This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Joint Visual and Audio Learning for Video Highlight Detection

Taivanbat Badamdorj[1], Mrigank Rochan[2], Yang Wang[2,3], and Li Cheng[1]

[1]Department of Electrical and Computer Engineering, University of Alberta, Canada
[2]Huawei Noah's Ark Lab, Canada
[3]Department of Computer Science, University of Manitoba, Canada

Abstract

In video highlight detection, the goal is to identify the interesting moments within an unedited video. Although the audio component of the video provides important cues for highlight detection, the majority of existing efforts focus almost exclusively on the visual component. In this paper, we argue that both audio and visual components of a video should be modeled jointly to retrieve its best moments. To this end, we propose an audio-visual network for video highlight detection. At the core of our approach lies a bimodal attention mechanism, which captures the interaction between the audio and visual components of a video, and produces fused representations to facilitate highlight detection. Furthermore, we introduce a noise sentinel technique to adaptively discount a noisy visual or audio modality. Empirical evaluations on two benchmark datasets demonstrate the superior performance of our approach over the state-of-the-art methods.

1. Introduction

We have witnessed an explosion of online video content in recent years, which may be partly attributed to the rapid adoption of video based social networks like Instagram and TikTok. This has led to increasing demands for video highlight detection, which aims to automatically detect interesting moments (called "highlights") within a video. Highlight detection is important due to its broad range of downstream applications including video summarization, recommendation, editing, and browsing. In consequence, there has been significant progress in the field in recent years [37, 12, 46, 50, 17, 49, 7, 44, 31, 15, 42].

However, the majority of existing research efforts focus on visual highlight detection. Audio-visual highlight detection is largely unexplored territory. In this paper, we posit that interesting moments can be identified from both visual and audio information. For example, Fig. 1, shows a video of a car crashing during a race. We may identify the interesting part of the video by seeing the crash. On the other hand, hearing people stop cheering and start talking could also be an indicator that the moment is interesting. While we can process the visual and audio information separately, our insight is that they also interact with each other. We can imagine that if we were in the crowd depicted in Fig. 1, we would look to see what had happened if everybody suddenly stopped cheering. It can also work the other way around: what we hear can reinforce our beliefs that the given moment is interesting. These observations moti-

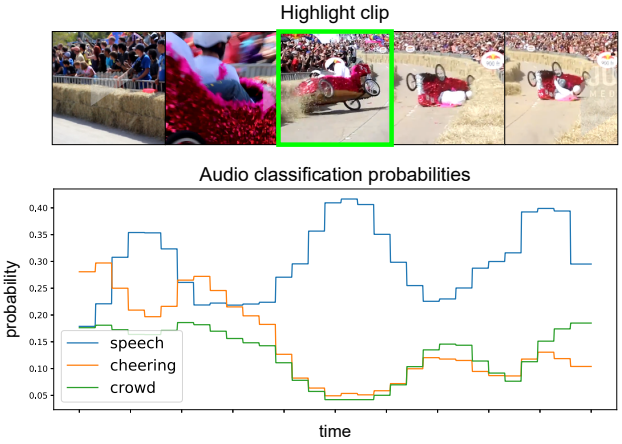

Figure 1. Audio-visual highlight detection: audio can be informative about which part of the video is a highlight. In this video, a car crashes during a race in front of a large crowd. The highlight (in green) is the crash, and we also show the top three audio class probabilities from a pre-trained audio classification network [22]. The class probabilities (speech, cheering, and crowd) show that the cheering and crowd noise dies out during the crash, while talking increases. Intuitively, we understand that if people stop cheering suddenly, something must have happened. In this work, we learn to utilize such audio cues.

vate us to propose an approach that jointly learns from visual and audio information to detect highlights in videos. It consists of two different attention mechanisms: a unimodal self-attention mechanism that models the relationships between moments belonging to the same modality; and a bimodal attention mechanism that models the interaction between the two modalities.

Furthermore, we impart the ability to ignore a modality on our model. Intuitively, if we hear something interesting, but do not see anything interesting when we look, we would like to be able to ignore what we hear. In addition, if we do not hear anything of interest, it is also not worth looking. Inspired by the visual sentinel [24], we introduce a noise sentinel in the bimodal attention mechanism that allows our model to "look-away" from a modality by attending to the noise sentinel instead.

The proposed attention mechanisms and the noise sentinel are novel and effective as empirically indicated in the ablation studies. We demonstrate our model's superior performance on three well-known benchmarks, where our approach significantly outperforms the state-of-the-art methods.

2. Related Work

Video Highlight Detection: Early efforts in video highlight detection mainly deal with sports videos [41, 45, 38]. Later works were proposed to deal with videos from social media [37], as well as first-person videos [47].

Some works formulate highlight detection as a classification problem [31], while it is also popular to formulate highlight detection as a ranking problem [37, 7, 17, 12, 49, 42], where a ranking network is trained to rank highlight clips higher than non-highlight clips.

Recently, weakly-supervised highlight detection methods have been proposed [44, 15], where the training label is only available at the entire video level. The work of [44] takes advantage of the fact that clips from shorter videos are more likely to be highlights to train a ranking network. MN [15], on the other hand, proposes a multiple instance ranking framework that learns to rank clips from a given category higher than clips from other categories.

MN [15] is the only other work that proposes an audio-visual framework. Their network operates independently for each clip, and uses a variant of simple concatenation to produce a fused audio-visual feature. In contrast, our model can utilize the context within the entire video through the self-attention and bimodal attention layers. In addition, while their work assumes that audio is useful only as a complementary feature to the visual features, we make no such assumption, and allow each modality to modulate the other.

Video Summarization is a closely related task aimed at producing a compact and cohesive summary of a given video. Early works in video summarization are predominantly unsupervised [19, 20, 23, 25, 26, 27, 29, 34, 33, 57, 53], and as such, many rely on heuristics such as diversity and representativeness to obtain a summary video.

Weakly supervised methods [30, 19, 20, 34, 28, 3, 32] have also been developed to utilize video-level information. Benefiting from the massive user tagging of online videos, research efforts in supervised learning [9, 10, 11, 33, 51, 52, 56, 5] are also progressing rapidly.

