Probing Visual-Audio Representation for Video Highlight Detection via Hard-Pairs Guided Contrastive Learning

Shuaicheng Li*[1]
lishuaicheng@sensetime.com
Feng Zhang*[1]
zhangfeng4@sensetime.com
Kunlin Yang[1]
yangkunlin@sensetime.com
Lingbo Liu[2]
liulingbo918@gmail.com
Shinan Liu[1]
liushinan@sensetime.com
Jun Hou†[1]
houjun@sensetime.com
Shuai Yi[1]
yishuai@sensetime.com

[1] Sensetime Research
[2] The Hong Kong Polytechnic University
Hong Kong, China

Abstract

Video highlight detection (VHD) is a crucial yet challenging problem which aims to identify the interesting moments in untrimmed videos. The key to this task lies in effective video representations that jointly pursue two goals, *i.e.*, 1) cross-modal representation learning and 2) fine-grained feature discrimination. To issue 1), the dominant VHD models adopt cross-attention based transformer to learn audio-visual information and inter-modality alignment. They always assume that multi-modal signals are synchronized which may not hold in practice due to spurious noise and appearance shift in untrimmed videos. To relieve this problem, we propose a cross-modality co-occurrence encoding by considering not only single visual/audio but asynchronous cross-modal correlations. We also explore the additional global contextual information abstracted from local region to further promote the inter-modality learning. To issue 2), to enlarge the discriminative power of feature embedding, we propose a hard-pairs guided contrastive learning (HPCL) scheme to reflect intrinsic semantic representation. A hard-pairs sampling strategy is employed in HPCL to mine the hard segment samples for improving feature discrimination and providing significant gradient information. Extensive experiments conducted on two benchmarks demonstrate the effectiveness and superiority of our proposed methods compared to other state-of-the-art methods.

* Equal contribution
† Corresponding author

Figure 1: Schematic depiction of multi-modal representation learning. (a) within-modality learning with cross-attention. (b) Our proposed method explores the inter and intra-relations by measuring intra-visual, intra-audio and cross-modality co-occurrence. And HPCL is introduced for feature discriminative.

1 Introduction

In recent years, posting well-edited video shining moments on social platforms, *e.g.*Youtube, Tiktok, has become our daily routine. Due to the labor for cropping untrimmed videos, video highlight detection task has drawn extensive attention from the research community. The goal of this task is to localize the highlight segments and trim shining moments from untrimmed long videos automatically, which has a wide range of downstream applications such as video summarization [17, 32] and detection [23, 34].

Most approaches make sorts of efforts to discriminate highlight and non-highlight clips. Pair-based approaches [18, 20, 31, 42] assumed that there exists distinguishable appearances between highlight and background segments. A ranking model was trained based on pairs (*highlight, non-highlight*) to rank segment scores and select shining moments [6, 40, 45]. [1, 17, 33] developed an audio-visual network to assemble multi-modal representations, indicating that better representation modeling benefits more for highlight detection performance. Therefore, an essential question can be asked: *How to fully exploit video representations?*

Intuitively, it should not only (1) capture multi-modality contextual information, but also (2) be well distinguishable to inter-segments. **To issue (1)**, a main stream of efforts delves into effective feature learning, *e.g.*, cross-modal signals fusion [1, 17, 24]. As shown in Figure 1(a) and (b), there are two directions on handing multi-modal data[1, 9, 33]. The first is modeling cross-modality representations by cross-attention modules (Figure 1 (a)) such as [1, 9, 33]. However, these methods are sub-optimal for exploiting the complex relationships between inter-modality since they are based on the assumption that multiple signals are synchronized, which may not hold in practice with spurious noise and indistinct correspondence between these modalities. **With regard to issue (2)**, prior studies [20, 31, 42] employ ranking models to facilitate segment pairs discrimination. They only push away the dissimilar pairs by ranking loss and do not reflect intrinsic semantic representation. Moreover, since there exists highly similar content for consistent video segments, it is essential to focus on the distinction between highlight segments and its surrounding non-highlight clips.

