Cross-Category Highlight Detection via Feature Decomposition and Modality Alignment

Zhenduo Zhang

Platform Technology Department, OVBU, PCG, Tencent, China
ericzdzhang@163.com

Abstract

Learning an autonomous highlight video detector with good transferability across video categories, called Cross-Category Video Highlight Detection(CC-VHD), is crucial for the practical application on video-based media platforms. To tackle this problem, we first propose a framework that treats the CC-VHD as learning category-independent highlight feature representation. Under this framework, we propose a novel module, named Multi-task Feature Decomposition Branch which jointly conducts label prediction, cyclic feature reconstruction, and adversarial feature reconstruction to decompose the video features into two independent components: highlight-related component and category-related component. Besides, we propose to align the visual and audio modalities to one aligned feature space before conducting modality fusion, which has not been considered in previous works. Finally, the extensive experimental results on three challenging public benchmarks validate the efficacy of our paradigm and the superiority over the existing state-of-the-art approaches to video highlight detection.

Introduction

Video highlight detection, which aims to automatically detect attractive moments within a video, has been a research hotspot in recent years. Video highlight detection has many downstream applications, including video summarization, video recommendation, and video editing. Despite the great successes in this field, the existing methods generally focused on training a highlight detector for a specific video category (e.g., surfing, skiing, parkour, etc.) while ignoring the transferability of a highlight detection model across different video categories (Gygli, Song, and Cao 2016; Hong et al. 2020; Badamdorj et al. 2021; Wei et al. 2022). The problem of learning a highlight detector with good transferability across video categories is called Cross-Category Video Highlight Detection(CC-VHD). The most related work in promoting the transferability of highlight detection is the work DL-VHD (Xu et al. 2021), treating the CC-VHD in an Unsupervised Domain Adaptation (UDA) (Pan and Yang 2010) way in which one adapts the knowledge learned from one labeled source domain (the source video category with supervision) to another one unlabeled target domain (the unsupervised target video category). The main issue of the proposed problem setting is that the highlight detector can be trained on only one specific category and transferred to other categories, which cannot fully utilize the annotations of the training data containing many video categories. In fact, it is a waste of resources for precious labeling data. Besides, when adopting the highlight detector in the actual scenario, it is hard and time-consuming to seek the best source categories for training.

Motivated by the weaknesses of the above approaches, we aim to overcome the poor transferability of a highlight detection model and fully utilize all the annotated training data to further promote the highlight detector's performance. The video category can be seen as the biased attribute in video highlight detection, and the highlight estimation should be bias-independent to guarantee transferability. It is feasible to remove the category bias and obtain a category-independent highlight feature representation to promote transferability. Hence, we propose a paradigm to treat the CC-VHD as a problem of learning category-independent highlight feature representation, using annotated video segments of all categories, instead of treating it in an Unsupervised Domain Adaptation (UDA) way.

To implement this paradigm, we propose a novel module named *Multi-task Feature Decomposition Branch*(MFDB), which disentangles video features into two independent components: highlight-related component and category-related component. The MFDB module jointly conducts three tasks: label prediction, cyclic feature reconstruction, and adversarial feature reconstruction, to guarantee the compactness and independence of feature decomposition.

Previous works have done many works in exploring how to fuse the information from different modalities to promote the highlight detection (Hong et al. 2020; Badamdorj et al. 2021), while they rarely take the alignment of multi-modal features into consideration. The audio encoder and visual encoder in this task are usually pretrained from different video sources, and their feature spaces may be far away from each other. ALBEF (Li et al. 2021) and CLIP2TV (Gao et al. 2021) show that aligning different modalities before fusion will promote the performance of video retrieval and other downstream tasks, such as VQA and NLVR. This motivates us to align the visual and audio modalities to one aligned

feature space before fusion in the highlight detection task.

Finally, we conduct extensive experiments on popular video highlight benchmarks to validate the effectiveness and superiority of our paradigm.

To sum up, the contributions of this work are:

- We first propose to formulate the Cross-Category Video Highlight Detection as a problem of learning category-independent highlight representation.
- We propose a novel module *Multi-task Feature Decomposition Branch* which jointly conducts label prediction, cyclic feature reconstruction, and adversarial feature reconstruction to guarantee the compactness and independence of highlight feature and category feature.
- We propose to align the visual and audio modalities to an aligned feature space before fusion to promote the performance of highlight detection.
- Extensive experiments on popular video highlight benchmarks demonstrate the effectiveness of our paradigm and the superiority over the other existing approaches.