Our work is influenced by the attention-based model of [5] that operates on sequences of clips as potential highlights. Unlike their work, our work is multi-modal, and uses a different formulation of the attention.

Multi-modal Learning: Recent progress has been made in multi-modal learning in various fields such as salient object detection [54, 55, 16], action recognition [6], and speech recognition [1, 36]. It has been observed [43] that simple fusion strategies often fail. We also observe in our experiments that late and early fusion strategies are suboptimal, and propose the use of a bimodal attention layer which captures the interaction between modalities.

Our work is also related to audio-visual speech recognition [1, 36]. The work of [1] uses an attention network but does not model the interplay between audio and visual channels, instead the output of the previous decoding layer is used as the query for both modalities. The work of [36] considers a one-way relation where the audio can be used to query the visual features, but not vice-versa. In contrast, our model is symmetric, and allows the visual features to query the audio features.

Visual Sentinel is a concept introduced in image caption generation by [24]. The key idea is that caption generation models do not always need to attend to the image: some words like "and","of", and "from" have no visual grounding within images. Therefore, the visual sentinel [24] is proposed as a "look-away" mechanism. When generating words without visual grounding, their model learns to attend to the visual sentinel instead of the image.

This inspires us to formulate a *noise sentinel* in our context to automatically "look away" from a modality if it is noisy. As we shall see in the experiments, this greatly improves the empirical performance of our approach.

3. Our Approach

We show the overall architecture of our framework in Fig. 2. We first split a video V into clips of a fixed length (*e.g.* 100 frames), resulting in N clips. We characterize the i-th clip by two vectors, $\boldsymbol{v}_i$ for the visual features, and $\boldsymbol{a}_i$ for the audio features, where $i \in 1, 2, ..., N$. These visual and audio features are extracted using pre-trained visual [13, 39] and audio feature extractors [22]. Therefore, the input video is sufficiently represented by two sets of feature vectors, the visual $\{\boldsymbol{v}_i\}_{i=1}^{N}$ and the audio features $\{\boldsymbol{a}_i\}_{i=1}^{N}$. Moreover, as explained in Fig. 2, the interactions among clips as well

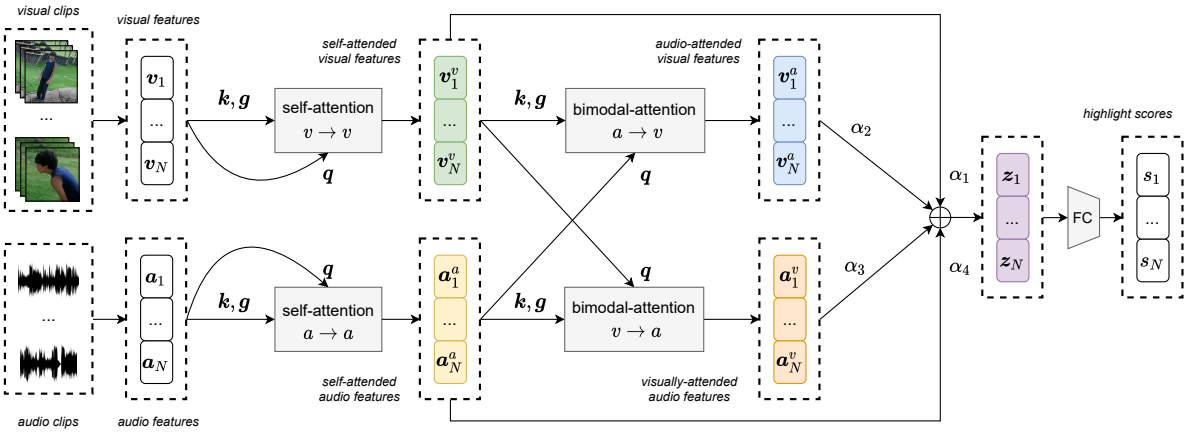

Figure 2. An overview of our framework. An input video is evenly split into clips of a fixed length, producing N total clips. The visual and audio modalities of the i-th clip are then represented as vectors, $\boldsymbol{v}_i$ and $\boldsymbol{a}_i$, respectively. For each modality, the relationship between clips are tackled separately in the self-attention layer; this leads to the self-attended features $\boldsymbol{v}_i^v$ for visual, and $\boldsymbol{a}_i^a$ for audio. It is then fed into the bimodal attention layers to model the cross-modality relationships, and produce the bimodally attended features $\boldsymbol{v}_i^a$ (audio attended visual features) and $\boldsymbol{a}_i^v$ (visually attended audio features). A more detailed inspection of the bimodal attention layer is shown in Fig. 3. Finally, the self-attended and bimodally attended features come together with a learned weighted sum, to produce the feature $\boldsymbol{z}_i$, which is processed to output the final clip highlight score s_i for the i-th clip.

as between the visual and audio modalities are captured by two different attention mechanisms, namely self-attention and bimodal attention. For each modality, the self-attention layer captures interactions among clips of the same modality. The output of both modalities are then fed to the two input branches of the bimodal attention layers which capture the interplay between the two modalities. These attended features are subsequently fused and fed to a classifier which produces a score s_i for the i-th clip, indicating whether this clip is a highlight or not.

Unimodal Self-Attention. A given clip by itself is often inadequate to determine whether it is a highlight; rather it is beneficial to consider the relationship of this clip with other clips in the video. The self-attention mechanism [40] and its variants have been shown to be effective in modeling such dependencies. Our self-attention mechanism is an adaptation of the disentangled attention block [48] to our context. Without loss of generality, we describe in what follows our self-attention mechanism for the visual features, $\{\boldsymbol{v}_i\}_{i=1}^N$.