To address the above challenges, this paper goes deeper into designing visual-audio architectures by two views: (1) *cross-modal relations alignment and learning* and (2) *inter-segment feature discrimination*. We propose a novel visual-audio framework for highlight detection. Specifically, in addition to extract the modal-wise information by self-attention mechanism, we explore the dependencies between within-modality features and exclude the unrelated clues to facilitate the specialized characteristic of inter-segment alignment by

cross-modality co-occurrence encoding. We further explore the additional learnable context to enhance intra-modality representations by implicitly modeling statistics over the entire training data. In the latter view (2), we propose a supervised dense hard-pairs guided contrastive loss (HPCL) for feature discrimination without requiring any additional data argument tricks as most prior works do in mainstream self-supervised works [8, 16]. This is achieved by a) using categorical information as a contrastive factor under a supervised setting and b) mining hard-pairs to provide a significant gradient contribution for enhancing discriminative power. In a), the data samples are trained in individual videos where positive and negative query is determined by its ground truth. HPCL shapes their embedding space in a discriminative manner by pulling in similar samples against dissimilar ones. In b), a hard-pair mining regularization strategy is introduced to make better use of informative video segments and let the model pay more attention to those discriminative-hard segments.

The main contributions can be summarized as follows:

- We propose a novel visual-audio VHD framework to capture intra-modality and inter-modality representations and exploit cross-modality relations and exclude unrelated clues for inter-segment alignment. With this simple and effective framework, semantic representations can be learned robustly, which is important for accurately identifying highlight segments.

- A supervised hard-pairs guided contrastive learning scheme is deployed to reflect structural representation of video sequence. Besides, a hard-pairs mining regularization is introduced to make better use of those discriminative-hard segments caused by temporal consistency in video sequence.

- Extensive experiments are conducted on the YouTube Highlights and TVsum benchmarks, and our proposed method outperforms other state-of-the-art methods. Detailed ablation studies demonstrate the effectiveness of our novel components.

2 Related Work

Video Highlight Detection. The goal of the video highlight detection task is to predict the highlight moments according to the semantic features on the untrimmed videos. [20, 31, 42] treat the video highlight detection task as a pair-based ranking to select shining moments Recent methods[1, 40] propose to use self-attention mechanism to capture contextual features. These methods utilize the temporal relations between segments and achieve excellent performance. Joint-VA [1] develops a cross-attention module following other works [29, 33] to exploit cross-modal features and then utilizes noise sentinel to relieve the feature confusion. And TCG[43] develops a low-rank audio-visual tensor fusion to capture the complex association between two modalities. These works are usually based on the assumption that audio and visual data are synchronized and highly correlated [22, 26]. It may not hold in practice with indistinct correspondence between inter-modality. We utilize the segment-wise attention to selectively capture the fine-grained relations between inter-modality and dampen the noise in both modalities.

Contrastive Learning. Many studies [7, 8, 16, 35] on unsupervised representation learning concentrate on the central concept: contrastive learning. They generate several positive augmented version by perturbations while negative data are randomly sampled from the other images. They typically consider contrastive learning as pre-training step and use the variant

versions as positive samples in unsupervised setting. Different from these methods, we raise a segment-wise dense contrastive learning scheme in the fully supervised setting with the known categorical information for contrastive factor. We also present a hard-pairs regularization strategy tailored for our video task to enlarge the discriminative power and specialize in hard shining moments caused by temporal consistency.

3 Approach

3.1 Architecture

Given an arbitrary unedited video sequence $V = \{v_t\}_{t=1}^{T}$ containing T segments, each segment v_t is annotated as binary label $y_t \in \{0,1\}$, indicating whether v_t contains the interesting part about categorical moments. Models are aiming to predict the label (*i.e., highlight or non-highlight*) of every segment. Our proposed method is illustrated in Figure 2. The CNN feature extractors output the visual and audio feature separately. Then, the inter and intra-relationships are explored through the inter and intra-modality encoding modules. Finally, the output representations of Co-occurrence Encoding, together with the Intra-Modality features are used to generate the segment-level confidence scores by three FC classifiers respectively and obtain the final highlight detection results by the weighted sum of final scores. For the output segment-wise representations , we utilize HPCL to compute segment-to-segment contrast to regularize the latent embedding space.

3.2 Feature Extractor

Given an untrimmed video sequence, the visual features are extracted by a pre-trained 3D backbone E_v, while audio information is extracted by a audio pre-trained network E_a following the previous methods [1, 38]. The visual and audio features of each segment are then flattened into a feature vector and are further transformed to the same embedding space with a linear layer respectively. Thus, the visual and audio features of the whole video can be denoted as $\mathbf{F}^v \in \mathbb{R}^{T\times d}$ and $\mathbf{F}^a \in \mathbb{R}^{T\times d}$ respectively. And $d = 256$ in this work.