Related Work

Video Highlight Detection

Video highlight detection is a task that assigns each video segment a highlight score. The existing works can be divided into two categories according to the way of supervision. The supervised methods (Gygli, Song, and Cao 2016; Jiao et al. 2018; Sun et al. 2016; Yao, Mei, and Rui 2016) need the highlight annotations of all segments in a video. Since the annotation for training videos is a time-consuming and laborious task, weakly-supervised approaches have played an important role in recent years. Different effective weak supervisory signals have been employed to define highlights, such as the frequent occurrence of specific segments within a video category (Panda et al. 2017a; Potapov et al. 2014a; Yang et al. 2015b), the video duration (Xiong et al. 2019) and the segment bag information (Hong et al. 2020). From the perspective of the training task, the training task of video highlight detection can be divided into two classes: classification task (Rochan et al. 2020) and ranking task(Sun et al. 2016; Garcia del Molino and Gygli 2018; Jiao et al. 2018; Gygli, Song, and Cao 2016; Wang et al. 2020). For the classification task, the network tries to classify video clips according to whether they are highlights or not. It is also popular to adopt a ranking training task, where we train the ranking network to rank highlight clips higher than non-highlight clips. The transferability of highlight detection has attracted the attention of researchers recently. The most related work in promoting the transferability of highlight detection is the work (Xu et al. 2021), treating the task in an Unsupervised Domain Adaptation (UDA) way and adapting the knowledge learned from the source video category to the target video category.

Feature Decomposition

Feature Decomposition (Zhang et al. 2019; Bao et al. 2018; He et al. 2019; Singh, Ojha, and Lee 2019; Tran, Yin, and Liu 2017) models the explanatory factors from diverse data variation, which has drawn considerable attention in many fields, such as face recognition, person re-identification, and generative tasks. Previous works utilized annotated attribute data to decompose representations into identity-related and identity-independent information (pose, viewpoint, illumination, etc.) for recognition and identification tasks. (He et al. 2019) learned invariant feature representations of heterogeneous face images by minimizing Wasserstein distance between cross-modality distributions. (Hou, Li, and Wang 2021) decomposed the feature representation into age-related feature and identity-related feature by minimizing the mutual information between two components.

Approach

Figure 1 illustrates the overall framework of our proposed approach. The framework inputs are the sampled frames from the video and the audio signal. We denote the training batch size is $\mathcal{B}$ and the number of sampled frames per video is N_f. The frameset of the k_{th} video is denoted as I_k. We feed the I_k to the video encoder, a vision transformer (Vaswani et al. 2017; Dosovitskiy et al. 2020), and obtain the feature of N_f frames. Then the N_f frame features are processed by the temporal transformer to capture the relationship among the sampled frames and we get the N_f frame embeddings $\{v_k^i\}_{i=1}^{N_f}$. We note that the parameter set of the overall visual feature extractor is θ_V.

We feed the audio signal of the k_{th} video to the audio encoder and a projection MLP to obtain the audio embeddings and project them to have the same dimension as the visual embeddings. Since the audio signals have different lengths, the number of audio embeddings differs within the batch. To perform batch training, we need to ensure that the number of audio embeddings is the same for each video. Therefore, we first determine the minimum number N_a of audio embeddings for each video in the current batch. Then we uniformly sample the audio embeddings for each video to obtain N_a audio embeddings. The final audio embeddings of the k_{th} video is denoted as $\{a_k^i\}_{i=1}^{N_a}$. We note that the parameter set of the overall audio feature extractor is θ_A.

We attempt to fuse the visual and audio modalities via the Merged Attention Model (Dou et al. 2022), where the visual and audio features are simply concatenated together, then fed into a single transformer (Vaswani et al. 2017) block. We note that the parameter set of the Merged Attention Model is θ_M.

The Multi-task Feature Decomposition Branch(MFDB) can decompose a feature z_n into two independent components: the category-related feature c_n and the highlight-related feature h_n. We employ two MFDBs on top of the visual encoder and the Merged Attention Model, and the two MFDBs do not share parameters.

The Visual-Audio Alignment procedure aligns the visual and audio modalities to an aligned feature space before fusion, which is implemented via a contrastive learning paradigm.