Let us denote the visual self-attention relationship with $v \rightarrow v$ (we attend from visual to visual features). We denote by $W_{q,v\rightarrow v}$, $W_{k,v\rightarrow v}$, and $W_{m,v\rightarrow v}$ linear projection matrices, and define the following quantities:

$$\boldsymbol{q}_{v\rightarrow v}(\boldsymbol{v_i}) = \hat{\boldsymbol{q}}_{v\rightarrow v}(\boldsymbol{v_i}) - \frac{1}{N}\sum_{l=1}^{N}\hat{\boldsymbol{q}}_{v\rightarrow v}(\boldsymbol{v_l}), \quad (1)$$

$$\boldsymbol{k}_{v\rightarrow v}(\boldsymbol{v_j}) = \hat{\boldsymbol{k}}_{v\rightarrow v}(\boldsymbol{v_j}) - \frac{1}{N}\sum_{l=1}^{N}\hat{\boldsymbol{k}}_{v\rightarrow v}(\boldsymbol{v_l}), \quad (2)$$

Here $\boldsymbol{q}_{v\rightarrow v}(\boldsymbol{v_i})$ and $\boldsymbol{k}_{v\rightarrow v}(\boldsymbol{v_j})$ are called the query and the key in self-attention. $\hat{\boldsymbol{q}}_{v\rightarrow v}(\boldsymbol{v_i})$ and $\hat{\boldsymbol{k}}_{v\rightarrow v}(\boldsymbol{v_j})$ represent linear projections. We subtract the mean of these linear projections from each quantity in Eq. (1) and Eq. (2).

$$\hat{\boldsymbol{q}}_{v\rightarrow v}(\boldsymbol{v_i}) = W_{q,v\rightarrow v}\boldsymbol{v}_i, \quad (3)$$

$$\hat{\boldsymbol{k}}_{v\rightarrow v}(\boldsymbol{v_j}) = W_{k,v\rightarrow v}\boldsymbol{v}_j. \quad (4)$$

We also define the term m_j, which is a linear projection of each key into a scalar real. m_j is called the unary term.

$$m_j = W_{m,v\rightarrow v}\boldsymbol{v}_j. \quad (5)$$

Now, the visual self-attention score $\omega_{v\rightarrow v}(\boldsymbol{v}_i, \boldsymbol{v}_j)$ between two clips $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ is defined by

$$\begin{aligned}\omega_{v\rightarrow v}(\boldsymbol{v}_i, \boldsymbol{v}_j) =& \text{softmax}(c\,\boldsymbol{q}_{v\rightarrow v}(\boldsymbol{v}_i)^\intercal \boldsymbol{k}_{v\rightarrow v}(\boldsymbol{v}_j)) \\ &+ \text{softmax}(m_j), \end{aligned} \quad (6)$$

where the softmax is over the key indices j (over the attended sequence). Following [40], the constant c is used to normalize the dot product and to improve gradient flow through the softmax operator.

Intuitively, the attention score $\omega_{v\rightarrow v}(\boldsymbol{v}_i, \boldsymbol{v}_j)$ captures the amount of dependency of $\boldsymbol{v}_i$ on $\boldsymbol{v}_j$ relative to other clips in the video. As suggested by [48], the pairwise term $\boldsymbol{q}_{v\rightarrow v}(\boldsymbol{v}_i)^\intercal \boldsymbol{k}_{v\rightarrow v}(\boldsymbol{v}_j)$ captures the relationship between clips, while the unary term m_j represents the saliency of every clip within the video. This allows our module to

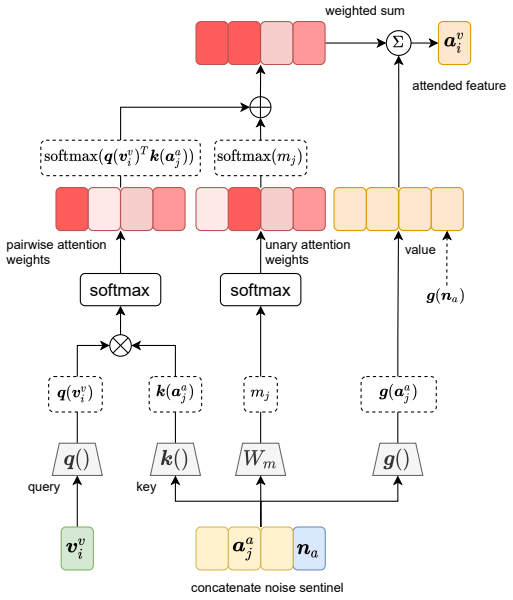

Figure 3. Detailed illustration of the bimodal attention layer. It aims to capture the relationship between the audio and visual modalities. The figure showcases attending to audio using visual features ($v \rightarrow a$). The noise sentinel is a part of the bimodal attention layer, formed by concatenating along the time axis to the modality attended to. The bimodal attention is then computed according to Eq.(19).

learn a representation that takes into account the relationship between clips as well as the overall importance of each clip in the entire video.

We define the linear projection matrix, $W_{g,v\rightarrow v}$. The attended features are then produced by applying attention scores to re-weight the input features,

$$g_{v\rightarrow v}(\boldsymbol{v}_j) = W_{g,v\rightarrow v}\boldsymbol{v}_j, \tag{7}$$

$$\boldsymbol{v}_i^v = \sum_{j=1}^{N} \omega_{v\rightarrow v}(\boldsymbol{v}_i, \boldsymbol{v}_j) g_{v\rightarrow v}(\boldsymbol{v}_j) + \boldsymbol{v}_i, \tag{8}$$

where $g_{v\rightarrow v}(\boldsymbol{v}_j)$ is called the value embedding of $\boldsymbol{v}_j$. The attended feature vector $\boldsymbol{v}_i^v$ is the combination of the original feature $\boldsymbol{v}_i$ and the sum of other clip features weighted by their corresponding attention scores. Here we adopt the convention of using superscript to denote the modality that is used as a query. For example, $\boldsymbol{v}_i^v$ refers to the visually-attended visual features of the i-th clip.

