the positive bag and negative bag into the proposed model to estimate the highlight scores of instances (*i.e.*, $\{\mathcal{E}_p^i\}_{i=1}^N, \{\mathcal{E}_n^i\}_{i=1}^N$) and event prediction (*i.e.*, interest event or non-interest event) of two bags (*i.e.*, $y_{\mathcal{B}_p}, y_{\mathcal{B}_n}$) as follows:

$$\begin{aligned} \{\mathbf{f}_p^i\}_{i=1}^N, \{\mathbf{f}_n^i\}_{i=1}^N &= f^F(\{\mathcal{I}_p^i\}_{i=1}^N, \{\mathcal{I}_n^i\}_{i=1}^N | \theta^F), \\ \{\mathcal{E}_p^i\}_{i=1}^N, \{\mathcal{E}_n^i\}_{i=1}^N &= f^E(\{\mathbf{f}_p^i\}_{i=1}^N, \{\mathbf{f}_n^i\}_{i=1}^N | \theta^E), \\ y_{\mathcal{B}_p}, y_{\mathcal{B}_n} &= f^C(\{\mathbf{f}_p^i\}_{i=1}^N, \{\mathbf{f}_n^i\}_{i=1}^N, \{\mathcal{E}_p^i\}_{i=1}^N, \{\mathcal{E}_n^i\}_{i=1}^N | \theta^C), \end{aligned} \tag{1}$$

where $f^F(\cdot)$ is the vision-audio fusion module parameterized by θ^F. The vision-audio fusion module takes each segment's vision feature and audio feature as input to encode the vision-audio fusion feature that contains both vision information and audio information (*i.e.*, $\{\mathbf{f}_p^i\}_{i=1}^N, \{\mathbf{f}_n^i\}_{i=1}^N$ are vision-audio fusion features for the positive bag and negative bag). The encoded fusion features are input into the highlight estimation module $f^E(\cdot)$ parameterized by θ^E to predict their highlight scores. The bag classification module $f^C(\cdot)$ takes as input the vision-audio fusion features of all segments and their immediate highlight score to estimate the event category of both the positive bag and the negative bag.

[6] We use the term event-specific to mean that there is event/category of interest specified by keyword(s) like "surfing", following [27,30].

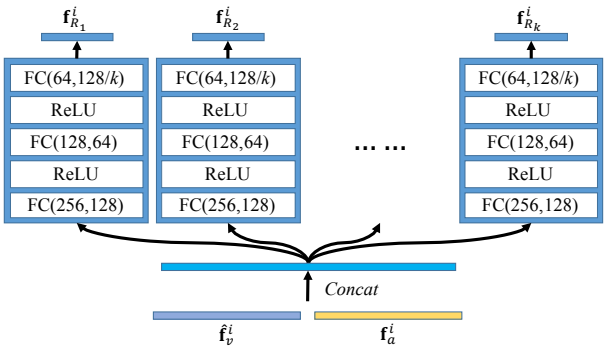

Fig. 3: Illustration of the vision-audio fusion submodule. The dimension of both vision feature $\hat{\mathbf{f}}^i_v$ and audio feature $\mathbf{f}^i_a$ are 128. k is the number of fusion submodule. The "FC" and "ReLU" represent fully connection and rectified linear unit activation, respectively.

To facilitating distinguishing positive bags from negative bags, we introduce two loss functions, *i.e.*, the max-max ranking loss and the binary bag event classification loss, to effectively train the whole multiple instance learning framework. The illustration shown in Figure 2 provides an overview of our proposed method.

3.1 Vision-audio Fusion Module $f^F(\cdot)$

Given a bag of segments, instead of using the visual information to estimate the highlight score individually, we consider using both visual and audio information as visual and audio events tend to occur together, and it has been shown that audio can be adopted to assist computer vision tasks[1,10,25]. For instance, a scene of people surfing is usually accompanied by the sound of waves. To this end, we design a vision-audio fusion module to encode visual-audio fusion representations for video highlight detection.