3.3 Feature Encoding Module

Intra-Modality Encoding. We leverage the standard self-attention mechanism [37] to model within-modality relations and dampen the irrelevant modality. The modality-wise attention are deployed to embed contextual features. For either of the two modality, it models the relation between different segments and outputs a feature sequence $\hat{\mathbf{F}} \in \mathbb{R}^{T\times d}$ enhanced with temporal context. It's not enough to just capture the relationships within the segments. Motivated by the learned query proposed by [4], we introduce a decoder to parse the uni-modal features $\hat{\mathbf{F}}$ (omit the layer number n for clarity) and contextualize the global features where the decoder is implemented with the pure transformer decoder. For visual stream, we formulate a learnable parameters $\mathbf{G}_{\texttt{init}} \in \mathbb{R}^{d}$ as initial global context, thus implicitly modeling statistics over the entire training data. This global "query" specializes in abstracting statistics-based global embedding over all videos instead of depending on a certain global context for the corresponding video. It is different from and complementary to the previous attention-based multi-modal methods. In detail, the decoder takes the updated uni-modal

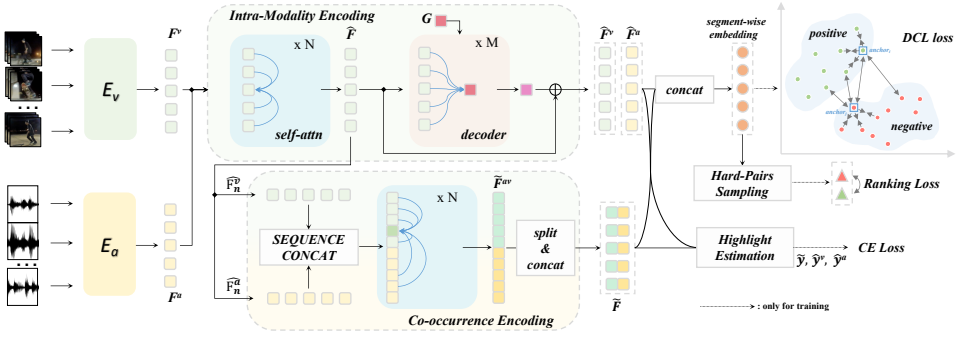

Figure 2: Illustration of our proposed method. E_v and E_a extract the high-level visual and audio features respectively and then followed the intra-modality encoding for modeling visual and audio representation separately. In addition, the cross-modality co-occurrence encoding is employed to exploit inter-modality relations. Lastly, for the output embeddings, we view segment-level representation as a point, a dense contrastive learning is proposed to shape the structural information in a discriminative manner.

features $\hat{\mathbf{F}}$ and $\mathbf{G}_{\texttt{init}}$ as input. It views $\mathbf{G}_{\texttt{init}}$ as *query* and uni-modal features $\hat{\mathbf{F}}$ as *values*, and then the *query* aggregates the uni-modal context information and abstracts global informative representation represented as $\mathbf{G}$. Finally, a straightforward method is to directly sum the global context and video representations, which can be formulated as $\hat{\mathbf{F}} = \hat{\mathbf{F}} + \mathbf{G}$. Similarly, in the audio stream, we perform mirroring operations on audio features, which will not be repeated due to space reasons.

Co-occurrence Encoding. Previous works [1, 2, 17] usually utilize cross-modal encoder to capture semantic associations based on these multi-modal signals. However, it can be sub-optimal since these works are usually based on the assumption that multi-modal signals are synchronized which may not hold in practice due to indistinct correspondence between these modality. Additionally, the cross-attention may introduce the cluttered background and inaccurate modality content since it restricts that modal A must build correlations with modal B. Our proposed multi-modality encoding method relaxes this condition to allow cases where only visual or audio modal is useful. This simple and effective module is useful to robustly learn semantic representation and is complementary to intra-modality encoding. Ideally, we would like the proposed method can dampen the noise and selectively choose effective information from multi-modality instead of all in them. It can relieve the inter-modality asynchronization by learning to ignore the cross-modal segment features with spurious noise and augment the intimate ones. Our cross-modality co-occurrence encoding is built upon the attention mechanism in canonical transformer decoder and takes the full sequence of segment embeddings corresponding to all visual and audio features $\hat{\mathbf{F}}^v = \{f_{1'}^v, ..., f_T^{v'}.\}, \hat{\mathbf{F}}^a = \{f_{1'}^a, ..., f_T^{a'}.\}$ as input. Assume the sequence modalities input $\mathbf{F}^{va} = \{f_{1'}^v, ..., f_T^{v'}, f_1^{a'}, ..., f_T^{a'}\} \in \mathbb{R}^{2T \times d}$, we alternatively contextualize the co-occurrence