The self-attended features of the audio modality ($a \rightarrow a$) can be similarly defined as:

$$\boldsymbol{a}_i^a = \sum_{j=1}^{N} \omega_{a\rightarrow a}(\boldsymbol{a}_i, \boldsymbol{a}_j) g_{a\rightarrow a}(\boldsymbol{a}_j) + \boldsymbol{a}_i. \tag{9}$$

Bimodal Attention. The self-attention only captures the clip interactions within one modality. To capture the interactions across modalities, we extend the self-attention mechanism and devise a bimodal attention, shown in Fig. 3. In our bimodal attention, we compute the query from one modality, and the key and value from the other modality. This is presented in the following case of using visual features to attend to the audio, i.e. $v \rightarrow a$, as:

$$\boldsymbol{a}_i^v = \sum_{j=1}^{N} \omega_{v\rightarrow a}(\boldsymbol{v}_i, \boldsymbol{a}_j) g_{v\rightarrow a}(\boldsymbol{a}_j) + \boldsymbol{a}_i^a. \tag{10}$$

Here $\omega_{v\rightarrow a}(\boldsymbol{v}_i, \boldsymbol{a}_j)$ and $g_{v\rightarrow a}(\boldsymbol{a}_j)$ are similarly defined w.r.t. Eqs. (1)–(6). The bimodal attention mechanism allows the information from two different modalities to influence each other.

Noise Sentinel. Given the features from the self-attention and bimodal attention, a simple strategy is to directly fuse these features (e.g. via concatenation) for final prediction. Empirically, we have found that this strategy does not give the best performance. This is possibly due to the noise in the features. For example, if the audio features are noisy, the visually attended audio features from the bimodal attention layer are not reliable. In addition, using the noisy audio to query the visual features would also result in unreliable visual features. Intuitively, we would like our model to have the capacity to ignore a certain modality when it is noisy. Inpired by the visual sentinel [24] used in image captioning, we present a novel noise sentinel mechanism for this purpose.

The noise sentinel is a parameter within the bimodal attention layer. It has the same channel-dimension as the attended features within the attention layer, and is initialized to zero. It is then concatenated to each of the self-attended sequences as follows:

$$[\boldsymbol{a}_1^a, \boldsymbol{a}_2^a, \ldots, \boldsymbol{a}_N^a, \boldsymbol{n}_a], \tag{11}$$

$$[\boldsymbol{v}_1^v, \boldsymbol{v}_2^v, \ldots, \boldsymbol{v}_N^v, \boldsymbol{n}_v], \tag{12}$$

where $\boldsymbol{n}_a$ and $\boldsymbol{n}_v$ represent the noise sentinel parameters. We use the self-attended visual features $\boldsymbol{v}_i^v$ as queries to attend to the sequence in Eq. (11). Symmetrically, we use the the self-attended audio features $\boldsymbol{a}_i^a$ to attend to the sequence in Eq. (12). Note we do not use the noise embedding as a query. In what follows, we derive the visually-attended audio features $\boldsymbol{a}_i^v$. The audio-attended visual features $\boldsymbol{v}_i^a$ can be obtained by following-suit.

Let us denote the key for the audio noise embedding $\boldsymbol{n}_a$ as $\boldsymbol{k}_{\text{noise}}$, and the unary term for the noise embedding as m_{noise}. We define a few more variables for notational con-

venience.

$$\boldsymbol{q}_i = \boldsymbol{q}_{v\to a}(\boldsymbol{v}_i^v), \tag{13}$$
$$\boldsymbol{k}_j = \boldsymbol{k}_{v\to a}(\boldsymbol{a}_j^a),\ \boldsymbol{k}_{\text{noise}} = \boldsymbol{k}_{v\to a}(\boldsymbol{n}_a), \tag{14}$$
$$m_j = W_{m,v\to a}\boldsymbol{a}_j^a,\ m_{\text{noise}} = W_{m,v\to a}\boldsymbol{n}_a, \tag{15}$$
$$g(\boldsymbol{a}_j^a) = W_{g,v\to a}\boldsymbol{a}_j^a,\ g(\boldsymbol{n}_a) = W_{g,v\to a}\boldsymbol{n}_a. \tag{16}$$

We are ready to define the attention scores. The attention score from the visual to audio clips is

$$\omega_{v\to a}(\boldsymbol{v}_i^v, \boldsymbol{a}_j^a) = \frac{\exp(c\boldsymbol{q}_i^\intercal \boldsymbol{k}_j)}{\exp(c\boldsymbol{q}_i^\intercal \boldsymbol{k}_{\text{noise}}) + \sum_{l=1}^N \exp(c\boldsymbol{q}_i^\intercal \boldsymbol{k}_l)} + \frac{\exp(m_j)}{\exp(m_{\text{noise}}) + \sum_{l=1}^N \exp(m_l)}; \tag{17}$$

and the attention score between the visual clip and audio noise embedding is

$$\omega_{v\to a}(\boldsymbol{v}_i^v, \boldsymbol{n}_a) = \frac{\exp(c\boldsymbol{q}_i^\intercal \boldsymbol{k}_{\text{noise}})}{\exp(c\boldsymbol{q}_i^\intercal \boldsymbol{k}_{\text{noise}}) + \sum_{l=1}^N \exp(c\boldsymbol{q}_i^\intercal \boldsymbol{k}_l)} + \frac{\exp(m_{\text{noise}})}{\exp(m_{\text{noise}}) + \sum_{l=1}^N \exp(m_l)}. \tag{18}$$

Concretely, the visually attended audio features with noise sentinel is formulated by

$$\boldsymbol{a}_i^v = \sum_{j=1}^N \omega_{v\to a}(\boldsymbol{v}_i^v, \boldsymbol{a}_j^a) g_{v\to a}(\boldsymbol{a}_j^a) + \boldsymbol{a}_i^a + \omega_{v\to a}(\boldsymbol{v}_i^v, \boldsymbol{n}_a) g_{v\to a}(\boldsymbol{n}_a). \tag{19}$$

If the attended modality is noisy, our network can learn to ignore it by attending to the noise embedding instead. *e.g.* by setting the attention weight for the noise sentinel $\omega(\boldsymbol{v}_i^v, \boldsymbol{n}_a)$ much higher than the attention weights on the audio clips, $\omega(\boldsymbol{v}_i^v, \boldsymbol{a}_j^a)$.