Given the pre-extracted vision feature $\mathbf{f}^i_v \in \mathbb{R}^{512}$ and audio feature $\mathbf{f}^i_a \in \mathbb{R}^{128}$ of a segment $\mathcal{I}_i$ in a bag (*i.e.*, positive bag or negative bag), as the dimensions of both features are not the same, we first employ two fully connected layers to transform $\mathbf{f}^i_v$ to a 128-dimensional vector, denoted as $\hat{\mathbf{f}}^i_v$. We then encode the vision-audio relation feature $\mathbf{f}^i_R$ and employ the residual connection to merge the vision-audio relation feature and vision feature $\hat{\mathbf{f}}^i_v$, yielding the vision-audio fusion feature $\mathbf{f}^i = \hat{\mathbf{f}}^i_v + \mathbf{f}^i_R$.

To encode $\mathbf{f}^i_R$, we concatenate the vision feature $\hat{\mathbf{f}}^i_v$ and audio feature $\mathbf{f}^i_a$ and feed the concatenated feature into k parallel fusion submodules to transform the concatenated feature to k relation features $\mathbf{f}^i_{R_k}$. We then concatenate the k relation feature to form the vision-audio relation feature $\mathbf{f}^i_R$.

We show the architecture of the submodules in Figure 3. Each fusion submodule contains 3 fully connected layers and two activation operators to transform a 256-dimensional concatenated feature into a $\frac{128}{k}$-dimensional relation feature. In

this way, the vision-audio fusion feature can be rewritten as follows:

$$\mathbf{f}^i = \hat{\mathbf{f}}_v^i + Concat[\mathbf{f}_{R_1}^i, \ldots, \mathbf{f}_{R_k}^i], (i = 1, \ldots, N). \tag{2}$$

In this way, the k parallel relation submodules allow the vision-audio fusion module to learn various types of relations between vision and audio. Additionally, encoding two sources of features (*i.e.*, vision and audio) enables the vision-audio fusion module to automatically activate the audio information if the audio is useful for the interest event and suppress the audio information if the audio is noisy or not helpful.

3.2 Highlight Estimation Module $f^E(\cdot)$

To predict the highlight score, we feed the vision-audio fused feature $\mathbf{f}^i$ into the highlight estimation module, where we transform $\mathbf{f}^i$ into a score value that will be used for bag classification and computing the proposed max-max ranking loss in later sections. More specifically, we first compute the initial highlight score by:

$$\hat{\mathcal{E}}^i = W_H(ReLU(W_S\mathbf{f}^i)), \tag{3}$$

where W_S is a matrix projecting the vision-audio fusion feature into a subspace, the ReLU activation operator activates the effective elements, and the matrix W_H is applied to measure the highlight score.

Rather than simply using $\hat{\mathcal{E}}^i$ as the highlight score, we consider estimating the final highlight score using the scores of all segments in a bag since the highlight score of one segment is related to other segments in the same video. Therefore, we formulate the final score as:

$$\mathcal{E}^i = \left(\sum_{t=1}^{N} exp(\hat{\mathcal{E}}^t)\right)^{-1} exp(\hat{\mathcal{E}}^i), \tag{4}$$

In this way, $\mathcal{E}^i$ is normalized in a bag and can be compared with the score of a segment in another bag.

3.3 Bag Classification Module $f^C(\cdot)$

Apart from estimating highlight scores of individual segments, we find that the event category can also be used as a supervision signal for training. The event category label can be more easily collected as all videos can be collected by specific query tags, and the tags can be used to generate the binary event label (*i.e.*, interest event or non-interest events). In addition, it is the fact that a video may contain highlights of various events while we are only interested in a specific event's highlights. This means that correctly classifying the event category (interest event or non-interest events) can be a useful inductive bias for event-specific highlight detection.

More specifically, we first label positive videos (videos of interest event) as 1 and negative videos (videos of non-interest events) as 0, *i.e.*, $Y_{\mathcal{B}_p} = 1$ for $\mathcal{B}_p$

and $Y_{\mathcal{B}_n} = 0$ for $\mathcal{B}_n$. To classify the event category of each bag, we aggregate the vision-audio fusion features of all instances weighted by their immediate estimated highlight scores to generate the bag representation:

$$\mathbf{f}_{\mathcal{B}} = \sum_{i=1}^{N} \mathcal{E}^i \mathbf{f}^i. \tag{5}$$

In this way, the generated bag representation could be highly informative for the event classification of each bag, as it mainly relies on the vision-audio fusion feature of the instance with high highlight scores.