information for each modality. In the visual modality, the process can be defined as,

$$Q^{v}_{\text{dec}} = W^{q}_{\text{dec}}\mathbf{F}^{v}_{n}, \quad K^{va}_{\text{dec}} = W^{k}_{\text{dec}}\mathbf{F}^{va}, \quad S^{va}_{\text{dec}} = W^{s}_{\text{dec}}\mathbf{F}^{va} \tag{1}$$

$$\tilde{\mathbf{F}}^{v} = \text{softmax}(\frac{Q^{v}_{\text{dec}}K^{va\ \text{T}}_{\text{dec}}}{\sqrt{d_k}})S^{va}_{\text{dec}} \tag{2}$$

where $W^{q}_{\text{dec}}, W^{k}_{\text{dec}}, W^{s}_{\text{dec}}$ are learnable parameters and used to linearly transform the input to the query, key and value. The other decoder is also applied to exploit the associations between the audio features $\hat{\mathbf{F}}^{a}$ and the sequence modality features $\mathbf{F}^{va}$ and then generate the co-occurrence representations $\tilde{\mathbf{F}}^{a}$.

3.4 Hard-Pairs Guided Contrastive Learning

Segment-wise Contrastive Loss. Our HPCL replaces the current image-wise training strategy with a segment-to-segment intra-video dense paradigm. The HPCL is applied to regularize the output feature embedding space using categorical information as contrastive factor. In the training phase, given a target video sequence containing T segments with labels $\{y_i\}_{i=1}^{T}$, we first aggregate the input embeddings $\hat{\mathbf{F}}^{v}, \hat{\mathbf{F}}^{a}$ into the segment-wise representations $\hat{\mathbf{F}} = \{\hat{f}_i\}_{i=1}^{T} \in \mathbb{R}^{T\times 2d}$. Then, for the segment *query* with label y, the positive *keys* are the other segments labeled y while the negative *keys* are the segments belonging to the other class. Our dense segment-level loss aims to contrast positive keys against negative ones. Formally, it can be defined as,

$$\mathcal{L}_{\text{HPCL}} = \frac{1}{|T|}\sum_{q\in\hat{F}}\frac{1}{\Gamma^{P}_{q}}\sum_{k_+\in\Gamma^{P}_{q}} -\log\frac{exp(q\cdot k_+/\tau)}{exp(q\cdot k_+) + \sum\limits_{k_-\in\Gamma^{N}_{q}} exp(q\cdot k_-/\tau)} \tag{3}$$

where the video sequence contains T segments, $\Gamma^{P}_{q}, \Gamma^{N}_{q}$ represent the segment representation sets of positive and negative keys for the query $q \in \hat{\mathbf{F}}$ separately.

Hard-Pairs Regularization Strategy. Previous methods [19, 25, 30] verify that mining negative samples are likely to be more useful and provide significant gradient information. In our fully supervised setting, the negative data in contrastive learning are *true negative* exactly. Thus, we would ask *what makes a good negative samples in supervised learning?* The most useful negative samples are ones that the embedding currently believes to be similar to the query since the hardest points are those close to the query, and are expected to have a high propensity to have the same label. In order to improve the feature discriminating power in HPCL, we first sample these hard-pairs for video sequence and then utilize the ranking loss to optimize them. Specifically, given a video sequence with T segments and positive masks $\{y_i \in \{0,1\}\}_{i=1}^{T}$, the water-sheds formulated as the boundaries from *positives vs. negatives* are identified and denoted as $\{c_j\}_{j=1}^{W}$. Here c_j is the index of video segments and W represents the number of the water-sheds. For a water-shed c_j, we sample indexes according to c_j including $\Upsilon_1 = \{c_j - k\}_{k=1}^{L}$ and $\Upsilon_2 = \{c_j + k\}_{k=1}^{L}$, where it would be replaced with c_j if $c_j - k < 0$. $L = 3$ is the region size. The hard-pairs are represented by $\Upsilon = \{(c_j - k, c_j + k)\}_{k=1}^{L}$. The loss is employed to optimize these hard-pairs, which is formulated as,

$$\mathcal{L}_{\text{rank}} = \sum_{p\in\Upsilon} max(margin - d(p), 0) \tag{4}$$

where $d(p)$ represents the Euclidean distance between the features indexed by the pairs p. *margin* is a hyper-parameter. We set $margin = 0.7$.