Classifier. Let $\{\alpha_k\}_{k=1}^4$ be the set of learned weights with convex sum, $\sum_{k=1}^4 \alpha_k = 1$. The final feature representation for the i-th clip, $\boldsymbol{z}_i$, is then computed as a weighted sum of the unimodal self-attended and bimodally attended features, $\boldsymbol{z}_i = \alpha_1 \boldsymbol{v}_i^v + \alpha_2 \boldsymbol{v}_i^a + \alpha_3 \boldsymbol{a}_i^v + \alpha_4 \boldsymbol{a}_i^a$. The final bit of our framework is a two-layer network that outputs the highlight score of the i-th clip, as $s_i = \sigma\left(f_2\left(\text{ReLU}\left(f_1(\boldsymbol{z}_i)\right)\right)\right)$. Here each layer $f_i(\cdot)$, with $i \in \{1, 2\}$, contains layer-norm [2] and dropout [35] ($p = 0.5$), which is followed by a linear projection matrix W_i. The second layer $f_2(\cdot)$ specifically projects our features into a scalar, followed by the sigmoid activation $\sigma(\cdot)$ to produce the score s_i, indicating the probability of the i-th clip being a highlight.

Model Learning. The highlight detection datasets are highly imbalanced, since the majority of training clips are not highlights. We address this issue by adopting a weighted binary cross entropy loss, where a positive example (highlight) carries a higher weight (w_p) than a negative example (non-highlight). Our final training loss becomes $\mathcal{L} = \sum_{i=1}^N -[w_p y_i \log(s_i) + (1 - y_i)\log(1 - s_i)]$, with $y_i \in \{0, 1\}$ the ground-truth label of the i-th clip.

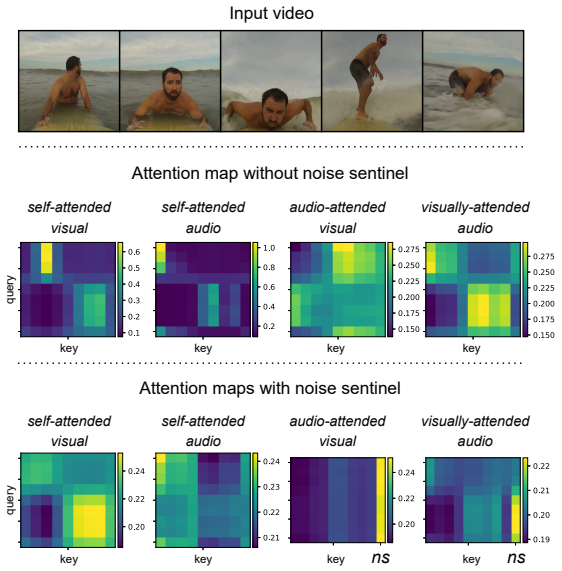

Figure 4. Attention maps for an example video (top row) without noise sentinel (middle row) and with noise sentinel (bottom row). *ns* denotes the attention weights placed on the noise sentinel (last column of the bimodal attention maps). This surfing video has noisy audio - the microphone constantly has water splashing against it, or is submerged in water. Consequently, our model chooses to attend largely to the noise sentinel for the audio attended visual features and visually attended audio features (bottom row) - since the audio is unreliable both as a query (there are no audio cues) and as a key (there is nothing of interest to attend to in the audio).

4. Experiments

4.1. Datasets and Setup

We utilize three datasets, namely the YouTube Highlights dataset [37], TVSum dataset [34], as well as the Video Titles in the Wild (VTW) dataset [50]. The YouTube Highlights dataset [37] contains six different categories: dog, gymnastics, parkour, skating, skiing, and surfing. There are approximately 100 videos for each category. Following the practice of prior efforts, we train a highlight detector for each category. The TVSum dataset contains 50 videos across ten categories. We follow prior works and train a highlight detector for each category using a random 80/20 split. The Video Titles in the Wild (VTW) dataset [50] contains highlight labels, but does not have categorical information. We follow the work of [50] and adopt their split of

	RRAE (V) [46]	GIFs (V) [12]	LSVM (V) [37]	CLA (V) [44]	LM (V) [44]	MN (V) [15]	MN (AV) [15]	Trail. (V) [42]	SA (V)	Ours (AV)
dog	0.49	0.308	0.60	0.502	0.579	0.5368	0.5816	0.633	**0.649**	0.645 (↓)
gym.	0.35	0.335	0.41	0.217	0.417	0.5281	0.6165	**0.825**	0.715	0.719 (↑)
park.	0.50	0.540	0.61	0.309	0.670	0.6888	0.7020	0.623	0.766	**0.808** (↑)
ska.	0.25	0.554	0.62	0.505	0.578	0.7094	**0.7217**	0.529	0.606	0.620 (↑)
ski.	0.22	0.328	0.36	0.379	0.486	0.5834	0.5866	**0.745**	0.712	0.732 (↑)
surf.	0.49	0.541	0.61	0.584	0.651	0.6383	0.6514	**0.793**	0.782	0.783 (↑)
Avg.	0.383	0.464	0.536	0.416	0.564	0.6138	0.6436	0.691	0.705 $^{\pm 0.004}$	**0.718** $^{\pm 0.006}$ (↑)

Table 1. Highlight detection results (mAP) on the YouTube dataset. Our visual only model, SA (V), outperforms all prior methods, and our full audio-visual model achieves state-of-the-art performance. Best and second-best results are in bold and with underline, respectively. (↑)/(↓) indicate an improvement/decline relative to the unimodal visual baseline, SA (V). We show the modalities used for each method in brackets: (V) for visual, and (AV) for audio-visual.

Method	mAP
VTW (V) [50]	0.583
SA (V)	0.722 ± 0.003
Ours (AV)	$\mathbf{0.812 \pm 0.002}$

Table 2. Highlight detection results (mAP) on the VTW dataset. Our approach outperforms baselines, and outperforms the prior state-of-the-art method [50] by 22.9% mAP. We show the modalities used for each method in brackets: (V) for visual, and (AV) for audio-visual.

2,000 videos for training, 300 for validation, and 2,000 for testing.