We then feed the generated bag feature $\mathbf{f}_{\mathcal{B}}$ into an event classifier that consists of two fully connected layers. We apply the softmax function to estimate the event categories for both the positive bag $y_{\mathcal{B}_p}$ and the negative bag $y_{\mathcal{B}_n}$.

3.4 Objective Functions

After obtaining the predicted highlight scores of segments and the estimated event categories of the positive bag and negative bag, we introduce two objective functions (*i.e.*, max-max ranking loss and bag event classification loss) to effectively train our MINI-Net.

- "max-max" ranking loss (MM-RL). To learn the highlight detection model, we expect that the highlight score of a ground-truth highlight segment is higher than the score of a non-highlight segment:

$$\mathcal{E}_{gt-H} > \mathcal{E}_{gt-N}, \tag{6}$$

where $\mathcal{E}_{gt-H}$ is the highlight score of a ground-truth highlight segment and $\mathcal{E}_{gt-N}$ is the score of a non-highlight segment.

However, the highlight annotations are not available during training. Considering that the positive video contains at least one highlight segment, and the negative video does not have any highlights of the interest event, we thus believe the segment from a positive video with the highest score is the most likely to be a highlight, and the segment from the negative bag with the highest score can be assigned as a hardest non-highlight. We adapt Eq. 6 as follows for acquiring a reliable relative comparison between mostly likely positive instance and hardest negative instance:

$$\max_{\mathcal{I}_p^i \in \mathcal{B}_p} \mathcal{E}_p^i > \max_{\mathcal{I}_n^i \in \mathcal{B}_n} \mathcal{E}_n^i, \tag{7}$$

where *max* operators pick the maximum value from the highlight scores of all segments in a bag. Here, the highlights in the non-interest events' videos are viewed as non-highlights for the interest event. Using the segments of the non-interest events as a negative instance is more reliable than using the segment from the long-form interest event's video [27].

To instantiate Eq. 7, we introduce the max-max ranking loss (MM-RL) as:

$$\mathcal{L}_{MM}(\mathcal{B}_p, \mathcal{B}_n) = \max(0, \epsilon - \max_{\mathcal{I}_p^i \in \mathcal{B}_p} \mathcal{E}_p^i + \max_{\mathcal{I}_n^i \in \mathcal{B}_n} \mathcal{E}_n^i), \tag{8}$$

where $\mathcal{L}_{AH}$ is applied to ensure that $\max_{\mathcal{I}_p^i \in \mathcal{B}_p} \mathcal{E}_p^i$ is larger than $\max_{\mathcal{I}_n^i \in \mathcal{B}_n} \mathcal{E}_n^i$ with a margin of ϵ. ϵ is a hyperparameter and is equal to 1 in this work.

- Bag Event Classification Loss. As mentioned in Sec. 3.3, in addition to the MM-RL loss, we expect the bag event classification loss can enforce the model to produce more informative bag representation for the specific event. To this end, we apply the binary cross entropy loss function to the estimated event categories of both positive bag and negative bag for bag event classification. Finally, we add up both the MM-RL and the bag event classification loss to form the final loss:

$$\mathcal{L} = \mathcal{L}_{MM}(\mathcal{B}_p, \mathcal{B}_n) + \mathcal{L}_{CE}(y_{\mathcal{B}_p}, Y_{\mathcal{B}_p}) + \mathcal{L}_{CE}(y_{\mathcal{B}_n}, Y_{\mathcal{B}_n}), \tag{9}$$

where $\mathcal{L}_{CE}(\cdot)$ is the binary cross entropy loss function.

4 Experiments

In this section, we conduct extensive experiments on three public datasets to investigate the effectiveness of the proposed model. More experimental results and details are reported and analyzed in the Supplementary Material.

4.1 Datasets and Metrics

We evaluate our method on three public benchmarks datasets, *i.e.*, YouTube Highlights [21], TVSum [19] and CoSum [4], for video highlight detection.