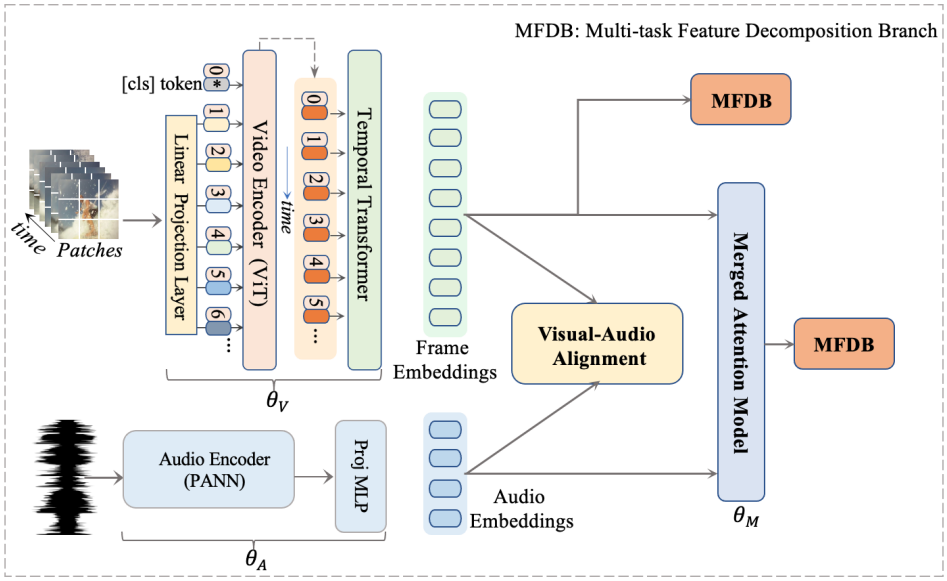


Figure 1: An overview of the proposed approach.

Multi-task Feature Decomposition Branch

The core of our approach to tackle the Cross-Category Highlight Detection problem is to decompose the original feature of the n_{th} video, z_n, into two independent components: the category-related feature c_n and the highlight-related feature h_n. When inferring, we only utilize the highlight-related feature h_n to obtain good transferability across video categories. We denote the feature decomposition module as $\Pi = (\Pi_c, \Pi_h)$, where Π_c and Π_h are modeled by two-layer MLPs. Given an original feature z_n, Π decomposes z_n into two components c_n and h_n as $(c_n, h_n) = \Pi(z_n) = (\Pi_c(z_n), \Pi_h(z_n))$.

We consider that the original feature z_n is properly decomposed by Π if the following four constraints are satisfied:

$\mathcal{C}_1$: c_n and h_n can reconstruct the original feature z_n via proper function Π' to guarantee the compactness of decomposition.

$\mathcal{C}_2$: We suppose that the highlight feature from the m_{th} video($m \neq n$) is h_m. The n_{th} video and m_{th} video have the same highlight label $y_m^h = y_n^h$. Then c_n and h_m can reconstruct the original feature z_n via Π' to guarantee the highlight invariance to category.

$\mathcal{C}_3$: We cannot neither reconstruct c_n from h_n via a function τ_c nor reconstruct h_n from c_n via a function τ_h, which can guarantee the independence of the two components.

$\mathcal{C}_4$: c_n can predict the category label and h_n can predict the highlight label to guarantee c_n and h_n encode the category information and highlight information,respectively.

NOTE: *We assume that the parameter set of the feature extractor that generates z_n is θ_z. If the MFDB is employed in the visual branch, θ_z is the parameter set of visual encoder $\theta_z = \theta_V$. If the MFDB is employed in the fusion branch, θ_z is the parameter set of visual encoder, audio encoder and the Merged Attention Model, $\theta_z = \{\theta_V, \theta_A, \theta_M\}$.*

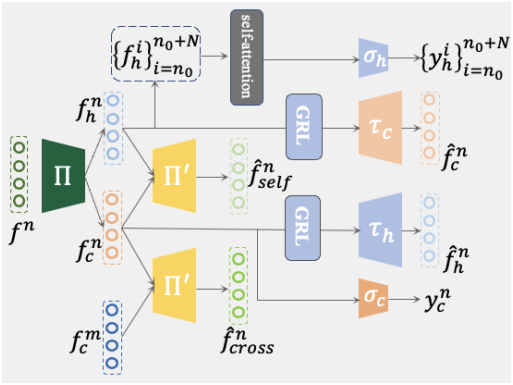


Figure 2: An overview of the proposed Multi-task Feature Decomposition Branch.