Features: On the YouTube Highlights and TVSum datasets, we follow the work of [42] and use a 3D CNN [13] with ResNet-34 [14] backbone pretrained on the Kinetics-400 dataset [4] to obtain the visual frame-level features. Since the 3D CNN performs temporal convolution over 16 consecutive frames, we consider a 3D feature to be a part of a clip if it overlaps by at least 50% with the clip.

On the VTW dataset, we follow the original work of [50], and use a C3D network [39] pretrained on Sports-1M [18] to obtain visual features. Each video is also divided into clips of 100 frames following [50].

For all datasets, we use a PANN audio network [22] pretrained on AudioSet [8] to obtain audio features that align with the visual clips. Frame-level features are average-pooled within each clip for both audio and visual features to generate a clip-level feature.

Implementation Details: We train our model using Adam [21], with a learning rate of 5×10^{-5}. We train for 30 epochs on the YouTube and TVSum datasets, and 5 epochs on the VTW dataset. Before the attention modules, we project each modality into a vector of fixed-size length 512. The key, query, and value vectors all follow the same size. The constant c for all attention modules is set to 0.06. We set the positive class weight $w_p = 5$ for the loss term.

Evaluation Metrics: We adopt the widely-used mean Average Precision or mAP as the evaluation metric for the YouTube and VTW datasets. On the TVSum dataset, we follow prior works [44] and adopt the mAP at top-5 metric. Following prior studies [37, 42, 12], a mAP/maP at top-5 score is separately computed for every video, because a highlighted moment in one video is not necessarily more interesting than non-highlight moments in other videos; we report the average mAP over all videos. We repeat each experiment 10 times with different random seeds, and report the average performance and standard deviation over these 10 trials.

4.2. Baselines

We define the following baseline approaches for comparison:

- SA (A) and SA (V): In the unimodal case (audio or visual features only), we use self-attention to obtain the self-attended features, and use the same classifier architecture to classify each clip. We dub this model SA (V) for self-attended visual model, and SA (A) for self-attended audio model. The SA (V) model is directly comparable to prior works

- SA (AV)$^{\text{early}}$: We try out a self-attention based audio-visual network that concatenates the audio and visual features in an early-fusion fashion, then does self-attention and uses the same classifier to classify each clip. We call this network SA (AV)$^{\text{early}}$.

- SA (AV)$^{\text{late}}$: In addition, we try out a self-attention based audio-visual network that sums the self-attended features from each modality in a late-fusion fashion, then uses the same classifier to classify each clip. We call this network SA (AV)$^{\text{late}}$.

	sLSTM (V) [52]	SM (V) [11]	SG (V) [26]	LM (V) [44]	DSN (V) [28]	VESD (V) [3]	Trail. (V) [42]	MN (AV) [15]	SA (V)	Ours (AV)
VT	0.411	0.415	0.423	0.559	0.373	0.447	0.613	0.8062	0.8337	**0.8370** (↑)
VU	0.462	0.467	0.472	0.429	0.441	0.493	0.546	**0.6832**	0.6469	0.5726 (↓)
GA	0.463	0.469	0.475	0.612	0.428	0.496	0.657	0.7821	**0.8444**	0.7845 (↓)
MS	0.477	0.478	0.489	0.540	0.436	0.503	0.608	0.8183	**0.8651**	0.8605 (↓)
PK	0.448	0.445	0.456	0.604	0.411	0.478	0.591	0.7807	0.7032	**0.8009** (↑)
PR	0.461	0.458	0.473	0.475	0.417	0.485	**0.701**	0.6584	0.6749	0.6922 (↑)
FM	0.452	0.451	0.464	0.432	0.412	0.487	0.582	0.578	0.6690	**0.7003** (↑)
BK	0.406	0.407	0.417	0.663	0.368	0.441	0.647	**0.7502**	0.6808	0.7300 (↑)
BT	0.471	0.473	0.483	0.691	0.435	0.492	0.656	0.8019	0.9496	**0.9741** (↑)
DS	0.455	0.453	0.466	0.626	0.416	0.488	**0.681**	0.6551	0.6079	0.6747 (↑)
Avg.	0.451	0.461	0.462	0.563	0.424	0.481	0.628	0.7324	$0.7476^{\pm 0.021}$	$\mathbf{0.7627}^{\pm 0.020}$ (↑)

Table 3. Highlight detection results (top-5 mAP) on the TVSum dataset. Our visual only model, SA (V), outperforms all prior methods, and our full audio-visual model achieves state-of-the-art performance. Best and second-best results are in bold and with underline, respectively. (↑)/(↓) indicate an improvement/decline relative to the unimodal visual baseline, SA (V). We show the modalities used for each method in brackets: (V) for visual, and (AV) for audio-visual. We see an improvement in seven out of ten categories over the unimodal SA (V) baseline with our full model.

Method	TVSum	YouTube	VTW
SA (V)	$0.748^{\pm 0.02}$	$0.705^{\pm 0.004}$	$0.722^{\pm 0.003}$
SA (A)	$0.687^{\pm 0.03}$	$0.670^{\pm 0.008}$	$0.794^{\pm 0.002}$
SA (AV)$^{\text{early}}$	$0.750^{\pm 0.01}$	$0.703^{\pm 0.004}$	$0.798^{\pm 0.003}$
SA (AV)$^{\text{late}}$	$0.750^{\pm 0.02}$	$0.709^{\pm 0.004}$	$0.805^{\pm 0.003}$
Ours (AV)	$\mathbf{0.763}^{\pm 0.02}$	$\mathbf{0.718}^{\pm 0.006}$	$\mathbf{0.812}^{\pm 0.002}$

Table 4. Comparison to Baselines: mAP/top-5 mAP with standard deviation on TVSum, YouTube, and VTW datasets for several baselines and our full model.

We also compare with state-of-the-art methods on each dataset. For each method, we use "(V)" if it only uses the visual signal, and "(AV)" if it is an audio-visual method.

4.3. Highlight Detection Results

YouTube Highlights: We show the main results on the YouTube dataset in Table 1. We achieve state-of-the-art performance on the YouTube dataset using our visual only model SA (V) model, outperforming a prior audio-visual network, MN (AV) [15]. We see a further gain of 1.3% when we use the proposed architecture. Our final architecture achieves better performance in five out of six categories, while maintaining reasonably good performance in the other category, dog.