- YouTube Highlights contains six evnet-specific categories, *i.e.*, dogs, gymnastics, parkour, skating, skiing and surfing, and there are approximately 100 videos in each event. The given label for YouTube highlights indicates whether a segment is a ground-truth highlight segment.

- TVSum is an available video summarization benchmark dataset that is collected from *YouTube* and crawled by an event-specific queried tag. The dataset consists of 50 videos grouped by 10 categories (5 videos per category). We follow [2,27] and select the top 50% shots in terms of the score provided by annotators for each video as a human-created summary.

- CoSum has 51 videos covering 10 events. We follow [16,2] and compare each generated highlights with three human-created summaries.

4.2 Compared Methods

To further demonstrate the capacity of our method, we compare our method with numerous different methods on three datasets for video highlight detection.

- Weakly supervised methods. The compared methods include RRAE [30], MBF [4], SMRS [6], Quasi [13], CVS [17], SG [14], and LIM-s [27], and two weakly supervised methods, VESD [2] and DSN [16]. Although most of these methods are used for video summarization, their performance is evaluated using the same metrics as the metrics used in this study.

- Supervised methods Additionally, there are several supervised methods (*i.e.*, GIFs [9],LSVM [21], KVS [18], DPP [7], sLstm [32] and SM [8]) that are applied in video highlight detection and video summarization. We compare these methods using the same matrices mentioned above.

Topic	Supervised Methods		Weakly supervised Methods		Weakly supervised	
	GIFs	LSVM	RRAE	LIM-s	MINI-Net$^{\text{w/o audio}}$	MINI-Net
dog	0.308	**0.60**	0.49	0.579	0.5368	0.5816
gymnastics	0.335	0.41	0.35	0.417	0.5281	**0.6165**
parkour	0.540	0.61	0.50	0.670	0.6888	**0.7020**
skating	0.554	0.62	0.25	0.578	0.7094	**0.7217**
skiing	0.328	0.36	0.22	0.486	0.5834	**0.5866**
surfing	0.541	0.61	0.49	0.651	0.6383	**0.6514**
Average	0.464	0.536	0.383	0.564	0.6138	**0.6436**

Table 1: Experimental results (mAP) on the YouTube Highlights dataset. Our method outperforms all of the compared methods, including the state-of-the-art weakly supervised ranking-based method [27].

Topic	Supervised Methods					Weakly supervised/Un Methods							Weakly supervised	
	KVS	DPP	sLstm	SM	SMRS	Quasi	MBF	CVS	SG	LIM-s	DSN	VESD	MINI-Net$^{\text{w/o audio}}$	MINI-Net
VT	0.353	0.399	0.411	0.415	0.272	0.336	0.295	0.328	0.423	0.559	0.373	0.447	0.8028	**0.8062**
VU	0.441	0.453	0.462	0.467	0.324	0.369	0.357	0.413	0.472	0.429	0.441	0.493	0.6527	**0.6832**
GA	0.402	0.457	0.463	0.469	0.331	0.342	0.325	0.379	0.475	0.612	0.428	0.496	0.7535	**0.7821**
MS	0.417	0.462	0.477	0.478	0.362	0.375	0.412	0.398	0.489	0.540	0.436	0.503	0.8128	**0.8183**
PK	0.382	0.437	0.448	0.445	0.289	0.324	0.318	0.354	0.456	0.604	0.411	0.478	0.7801	**0.7807**
PR	0.403	0.446	0.461	0.458	0.276	0.301	0.334	0.381	0.473	0.475	0.417	0.485	0.5446	**0.6584**
FM	0.397	0.442	0.452	0.451	0.302	0.318	0.365	0.365	0.464	0.432	0.412	0.487	0.5586	**0.5780**
BK	0.342	0.395	0.406	0.407	0.297	0.295	0.313	0.326	0.417	0.663	0.368	0.441	0.7174	**0.7502**
BT	0.419	0.464	0.471	0.473	0.314	0.327	0.365	0.402	0.483	0.691	0.435	0.492	0.7686	**0.8019**
DS	0.394	0.449	0.455	0.453	0.295	0.309	0.357	0.378	0.466	0.626	0.416	0.488	0.5911	**0.6551**
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	0.6979	**0.7324**

Table 2: Experimental results (top-5 mAP score) on the TVsum dataset. Our method outperforms all of the compared methods by a large margin.