Category	Uni-Modality					Multi-Modality				
	RRAE [40]	LIM-s [35]	Video2GIF [8]	LSVM [26]	SL[20]	MN[8]	Joint-VA [9]	TCG [31]	Ours	Ours*
dog	0.49	0.579	0.308	0.60	**0.708**	0.537	0.645	0.553	0.678	0.690
gymnastics	0.35	0.417	0.335	0.41	0.532	0.528	**0.719**	0.626	0.681	0.660
parkour	0.50	0.670	0.540	0.61	0.772	0.689	0.808	0.709	0.791	**0.890**
skating	0.25	0.578	0.554	0.62	0.725	0.709	0.620	0.691	0.740	**0.741**
skiing	0.22	0.486	0.328	0.36	0.661	0.583	**0.732**	0.601	0.719	0.690
surfing	0.49	0.651	0.541	0.61	0.762	0.638	0.783	0.598	**0.822**	0.811
Average	0.383	0.564	0.464	0.536	0.693	0.614	0.718	0.630	0.739	**0.747**

Table 1: Experimental results comparisons of highlight detection on YouTube Highlight dataset in terms of mAP. Notice that the model 'Ours' utilizes 3D CNN [34] as visual feature extractor following previous work [1, 17], while 'Ours*' uses I3D [4] to extract visual features. Uni-Modality represents the methods that only employing visual features while Multi-Modality represents those visual-audio methods.

The proposed HPCL scheme and the segment-wise cross-entropy loss are complementary to each other. They can fully exploit the meaningful features for highlight detection. For the multi-modal predicted scores $\tilde{y}, \hat{y}^v, \hat{y}^a$, the weighted sum of training target are :

$$\mathcal{L}_{\text{ce}} = L(\tilde{y}, y) + L(\hat{y}^v, y) + L(\hat{y}^a, y) \tag{5}$$

where y is the target lables and $L(\cdot)$ denotes the CE loss. Thus, the overall loss function is formulated as,

$$\mathcal{L}_{\text{hld}} = \lambda_1 \mathcal{L}_{\text{ce}} + \lambda_2 \mathcal{L}_{\text{HPCL}} + \lambda_3 \mathcal{L}_{\text{rank}} \tag{6}$$

where $\lambda_1, \lambda_2, \lambda_3$ denotes the hyper-parameter to balance the terms. We set $\lambda_1 = 1, \lambda_2 = 0.3, \lambda_3 = 0.1$.

4 Experiments

In this section, we conduct extensive experiments on two challenging benchmarks, *i.e.*, YouTube highlights[31] and TVSum[28] to demonstrate the effectiveness of the proposed method. In our experiments, we use the standard networks 3D CNN [34] with ResNet-34 [15] and I3D[4] which pretrained on the Kinetics-400[5] dataset as our visual backbones. The audio backbone network uses PANN [21] audio network pretrained on AudioSet [10]. More details on datasets and implementations are available in the supplementary material.

4.1 Comparison with State-of-the-Art

We compare our proposed method with the other state-of-the-art methods [1, 13, 17, 32, 39, 40, 41, 43, 44] on two widely adopted benchmarks, *i.e.*, YouTube Highlights and TVSum. Specifically, we also present the results using visual feature extractor I3D[4].

Results on YouTube Highlights. The results are listed in Table 1. Our methods achieve superior performance compared with all of the aforementioned methods with a considerable margin when we adopt the same visual feature extractor 3DCNN[34] as the works [1, 17]. For instance, our method improves the mAP of *skating* from 0.620 in Joint-VA to 0.740. The performance of *parkour* is boosted from 0.808 to 0.890. It indicates that fully exploiting intra-modality and inter-modality relations benefit the detection result. The average result can be further improved by 0.8% when we use the I3D [4] backbone for visual features.

Category	Uni-Modality							Multi-Modality				
	vsLSTM [47]	SM [12]	VESD [4]	LIM-s [39]	KVS [27]	DPP [11]	SL [40]	MN[17]	Joint-SA [1]	TCG [13]	Ours	Ours*
VT	0.411	0.415	0.447	0.559	0.353	0.399	0.865	0.806	0.837	0.850	0.894	**0.908**
VU	0.462	0.467	0.493	0.429	0.441	0.453	0.687	0.683	0.573	0.714	0.714	**0.728**
GA	0.463	0.469	0.496	0.612	0.402	0.457	0.749	0.782	0.785	0.819	0.844	**0.846**
MS	0.477	0.478	0.503	0.540	0.417	0.462	0.862	0.818	0.861	0.786	0.795	**0.850**
PK	0.448	0.445	0.478	0.604	0.382	0.437	0.790	0.781	0.801	**0.802**	0.779	0.783
PR	0.461	0.458	0.485	0.475	0.403	0.446	0.632	0.658	0.692	0.755	0.743	**0.780**
FM	0.452	0.451	0.487	0.432	0.397	0.442	0.589	0.578	0.700	0.716	0.704	**0.728**
BK	0.406	0.407	0.441	0.663	0.342	0.395	0.726	0.750	0.730	**0.773**	0.761	0.771
BT	0.471	0.473	0.492	0.691	0.419	0.464	0.789	0.802	**0.974**	0.786	0.891	0.895
DS	0.455	0.453	0.488	0.626	0.394	0.449	0.640	0.655	0.675	0.681	0.703	**0.723**
Average	0.451	0.452	0.481	0.563	0.395	0.440	0.733	0.731	0.763	0.768	0.783	**0.801**