TVSum: We show the results for the TVSum dataset in Table 3. Our visual-only model SA (V) outperforms the prior audio-visual state-of-the-art, and our audio-visual model improves performance by another 1.5%. We note that we see an improvement over the visual model in seven out of ten categories.

VTW: We show the results for the VTW dataset in Table 2. Our approach outperforms the visual only baseline by a large margin. This is due in part to the importance of audio on this dataset, as shown in Table 4, where the audio-only baseline SA (A) outperforms its visual counterpart by 7.2% mAP. Upon further investigation, we determined that this is because the dataset contains chiefly user generated videos, where highlight-worthy moments are typically accompanied by man-made sounds such as cheering and clapping. This strengthens our belief that audio can be a useful as a stand-alone signal for highlight detection.

Comparison to Baselines: We compare our full model to the three different baselines outlined in Sec. 4.2. The results are shown in Table. 4.

Our visual-only model SA (V) outperforms its audio counterpart SA (A) on the TVSum and YouTube datasets. On the VTW dataset SA (A) outperforms SA (V) by a large margin. Qualitative analysis of the VTW dataset determined that the interesting moments in the dataset were often accompanied by sounds like clapping and cheering.

Even so, our audio-based models SA (A) often perform on-par with visual state-of-the-art methods. On the TVSum and YouTube datasets, SA (A) outperforms all but the prior state-of-the-art on the respective dataset. This lends credence to the hypothesis that audio can be a useful stand-alone signal for highlight detection.

Our full model outperforms the multimodal early and late fusion baselines for all three datasets. This shows the effectiveness of our bimodal attention and noise sentinel components.

Qualitative Results: We visualize the attention maps produced by our approach in Fig. 4. The self-attention lay-

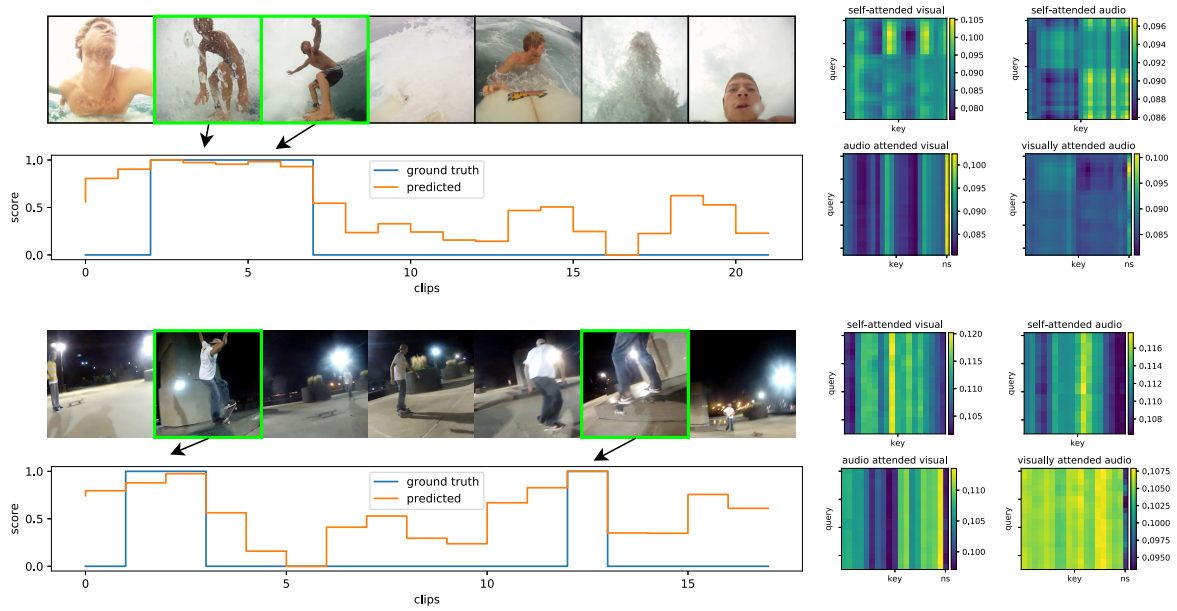

Figure 5. Qualitative results: Highlight moments are outlined in green, and the output of our model is plotted along with ground truth highlights below each video. The attention maps are visualized to the right of each video. The model attends to the noise sentinel *ns* for the surfing video (top), while attending to the opposite modality for the skating (bottom) video. The skating video has audio cues (cheering and skateboard noises) which our model utilizes.

ers produce similar attention maps. However, the attention maps of the bimodal attention layer exhibit a strong difference with and without the noise sentinel. With the noise sentinel, our model largely chooses to attend to the noise sentinel for the bimodally attended features. This particular video contains noisy audio, as the microphone is often submerged in water. Our approach learns to ignore the interaction between the modalities as the audio is not useful as a cue (query) nor as complementary information to the visual features (key, values). We present additional qualitative highlight detection results along with their attention maps in Fig. 5.

4.4. Ablation Studies

Impact of Disentangled Attention: In this experiment, we analyze the impact of disentangled attention compared with classic self-attention, where the attention score is formulated as $\omega_{v\to v}(\boldsymbol{v}_i, \boldsymbol{v}_j) = \text{softmax}(c\hat{\boldsymbol{q}}_{v\to v}(\boldsymbol{v}_i)^\intercal \hat{\boldsymbol{k}}_{v\to v}(\boldsymbol{v}_j))$ for self-attention and similarly for our bimodal attention layers. We show the result in Table 5. We see that disentangled attention improves our results for both datasets.

Impact of Noise Sentinel: In this experiment, we study the impact of the noise sentinel on the overall performance. We show the result in Table 6. The result demonstrates that using the noise sentinel improves the performance on both datasets.

Attention mechanisms	VTW	YouTube
Classic self-attention	0.808	0.705
Disentangled attention (ours)	**0.812**	**0.718**

Table 5. Ablation study on disentangled attention: We achieve the best result with the disentangled attention.