4.3 Highlight Detection Results

- Result for the YouTube Highlights dataset: We report our results in comparison with other researches [7]. For the sake of fairness, we also reported the results of a MINI-Net's variant, *i.e.*, MINI-Net$^{\text{w/o audio}}$, which removes the audio feature from the MINI-Net and replace the vision-audio fusion feature with vision feature (more analysis about MINI-Net$^{\text{w/o audio}}$ is reported in Section 4.4). We find that our method achieves the best result in terms of the average mAP over all events. Compared to the ranking-based weakly supervised method LIM-s and auto-encoder-based weakly supervised method RRAE, MINI-Net's average gains in mAP are 7.96% and 26.06%, respectively. The result strongly verifies that our weakly supervised method based on multiple instance learning has better capacity than the compared methods. It is noteworthy that the our result is even better than that achieved by supervised methods, *i.e.*, GIFs and LSVM, which are trained with event-specific manually annotated data. These results indicate that our MINI-Net can leverage unlabeled videos for video highlight detection more effectively than other methods without the need to spend a lot of manual labor on data annotation. We also find that our MINI-Net$^{\text{w/o audio}}$ outperforms all compared methods without audio feature. Such results indicate that proposed objective functions can improve the ability to distinguish of our model.

[7] The compared results are from original papers.

Topic	Supervised Methods					Weakly supervised							Weakly supervised	
	KVS	DPP	sLstm	SM	SMRS	Quasi	MBF	CVS	SG	LIM-s	VESD	DSN	MINI-Net$^{w/o\ audio}$	MINI-Net
BJ	0.662	0.672	0.683	0.692	0.504	0.561	0.631	0.658	0.698	–	0.685	0.715	0.7756	**0.8450**
BP	0.674	0.682	0.701	0.722	0.492	0.625	0.592	0.675	0.713	–	0.714	0.746	0.9628	**0.9887**
ET	0.731	0.744	0.749	0.789	0.556	0.575	0.618	0.722	0.759	–	0.783	0.813	0.7864	**0.9156**
ERC	0.685	0.694	0.717	0.728	0.525	0.563	0.575	0.693	0.729	–	0.721	0.756	0.9525	**1.0000**
KP	0.701	0.705	0.714	0.745	0.521	0.557	0.594	0.707	0.729	–	0.742	0.772	0.9585	**0.9611**
MLB	0.668	0.677	0.714	0.693	0.543	0.563	0.624	0.679	0.721	–	0.687	0.727	0.8686	**0.9353**
NFL	0.671	0.681	0.681	0.727	0.558	0.587	0.603	0.674	0.693	–	0.724	0.737	0.8972	**1.0000**
NDC	0.698	0.704	0.722	0.759	0.496	0.617	0.594	0.702	0.738	–	0.751	0.782	0.8901	**0.9536**
SL	0.713	0.722	0.721	0.766	0.525	0.551	0.624	0.715	0.743	–	0.763	0.794	0.7865	**0.8896**
SF	0.642	0.648	0.653	0.683	0.533	0.562	0.603	0.647	0.681	–	0.674	0.709	0.7272	**0.7897**
Average	0.684	0.692	0.705	0.735	0.525	0.576	0.602	0.687	0.720	–	0.721	0.755	0.8605	**0.9278**

Table 3: Experimental results (top-5 mAP score) on the CoSum dataset. Our method outperforms all of the compared methods by a large margin. The entries with "-" mean per-class results are not available for that method.

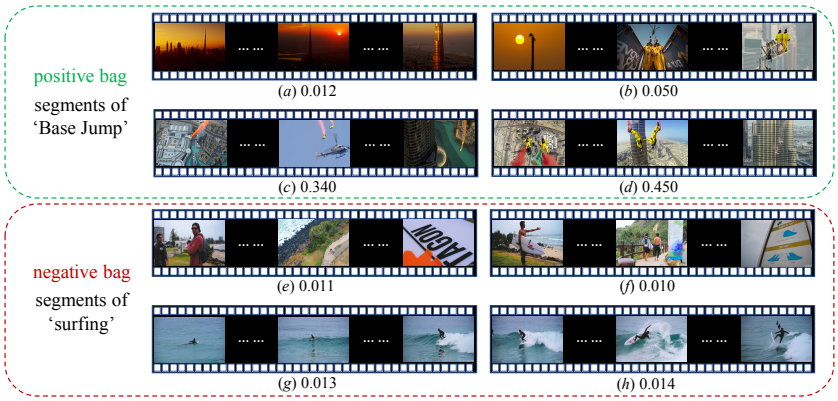

Fig. 4: The example of bag in our approach, and the highlight scores of each instance estimated by our MINI-Net trained for detecting "base jump" highlight.