We implement the above 4 constraints via four optimization objectives, where $\mathbb{E}[\cdot]$ calculates the mean loss within the batch.

To implement $\mathcal{C}_1$, we propose to employ a cyclic reconstruction task as Equation 1. Π' is modeled via a two-layer MLP and its input is the concatenation of c_n and h_n.

$$\begin{aligned}\min_{\theta_z,\theta_\Pi,\theta_{\Pi'}} \mathcal{L}_{reco}^{self} &= \mathbb{E}\left[\|z_n - \hat{z}_n^{self}\|_2\right] \\ &= \mathbb{E}\left[\|z_n - \Pi'([c_n|h_n])\|_2\right]\end{aligned} \tag{1}$$

To implement $\mathcal{C}_2$, we propose to employ an cyclic reconstruction task as Equation 2, where $m \neq n$ and $y_m^h = y_n^h$.

$$\begin{aligned}\min_{\theta_z,\theta_\Pi,\theta_{\Pi'}} \mathcal{L}_{reco}^{cross} &= \mathbb{E}\left[\|z_n - \hat{z}_n^{cross}\|_2\right] \\ &= \mathbb{E}\left[\|z_n - \Pi'([c_n|h_m])\|_2\right]\end{aligned} \tag{2}$$

To implement $\mathcal{C}_3$, we propose to employ an adversial reconstruction task as Equation 3. τ_c and τ_h are both modeled

by two-layer MLPs.

$$\begin{aligned}\min_{\theta_{\tau_c},\theta_{\tau_h}}\max_{\theta_z,\theta_\Pi}\mathcal{L}_{reco}^{adv} &= \mathbb{E}\left[\|c_n-\hat{c}_n\|_2+\|h_n-\hat{h}_n\|_2\right]\\ &= \mathbb{E}\left[\|c_n-\tau_c(h_n)\|_2+\|h_n-\tau_h(c_n)\|_2\right]\end{aligned} \tag{3}$$

At training time, we seek the parameters $\{\theta_z,\theta_\Pi\}$ to maximize the adversarial reconstruction loss $\mathcal{L}_{reco}^{adv}$ to make c_n and h_n independent to each other, while simultaneously seeking $\{\theta_{\tau_c},\theta_{\tau_h}\}$ to minimize $\mathcal{L}_{reco}^{adv}$ to reconstruct one component from another one. In order to optimize Equation 3, we exploit Gradient Reversal Layer(GRL) (Ganin et al. 2016) to connect τ_c and τ_h to the whole architecture. The GRL module behaves as the identity function during the forward pass and inverts the gradient sign during the backward pass, pushing the parameters to maximize the output loss. If we denote the GRLs of the category feature and highlight feature as g_c and g_h, the optimization objective of Equation 3 can be rewritten as Equation 4:

$$\begin{aligned}\min_{\theta_z,\theta_\Pi,\theta_{\tau_c},\theta_{\tau_h}}\mathcal{L}_{reco}^{adv} =&\mathbb{E}\left[\|c_n-\tau_c(g_h(h_n))\|_2\right]+\\ &\mathbb{E}\left[\|h_n-\tau_h(g_c(c_n))\|_2\right]\end{aligned} \tag{4}$$

To implement $\mathcal{C}_4$, we try to predict the category label using c_n and the highlight label using h_n. σ_c and σ_h denote the corresponding category classifier and the highlight classifier. σ_c is a multi-class classifier, and the classification loss for the category is calculated as Equation 5.

$$\min_{\theta_z,\theta_\Pi,\theta_{\sigma_c}}\mathcal{L}_{cls}^{c} =\mathbb{E}\left[-y_n^c\log\hat{y}_n^c\right]=\mathbb{E}\left[-y_n^c\log\sigma_c(c_n)\right] \tag{5}$$

σ_h is a binary-class classifier, and the output of it can either represent the highlight score in the ranking way or the probability of highlight class in the classification way. In order to better capture the highlight relationship between different video clips, we adopt an additional self-attention module(SA) (Badamdorj et al. 2021) when predicting the highlight label. The input to the SA module is a set of highlight features from different videos, i.e. $\{h_{n_0},...,h_{n_0+N_G}\}$ and the output highlight logits are $\left\{\hat{y}_{n_0}^h,...,\hat{y}_{n_0+N_G}^h\right\}=\sigma_h\left(\mathrm{SA}\left(\{h_{n_0},...,h_{n_0+N_G}\}\right)\right)$. N_G is the length of highlight feature set. The category classification loss is Equation 6.