Table 2: Comparison of the highlight detection performances with state-of-the-arts on the TVSum test split in terms of top-5 mAP.

Architecture Variants	Average Results	
	YouTube Highlights	TVSum
V Only	0.659	0.763
A Only	0.651	0.752
AV	0.675	0.784
CR-AV	0.697	0.789
CO-AV (ours)	**0.747**	**0.801**

Table 3: Ablation Study on the various modifications of our proposed method.

Learning Scheme	Average Results	
	YouTube Highlights	TVSum
$\mathcal{L}_{ce}$ (baseline)	0.702	0.766
$\mathcal{L}_{ce} + \mathcal{L}_{HPCL}$	0.733	0.792
$\mathcal{L}_{ce} + \mathcal{L}_{HPCL} + \mathcal{L}_{rank}$	**0.747**	**0.801**

Table 4: Ablation Study on the effect of dense contrastive learning scheme.

Results on TVSum. We also provide the detailed comparisons with previous works as shown in Table 2. The results with visual features extracted by [14] can reach 0.762, which outperforms most methods with the same backbones [1, 17]. A considerable improvement is achieved by using the backbone I3D [4]. We speculate that the I3D captures high-level features with larger receptive field, which benefits our feature discrimination using HPCL. It is noting that the cross-attention method [1] achieves 0.763 and performs lower than our performance 0.801, indicating the benefits of intra-modality and inter-modality learning and feature discrimination over the simple cross-attention mechanism.

4.2 Ablation Study

Architectures Variants. To intuitively show the effectiveness of the proposed method, we present the following various modifications of our proposed methods: 1) **A (V) Only:** we only utilize the audio (visual) signals for feature learning in our work and discard the visual (audio) stream and co-occurrence encoding. 2) **AV:** the visual and audio features extracted by feature extrator are simply aggregated by concatenation and then projected into the intra-modality module for feature modeling. 3) **CR-AV:** following the implementation of [1], we employ the cross-attention blocks to our cross-modality module for the audio-visual signals modeling. 4) **CO-AV**: Our final architecture with intra-modality and cross-modality co-occurrence encoding. It is worth noting that all variants introduce the HPCL for model optimization. The model setting follows our final architectures for all modifications. Table 3 summarizes the results of these architectures variants. The cross-modality representation modeling can generally improve the performance from 0.675 to 0.697 in YouTube Highlights as shown in the third and forth rows in Table 3. Furthermore, compared to the **CR-AV**, our method with co-occurrence encoding (**CO-AV**) can boost the performance from 0.697 to 0.747 in YouTube Highlights dataset, showing the superiority and effectiveness of our intra-

Methods	Initial Manner		Average Results	
	Random	Mean	Youtube Highlight	TVSum
Ours w/oG			0.710	0.789
Ours w/G	✓		0.744	**0.803**
Ours w/G		✓	**0.747**	0.801

Table 5: Ablation Study on the effect of Global Representation. *random* represents randomly initialize global embedding while *Mean* utilizes the mean of input segment-wise embedding.

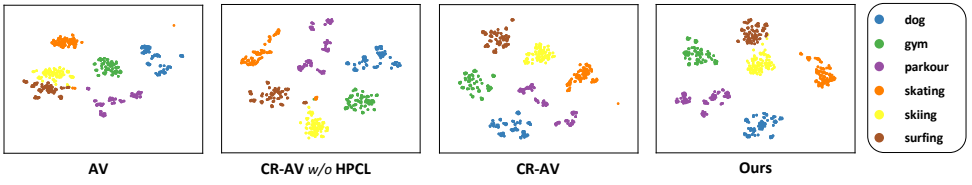

Figure 3: t-SNE plots of feature embedding for the testing split of YouTube Highlights. We visualize four architecture variants for better comparison. Each segment-level embedding is viewed as a point and the segment belonging the same category have the same color. Best view in color.

modality and cross-modality co-occurrence representation encoding.