- Result on TVSum dataset and CoSum dataset: The experimental results for our method on the TVSum dataset and the CoSum dataset are shown in Table 2 and Table 3, respectively. TVsum and CoSum are more challenging datasets as they have diverse videos. However, our method outperforms all of the baselines by a large margin on both the TVSum dataset and the CoSum dataset. Note that LIM-s [27], which is the most competitive ranking-based weakly supervised method, provides the average top-5 mAP, which is 16.94% less than the value achieved with our MINI-Net on the TVSum dataset. Our approach achieves a significant and consistent improvement over all the events in the two datasets. (*e.g.*, the top-5 mAP of our MINI-Net vs. that of VESD are 84.50% vs. 68.5% on the BJ event of CoSum dataset). These results show that the training model based on multiple instance learning using both interest events video data and non-interest events video data is more useful for video highlight detection. As these two datasets consist of long-form videos crawled from social media platforms, in addition to the highlights of the interest event,

Dataset	Gated-Attention [11]	DMIL-RM [20]	MINI-Net
YouTube	0.6289	0.6357	**0.6436**
TVSum	0.6533	0.6895	**0.7324**
Cosum	0.7516	0.7943	**0.9278**

Table 4: Comparisions with related multiple instance learning methods.

these videos inevitably contain video information of other events. Figure 4 shows segments and their highlight scores. We can determine that the segments in the non-interest event (*i.e.*, negative bag) are assigned low highlight scores (the segments (e)-(h) in Figure 4) and the highlights of the interest event achieve the highest scores (the segment (d) in Figure 4). The performances on the TVsum and CoSum datasets indicate that our model has the capacity to treat segments from non-interest events as non-highlights and only detect highlights from the interest event.

- Comparison with other multiple instance learning methods. To further prove that our proposed multiple instance learning framework is suitable for video highlight detection, we compare the other two multiple instance learning frameworks, *i.e.*, Gated-Attention [11] and DMIL-AM [20], which are adapted to video highlight detection. It is clearly shown in Table 4 that our method performs the best. *e.g.*, MINI-Net outperforms Gate-Attention and DMIL-RM by 17.62% and 13.35% on CoSum dataset, respectively. The results in Table 4 demonstrate that the architecture of MINI-Net is more suitable for video highlight detection.

4.4 Ablation Studies

We present an ablation study to evaluate each component of our model.

- Effect of bag modeling. Firstly, we evaluate the effect of bag classification module on the proposed model by removing the module, *i.e.*, MINI-Net$^{\text{w/o BCM}}$. Comparing the full model and our model without bag classification module, we clearly observe that the bag classification improves the performance (*e.g.*, "MINI-Net" improves the performance of "MINI-Net$^{\text{w/o BCM}}$" from 65.58% to 73.24% for TVSum dataset). This implies that our bag classification module is able to help select as many ground-truth highlights from the video as possible, which benefits video highlight detection.

- Effect of max-max ranking loss (MM-RL). Secondly, we evaluate the impact of MM-RL on our approach. MINI-Net$^{\text{w/o MM-RL}}$ indicates that we have removed the MM-RL from the Eq. 9. From Table 5, we also observe that adding max-max ranking loss can consistently boost the performance (*e.g.*, the results of "MINI-Net" vs. those of "MINI-Net$^{\text{w/o MM-RL}}$" are 92.78% vs. 77.59% for the CoSum dataset). This result indicates that forcing the most likely highlight segment and the hard non-highlight segment to be far apart in terms of highlight score can help the potential ground-truth highlight segment of the interest event obtain a relatively high score .