$$\min_{\theta_z,\theta_\Pi,\theta_{\sigma_h}}\mathcal{L}_{cls}^{h} =\mathbb{E}\left[-y_n^h\log\hat{y}_n^h\right] \tag{6}$$

The overall loss of MFDB module is Equation 7.

$$\mathcal{L}_{MFDB}=\mathcal{L}_{reco}^{self}+\mathcal{L}_{reco}^{cross}+\mathcal{L}_{reco}^{adv}+\mathcal{L}_{cls}^{c}+\mathcal{L}_{cls}^{h} \tag{7}$$

Visual-Audio Alignment

The audio and visual encoders are usually pretrained from different sources in practice. For instance, the audio encoder PANN (Kong et al. 2020) is pretrained on AudioSet (Gemmeke et al. 2017) and the visual encoder ViT (Dosovitskiy et al. 2020) is pretrained on Kinect-400 (Carreira and Zisserman 2017) dataset. The gap in the source domain leads to the fact that their feature spaces may be far from each other. The success in ALBEF (Li et al. 2021), CLIP2TV (Gao et al. 2021) motivates us to align the visual and audio modalities to an aligned feature space before fusion in the highlight detection task. We average the visual and audio embeddings to get the visual and audio representation of the video. $v_k=\frac{1}{N_f}\sum_{i=1}^{N_f}v_k^i$ and $a_k=\frac{1}{N_a}\sum_{i=1}^{N_a}a_k^i$. The Visual-Audio Alignment loss is calculated by Equation 8, where τ is the learnable temperature widely used in contrastive learning.

$$\mathcal{L}_{Align}=\frac{1}{2}\mathbb{E}\left[\frac{e^{v_i\cdot a_i/\tau}}{\sum_{j\sim\mathcal{B}}e^{v_i\cdot a_j/\tau}}+\frac{e^{a_i\cdot v_i/\tau}}{\sum_{j\sim\mathcal{B}}e^{a_i\cdot v_j/\tau}}\right] \tag{8}$$

Overall Loss Function

We add two different MFDB modules on the visual branch and the fusion branch, and the losses of the two MFDB modules are denoted as $\mathcal{L}_{MFDB}^{v}$ and $\mathcal{L}_{MFDB}^{f}$. Hence, the loss of the framework is shown in Equation 9, where $\{\lambda_i\}_{i=1}^3$ are the weights of the losses.

$$\mathcal{L}=\lambda_1\mathcal{L}_{MFDB}^{f}+\lambda_2\mathcal{L}_{MFDB}^{v}+\lambda_3\mathcal{L}_{Align} \tag{9}$$

Experiments

Experiment Settings and Compared Methods

We evaluate our approaches on three popular benchmark datasets, i.e., *YouTube Highlights* (Sun et al. 2016), *TVSum* (Song et al. 2015) and *CoSum* (Chu, Song, and Jaimes 2015), for video highlight detection. *YouTube Highlights* contains six event-specific categories, and there are approximately 100 videos in each event.*TVSum* is an available video summarization benchmark dataset, collected from YouTube and crawled by an event-specific queried tag. *TVSum* consists of 50 videos grouped into ten categories (5 videos per category). *CoSum* contains 51 videos covering 10 events.

We use the ViT-32 (Dosovitskiy et al. 2020) pretrained on the Kinect-400 (Carreira and Zisserman 2017) as the video encoder to extract visual features of the sampled frames and use the PANN (Kong et al. 2020) pretrained on the AudioSet (Gemmeke et al. 2017) to extract the audio embeddings of audio clips. We sample 16 frames uniformly from the video frames of a video. We train our model using Adam, with a learning rate of 1×10^{-4}. The weights of the losses in Equation 9 are $\{\lambda_i\}_{i=1}^3=\{1.0,0.5,1.0\}$, which are selected by GridSearch strategy. The size of the highlight feature set fed into the self-attention layer in the MFDB module, which is the N_G mentioned above, is set to 16. The source code will be released.