Effect of HPCL. We validate the design of our HPCL scheme as shown in Table 4. We formulate that the **Baseline** discards the contrastive loss and hard-pairs rank loss and only utilizes segments-wise cross-entropy loss for highlight detection. The results between the first and second rows suggest that applying contrative loss in supervised setting improve the performance from 0.702 to 0.733 in YouTube Highlights and 0.766 to 0.792 in TVsum, which demonstrate the superiority of our HPCL. Also, the hard-pairs sampling is further boost the performance by 1.4%, validating our analysis that mining hard pairs is helpful for discriminating power improvement.

Effect of Global Representation. In this experiment, we verify that global representation play an important role for semantic feature exploiting as shown in Table 5. As we can see, there is a little difference between the performances of the model with or without global decoder. The model using random initial global embedding performs slight worse than using the mean of segment-wise features.

Early Fusion & Late Fusion. Based on our proposed method, we obtain multi-modal feature embeddings from multiple modalities. We conducted a comparative experiment for early fusion and late fusion on multi-modal features. **Early fusion**: Multi-modal features are first concatenated and then process to the classifier for final prediction. **Late fusion**: Multi-modal features are first used to predict highlight scores by specific classifiers, and then fuse the final scores with the weighted terms. The results are shown in Table 6. Early fusion manner only performs slight better than the late fusion, verifying the flexibility of our network.

4.3 Visualization

Feature Distribution Visualization. We apply the t-SNE [36] to the aggregated visual and audio representations on the YouTube Highlights dataset. Figure 3 displays the t-SNE visualization of our architecture variants as illustrated in Sec. 4.2. We find **CR-AV** *w/ HPCL*

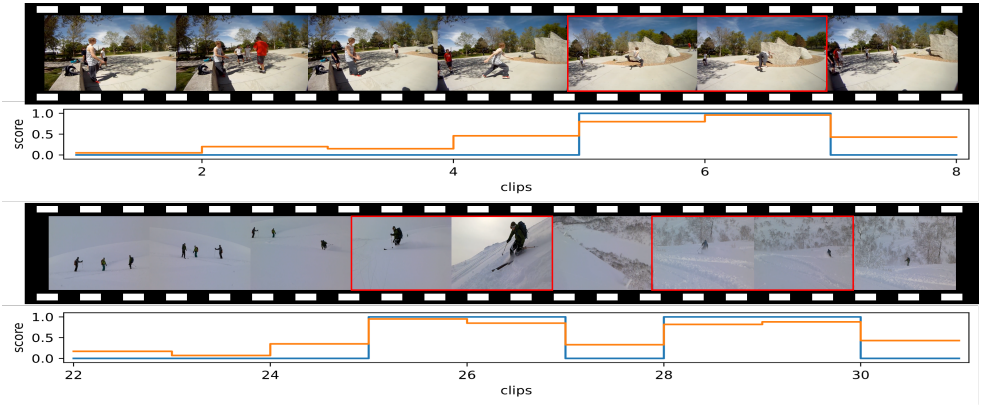

Figure 4: Qualitative results. We show highlight detection results on the test set of YouTube Highlights. The red box represent the ground truth segments.

Fusion Methods	Average Results	
	YouTube Highlights	TVSum
Early fusion	0.741	0.796
Late fusion	0.747	0.801

Table 6: Ablation Study for various fusion method.

performs well compared to the original **CR-AV** *w/o* HPCL, showing the strong intra-class compactness and inter-class dispersion . In addition, when integrating the HPCL and cross-modality co-occurrence as our final model, the features are better separated.

Qualitative Results. As shown in Figure 4, we display some qualitative results on the YouTube Highlights dataset. Our proposed method can successfully detect the shining moments, and the highlight moments and background scenes can be well distinguished.

5 Conclusion

This paper proposes a novel highlight detection methods, which aims to pursue two confounding goals: 1) cross-modal relations alignment and learning; 2) inter-segments feature discrimination. In the former case, we propose a visual-audio network to capture cross-modal representation by measuring within-modality relations. To enhance the video representation, we also introduce a global decoder to abstract global informative features by selectively integrating the segment-level representations. In the later case, a hard-pairs guided contrastive learning scheme is introduced to shape segment representations by improving intra-class compactness and inter-class dispersion in a discrminative manner with hard-pairs sampling strategy. Extensive experiments conducted on two widely adopted benchmarks demonstrate the effectiveness and superiority of our proposed method compared to previous methods.