- Effect of audio features. Finally, to verify that audio is beneficial in our work, we conduct an experiment that trains our model without audio features,

Dataset	MINI-Net$^{\text{w/o vision}}$	MINI-Net$^{\text{w/o audio}}$	MINI-Net$^{\text{w/o MM-RL}}$	MINI-Net$^{\text{w/o BCM}}$	MINI-Net
YouTube	0.5223	0.6138	0.6166	0.6113	**0.6436**
TVSum	0.5972	0.6979	0.6495	0.6558	**0.7324**
Cosum	0.6914	0.8605	0.7759	0.7823	**0.9278**

Table 5: Ablation study on three datasets.

i.e., MINI-Net$^{\text{w/o audio}}$ in Table. 5, and MINI-Net$^{\text{w/o vision}}$ indicates that we have removed the vision feature. More specifically, we use the audio or vision features after several layers of fully connected layers (we make the number of parameters consistent) to replace the fused features that are input to the subsequent network. In Table 5, we can find that our full method outperforms the alternative variants. In particular, comparing MINI-Net$^{\text{w/o audio}}$ and MINI-Net$^{\text{w/o vision}}$ for the three datasets, the MINI-Net$^{\text{w/o vision}}$ outperforms MINI-Net$^{\text{w/o audio}}$ by 9.15%, 10.07% and 16.91% for YouTube Highlights dataset, TVSum dataset and CoSum dataset, respectively. These results indicate that: 1) Even using only vision features, our method outperforms the compared methods in Table 1, Table 2 and Table 3. 2) Using audio alone can degrade the performance more than using video alone, as audio is sometimes not native, and music or a voiceover is applied by the video owner. Such audio cannot be utilized to improve the performance and introduce noise; 3) It is also verified that the combination of audio and vision can improve the performance of the model.

5 Conclusion

Compared to related work, to our best knowledge, this work is the first to cast the weakly supervised video highlight detection problem as a multiple instance ranking approach. The bag modeling in our multiple instance ranking network (MINI-Net) particularly solves the difficulty of localization of highlight segments of a specific event during training, because MINI-Net works on bag level, where it is only required to ensure a positive bag having a highlight segment of that event and a negative bag having relevant ones. Based on such bag setting, with a max-max ranking loss, our MINI-Net is able to effectively leverage and quantify all segment information of a video, and therefore the proposed MINI-Net manages to acquire reliable higher highlight scores for positive bags as compared to negative bags. The experimental results have validated the effectiveness of our approach.

6 Acknowledgements

This work was supported partially by the National Key Research and Development Program of China (2018YFB1004903), NSFC(U1911401,U1811461), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157), Guangdong NSF Project (No. 2018B030312002), Guangzhou Research Project (201902010037), and Research Projects of Zhejiang Lab (No. 2019KD0AB03).