We compare our methods with the following state-of-the-art video highlight detection baselines on three datasets. We introduce six weakly-supervised approaches, i.e. RRAE(V) (Yang et al. 2015a), SG(V) (Mahasseni, Lam, and Todorovic 2017), DSN(V) (Panda et al. 2017b), VESD(V) (Cai et al. 2018), LIM-s(V) (Xiong et al. 2019) and MINI-Net(V) (Hong et al. 2020) for comparison. Besides, nine supervised video highlight detection methods, i.e. Video2GIF(V) (Gygli, Song, and Cao 2016), LSVM(V) (Sun et al. 2016), KVS(V) (Potapov et al. 2014b), DPP(V) (Gong et al. 2014), sLSTM(V) (Zhang

et al. 2016), SM(V) (Gygli, Grabner, and Van Gool 2015), JVAL(VA) (Badamdorj et al. 2021), DL(V) (Xu et al. 2021) and PLD(V) (Wei et al. 2022) are also involved for comparison. *The "V" and "VA" in parentheses after each method indicate whether the corresponding method is based on visual modality or visual-audio multi-modality.*

Comparisons with the State-of-the-art Methods

We compare our approach with the current state-of-the-art methods on three popular benchmarks, *Youtube Highlights*, *TVSum* and *CoSum*, which are shown in Table 1, Table 2 and Table 3. Firstly, we compare our overall framework that utilizes both visual and audio modalities with the current state-of-the-art methods. On the Youtube Highlights Dataset, our overall framework gains the best performance in four of six categories. The average mAP in all the categories increases from 0.730 in PLD to 0.749. On the TVSum Dataset, our approach obtains the best performance on seven of ten categories, and the average top-5 mAP raises from 0.771 in PLD to 0.815. Similarly, our model surpasses or equal the state-of-the-art methods on seven of ten categories on the CoSum Dataset. The average top-5 mAP score is also improved, from 0.946 in PLD to 0.961. Although our model does not gain the best performance in some specific categories, our model can be employed to detect highlights for all video classes instead of training one expert model for each category, which contributes to the good transferability across video categories.

Since some of the SOTA methods for comparison, such as DL (Xu et al. 2021) and PLD (Wei et al. 2022), only use the visual modality rather than both visual and audio modalities, we also utilize a single visual modality to make fair comparisons with the current methods. When only using visual modality, the Visual-Audio Alignment, the Merged Attention Model, and the MFDB module for fused features are not included in our framework, and we only use the visual encoder and the MFDB module for the visual feature. From Table 1, our model uses a single visual modality and gain 0.737 average mAP over all categories, which also outperforms the other current methods significantly. Likewise, our model using a single visual modality can also exceed the other state-of-the-art methods on the average top-5 mAP on TVSum and CoSum datasets. The average top-5 mAP on TVSum and CoSum are 0.806 and 0.954, respectively.

The comparison between our model using a single visual modality and the one using both modalities demonstrates that the audio modality can help improving the highlight detection performance on the three datasets.

Ablation Studies of MFDB Module

Necessity of MFDB The Multi-task Feature Decomposition Branch aims to decompose the original feature into two nearly independent components: category-related and highlight-related features. Next, we will verify the necessity of using the Multi-task Feature Decomposition Branch. For a fair comparison, we replace the MFDB module with a simple Multi-Task Learning Branch(MTLB), which jointly predicts category labels and highlight labels and cannot conduct feature decomposition. The highlight features are also processed by a self-attention module to capture the relationship among clips. The details of the multi-task learning head are shown in Figure 3. It is proper to use the MTLB, which only lacks the feature decomposition process compared with the MFDB. Besides, multi-task learning is a typical transfer learning approach for learning better representation.

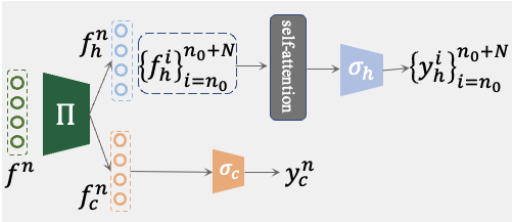
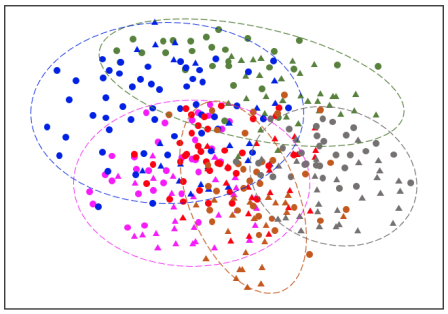
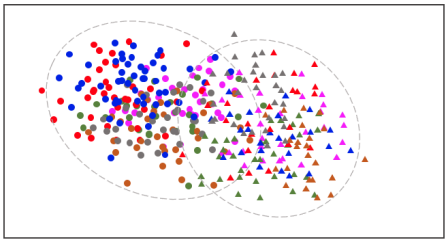


Figure 3: Details of the simple Multi-Task Learning Branch.