References

[1] Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, and Li Cheng. Joint visual and audio learning for video highlight detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8127–8137, 2021.

[2] Sijia Cai, Wangmeng Zuo, Larry S. Davis, and Lei Zhang. Weakly-supervised video using variational encoder-decoder and web prior. In *European Conference on Computer Vision (ECCV)*, 2018.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229, 2020.

[4] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

[6] Runnan Chen, Penghao Zhou, Wenzhe Wang, Nenglun Chen, Pai Peng, Xing Sun, and Wenping Wang. Pr-net: Preference reasoning for personalized video highlight detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7980–7989, 2021.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning (ICML)*, pages 1597–1607. PMLR, 2020.

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision (ECCV)*, pages 104–120, 2020.

[10] 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 *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[11] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video. In *Advances in Neural Information Processing Systems*, 2014.

[12] Michael Gygli, Helmut Grabner, and Luc Van Gool. Video summarization by learning submodular mixtures of objectives. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[13] Michael Gygli, Yale Song, and Liangliang Cao. Video2gif: Automatic generation of animated gifs from video. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1001–1009, 2016.

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

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 9729–9738, 2020.

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

[18] Yifan Jiao, Zhetao Li, Shucheng Huang, Xiaoshan Yang, Bin Liu, and Tianzhu Zhang. Three-dimensional attention-based deep ranking model for video highlight detection. *IEEE Transactions on Multimedia*, pages 2693–2705, 2018.

[19] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, pages 21798–21809, 2020.

[20] Hoseong Kim, Tao Mei, Hyeran Byun, and Ting Yao. Exploiting web images for video highlight detection with triplet deep ranking. *IEEE Transactions on Multimedia*, 20(9): 2415–2426, 2018.

[21] 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*, pages 2880–2894, 2020.

[22] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018.

[23] Shuaicheng Li, Qianggang Cao, Lingbo Liu, Kunlin Yang, Shinan Liu, Jun Hou, and Shuai Yi. Groupformer: Group activity recognition with clustered spatial-temporal transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13668–13677, 2021.

[24] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno, and Tetsuya Ogata. Audio-visual speech recognition using deep learning. *Applied Intelligence*, pages 722–737, 2015.

[25] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4004–4012, 2016.

[26] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[27] Danila Potapov, Matthijs Douze, Zaïd Harchaoui, and Cordelia Schmid. Category-specific video. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *European Conference on Computer Vision (ECCV)*, 2014.

[28] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[29] 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.

[30] Yumin Suh, Bohyung Han, Wonsik Kim, and Kyoung Mu Lee. Stochastic class-based hard example mining for deep metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7251–7259, 2019.

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

[32] Min Sun, Ali Farhadi, and Steven M. Seitz. Ranking domain-specific highlights by analyzing edited videos. In *European Conference on Computer Vision (ECCV)*, 2014.

[33] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *European Conference on Computer Vision (ECCV)*, pages 436–454. Springer, 2020.

[34] Praveen Tirupattur, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. Modeling multi-label action dependencies for temporal action localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1460–1470, 2021.

[35] Aaron Van den Oord, Yazhe Li, Oriol Vinyals, et al. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, page 4, 2018.

[36] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[38] Lezi Wang, Dong Liu, Rohit Puri, and Dimitris N Metaxas. Learning trailer moments in full-length movies with co-contrastive attention. In *European Conference on Computer Vision (ECCV)*, pages 300–316. Springer, 2020.

[39] Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. Less is more: Learning highlight detection from video duration. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[40] Minghao Xu, Hang Wang, Bingbing Ni, Riheng Zhu, Zhenbang Sun, and Changhu Wang. Cross-category video highlight detection via set-based learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7970–7979, 2021.

[41] Huan Yang, Baoyuan Wang, Stephen Lin, David P. Wipf, Minyi Guo, and Baining Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

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

[43] Qinghao Ye, Xiyue Shen, Yuan Gao, Zirui Wang, Qi Bi, Ping Li, and Guang Yang. Temporal cue guided video highlight detection with low-rank audio-visual fusion. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7950–7959, 2021.

[44] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *European Conference on Computer Vision (ECCV)*, 2016.

[45] Yingying Zhang, Junyu Gao, Xiaoshan Yang, Chang Liu, Yan Li, and Changsheng Xu. Find objects and focus on highlights: Mining object semantics for video highlight detection via graph neural networks. In *AAAI Conference on Artificial Intelligence*, pages 12902–12909, 2020.