References

1. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: International Conference on Computer Vision (2017)
2. Cai, S., Zuo, W., Davis, L.S., Zhang, L.: Weakly-supervised video summarization using variational encoder-decoder and web prior. In: European Conference on Computer Vision (2018)
3. Carbonneau, M.A., Cheplygina, V., Granger, E., Gagnon, G.: Multiple instance learning: A survey of problem characteristics and applications. Pattern Recognition **77**, 329–353 (2018)
4. Chu, W.S., Song, Y., Jaimes, A.: Video co-summarization: Video summarization by visual co-occurrence. In: Computer Vision and Pattern Recognition (2015)
5. Cinbis, R.G., Verbeek, J., Schmid, C.: Weakly supervised object localization with multi-fold multiple instance learning. Transactions on Pattern Analysis and Machine Intelligence **39**(1), 189–203 (2016)
6. Elhamifar, E., Sapiro, G., Vidal, R.: See all by looking at a few: Sparse modeling for finding representative objects. In: Computer Vision and Pattern Recognition (2012)
7. Gong, B., Chao, W.L., Grauman, K., Sha, F.: Diverse sequential subset selection for supervised video summarization. In: Advances in Neural Information Processing Systems (2014)
8. Gygli, M., Grabner, H., Van Gool, L.: Video summarization by learning submodular mixtures of objectives. In: Computer Vision and Pattern Recognition (2015)
9. Gygli, M., Song, Y., Cao, L.: Video2gif: Automatic generation of animated gifs from video. In: Computer Vision and Pattern Recognition (2016)
10. Hori, C., Hori, T., Lee, T.Y., Zhang, Z., Harsham, B., Hershey, J.R., Marks, T.K., Sumi, K.: Attention-based multimodal fusion for video description. In: International Conference on Computer Vision (2017)
11. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. arXiv preprint arXiv:1802.04712 (2018)
12. Jiao, Y., Li, Z., Huang, S., Yang, X., Liu, B., Zhang, T.: Three-dimensional attention-based deep ranking model for video highlight detection. Transactions on Multimedia **20**(10), 2693–2705 (2018)
13. Kim, G., Sigal, L., Xing, E.P.: Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In: Computer Vision and Pattern Recognition (2014)
14. Mahasseni, B., Lam, M., Todorovic, S.: Unsupervised video summarization with adversarial lstm networks. In: Computer Vision and Pattern Recognition (2017)
15. Meng, J., Wu, S., Zheng, W.S.: Weakly supervised person re-identification. In: Computer Vision and Pattern Recognition (2019)
16. Panda, R., Das, A., Wu, Z., Ernst, J., Roy-Chowdhury, A.K.: Weakly supervised summarization of web videos. In: International Conference on Computer Vision (2017)
17. Panda, R., Roy-Chowdhury, A.K.: Collaborative summarization of topic-related videos. In: Computer Vision and Pattern Recognition (2017)
18. Potapov, D., Douze, M., Harchaoui, Z., Schmid, C.: Category-specific video summarization. In: European Conference on Computer Vision (2014)
19. Song, Y., Vallmitjana, J., Stent, A., Jaimes, A.: Tvsum: Summarizing web videos using titles. In: Computer Vision and Pattern Recognition (2015)

20. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Computer Vision and Pattern Recognition (2018)
21. Sun, M., Farhadi, A., Seitz, S.: Ranking domain-specific highlights by analyzing edited videos. In: European Conference on Computer Vision (2014)
22. Tang, H., Kwatra, V., Sargin, M.E., Gargi, U.: Detecting highlights in sports videos: Cricket as a test case. In: International Conference on Multimedia and Expo (2011)
23. Ulges, A., Schulze, C., Breuel, T.: Multiple instance learning from weakly labeled videos. In: Workshop on Cross-media Information Analysis and Retrieval (2008)
24. Wang, J., Xu, C., Chng, E., Tian, Q.: Sports highlight detection from keyword sequences using hmm. In: International Conference on Multimedia and Expo (2004)
25. Wang, L., Sun, Z., Yao, W., Zhan, H., Zhu, C.: Unsupervised multi-stream highlight detection for the game "honor of kings". arXiv preprint arXiv:1910.06189 (2019)
26. Wu, J., Yu, Y., Huang, C., Yu, K.: Deep multiple instance learning for image classification and auto-annotation. In: Computer Vision and Pattern Recognition (2015)
27. Xiong, B., Kalantidis, Y., Ghadiyaram, D., Grauman, K.: Less is more: Learning highlight detection from video duration. In: Computer Vision and Pattern Recognition (2019)
28. Xiong, B., Kim, G., Sigal, L.: Storyline representation of egocentric videos with an applications to story-based search. In: International Conference on Computer Vision (2015)
29. Xiong, Z., Radhakrishnan, R., Divakaran, A., Huang, T.S.: Highlights extraction from sports video based on an audio-visual marker detection framework. In: International Conference on Multimedia and Expo (2005)
30. Yang, H., Wang, B., Lin, S., Wipf, D., Guo, M., Guo, B.: Unsupervised extraction of video highlights via robust recurrent auto-encoders. In: International Conference on Computer Vision (2015)
31. Yao, T., Mei, T., Rui, Y.: Highlight detection with pairwise deep ranking for first-person video summarization. In: Computer Vision and Pattern Recognition (2016)
32. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Video summarization with long short-term memory. In: European Conference on Computer Vision (2016)
33. Zhou, K., Qiao, Y., Xiang, T.: Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In: AAAI Conference on Artificial Intelligence (2018)