Noise sentinel	VTW	YouTube
W/o noise sentinel	0.804	0.716
W/ noise sentinel (ours)	**0.812**	**0.718**

Table 6. Ablation study on noise sentinel: We achieve superior performance with the noise sentinel.

5. Conclusion and Outlook

We proposed an audio-visual framework for video highlight detection. Our architecture models the relationship between audio and visual information through a bimodal attention layer to produce fused representations. We also introduced an adaptive noise sentinel mechanism that "looks away" from a noisy audio or visual modality. We empirically showed that our framework achieves superior performance on well-known benchmark datasets. While we focus on highlight detection, the proposed techniques could benefit other audio-visual tasks such as audio-visual speech recognition, action localization and video summarization.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 2

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 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. 2, 7

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017. 6

[5] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 39–54, 2018. 2

[6] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020. 2

[7] Ana Garcia del Molino and Michael Gygli. Phd-gifs: personalized highlight detection for automatic gif creation. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 600–608, 2018. 1, 2

[8] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 6

[9] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. *Proceedings of Advances in neural information processing systems (NeurIPS)*, 27:2069–2077, 2014. 2

[10] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 505–520, 2014. 2

[11] 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 (CVPR)*, pages 3090–3098, 2015. 2, 7

[12] 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 (CVPR)*, pages 1001–1009, 2016. 1, 2, 6

[13] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018. 2, 6

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[15] Fa-Ting Hong, Xuanteng Huang, Wei-Hong Li, and Wei-Shi Zheng. Mini-net: Multiple instance ranking network for video highlight detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 345–360, 2020. 1, 2, 6, 7

[16] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate RGB-D salient object detection via collaborative learning. In *ECCV*, pages 52–69, 2020. 2

[17] Yifan Jiao, Xiaoshan Yang, Tianzhu Zhang, Shucheng Huang, and Changsheng Xu. Video highlight detection via deep ranking modeling. In *Proceedings of the Pacific-Rim Symposium on Image and Video Technology*, pages 28–39, 2017. 1, 2

[18] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732, 2014. 6

[19] Aditya Khosla, Raffay Hamid, Chih-Jen Lin, and Neel Sundaresan. Large-scale video summarization using web-image priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2698–2705, 2013. 2

[20] Gunhee Kim and Eric P Xing. Reconstructing storyline graphs for image recommendation from web community photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3882–3889, 2014. 2

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 6

[22] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. 1, 2, 6

[23] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1346–1353, 2012. 2

[24] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 375–383, 2017. 2, 4

[25] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2714–2721, 2013. 2

[26] 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 (CVPR)*, pages 202–211, 2017. 2, 7

[27] Chong-Wah Ngo, Yu-Fei Ma, and Hong-Jiang Zhang. Automatic video summarization by graph modeling. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 104–109, 2003. 2

[28] 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 (ICCV)*, pages 3657–3666, 2017. 2, 7

[29] 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 (CVPR)*, pages 7083–7092, 2017. 2

[30] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. Category-specific video summarization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 540–555, 2014. 2

[31] Mrigank Rochan, Mahesh Kumar Krishna Reddy, Linwei Ye, and Yang Wang. Adaptive video highlight detection by learning from user history. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 261–278, 2020. 1, 2

[32] Mrigank Rochan and Yang Wang. Video summarization by learning from unpaired data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7902–7911, 2019. 2

[33] Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 347–363, 2018. 2

[34] 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 (CVPR)*, pages 5179–5187, 2015. 2, 5

[35] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 5

[36] George Sterpu, Christian Saam, and Naomi Harte. Attention-based audio-visual fusion for robust automatic speech recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 111–115, 2018. 2

[37] Min Sun, Ali Farhadi, and Steve Seitz. Ranking domain-specific highlights by analyzing edited videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 787–802, 2014. 1, 2, 5, 6

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

[39] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015. 2, 6

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. 3

[41] Jinjun Wang, Changsheng Xu, Engsiong Chng, and Qi Tian. Sports highlight detection from keyword sequences using hmm. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, volume 1, pages 599–602, 2004. 2

[42] Lezi Wang, Dong Liu, Rohit Puri, and Dimitris N Metaxas. Learning trailer moments in full-length movies. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 300–316, 2020. 1, 2, 6, 7

[43] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12695–12705, 2020. 2

[44] Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. Less is more: Learning highlight detection from video duration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1258–1267, 2019. 1, 2, 6, 7

[45] Ziyou Xiong, Regunathan Radhakrishnan, Ajay Divakaran, and Thomas S Huang. Highlights extraction from sports video based on an audio-visual marker detection framework. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 29–32, 2005. 2

[46] 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 (ICCV)*, pages 4633–4641, 2015. 1, 6

[47] 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 (CVPR)*, pages 982–990, 2016. 2

[48] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 191–207, 2020. 3

[49] Youngjae Yu, Sangho Lee, Joonil Na, Jaeyun Kang, and Gunhee Kim. A deep ranking model for spatio-temporal highlight detection from a 360○ video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 1, 2

[50] Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, and Min Sun. Generation for user generated videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 609–625, 2016. 1, 5, 6

[51] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Summary transfer: Exemplar-based subset selection for video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1059–1067, 2016. 2

[52] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 766–782, 2016. 2, 7

[53] Ke Zhang, Kristen Grauman, and Fei Sha. Retrospective encoders for video summarization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 383–399, 2018. 2

[54] Miao Zhang, Wei Ji, Yongri Piao, Jingjing Li, Yu Zhang, Shuang Xu, and Huchuan Lu. LFNet: Light field fusion network for salient object detection. *IEEE Transactions on Image Processing*, 29:6276–6287, 2020. 2

[55] Miao Zhang, Jingjing Li, Wei Ji, Yongri Piao, and Huchuan Lu. Memory-oriented decoder for light field salient object detection. In *NeurIPS*, pages 896–906, 2019. 2

[56] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. Hierarchical recurrent neural network for video summarization. In *Proceedings of the 25th ACM Internation Conference on Multimedia*, pages 863–871, 2017. 2

[57] Kaiyang Zhou, Yu Qiao, and Tao Xiang. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2