(a) Youtube Highlights-MLTB



(b) Youtube Highlights-MFDB

Figure 4: Highlight feature distribution of the randomly sampled video clips in the Youtube Highlights Dataset. Different colors represent different categories, and there are six categories in Youtube Highlights. The solid circles represent the highlight clips, and the solid triangles represent the non-highlight clips. The circles with dashed boundaries in Figure 4(a) represent the highlight feature set of each category and the circles with dashed boundaries in Figure 4(b) represent highlight feature set of different highlight classes.

The performance comparison between the MFDB and the MTLB is shown in Table 4. When using the MTLB directly, we only jointly conduct the category classification and highlight classification task without disentangling the category and highlight features from each other. When using the MFDB, we add a constraint on the independence of the category-related and the highlight-related components. Ideally, we assume that if the highlight feature is independent

Topic	Weakly supervised			Supervised						
	RRAE (V)	LIM-s (V)	MINI-Net (VA)	Video2GIF (V)	LSVM (V)	JVAL (VA)	DL (V)	PLD (V)	**Ours** (V)	**Ours** (VA)
dog	0.49	0.579	0.582	0.308	0.60	0.645	0.708	**0.749**	0.710	0.716
gymnastics	0.35	0.417	0.617	0.335	0.41	0.719	0.532	0.702	0.717	**0.728**
parkour	0.50	0.670	0.702	0.540	0.61	0.808	0.772	0.779	0.802	**0.817**
skating	0.25	0.578	0.722	0.554	0.62	0.620	0.725	0.575	0.714	**0.726**
skiing	0.22	0.486	0.587	0.328	0.36	0.732	0.661	0.707	0.725	**0.738**
surfing	0.49	0.651	0.651	0.541	0.61	0.783	0.762	**0.790**	0.757	0.769
Average	0.383	0.564	0.644	0.464	0.536	0.718	0.693	0.730	0.737	**0.749**

Table 1: Highlight detection results (mAP) of weakly-supervised and supervised methods on the YouTube Highlights dataset.

Topic	Weakly supervised/Un Methods					Supervised Methods						
	SG (V)	LIM-s (V)	DSN (V)	VESD (V)	MINI-Net (VA)	sLSTM (V)	SM (V)	JVAL (VA)	DL (V)	PLD (V)	**Ours** (V)	**Ours** (VA)
VT	0.423	0.559	0.373	0.447	0.806	0.411	0.415	0.837	0.865	0.845	0.858	**0.868**
VU	0.472	0.429	0.441	0.493	0.683	0.462	0.467	0.573	0.687	0.809	0.805	**0.814**
GA	0.475	0.612	0.428	0.496	0.782	0.463	0.469	0.785	0.749	0.703	0.783	**0.787**
MS	0.489	0.540	0.436	0.503	0.818	0.477	0.478	0.861	0.862	0.725	0.859	**0.869**
PK	0.456	0.604	0.411	0.478	0.781	0.448	0.445	0.801	0.790	0.764	0.795	**0.803**
PR	0.473	0.475	0.417	0.485	0.658	0.461	0.458	0.692	0.632	**0.872**	0.864	0.867
FM	0.464	0.432	0.412	0.487	0.578	0.452	0.451	0.700	0.589	0.719	0.720	**0.730**
BK	0.417	0.663	0.368	0.441	0.751	0.406	0.407	0.730	0.726	0.740	0.729	**0.742**
BT	0.483	0.691	0.435	0.492	0.802	0.471	0.473	**0.974**	0.789	0.744	0.895	0.911
DS	0.466	0.626	0.416	0.488	0.655	0.455	0.453	0.675	0.640	**0.791**	0.754	0.762
Average	0.462	0.563	0.424	0.481	0.732	0.451	0.461	0.762	0.733	0.771	0.806	**0.815**

Table 2: Experimental results (top-5 mAP score) the TVSum dataset.

Topic	Weakly supervised				Supervised Methods						
	SG (V)	VESD (V)	DSN (V)	MINI-Net (VA)	KVS (V)	DPP (V)	sLSTM (V)	SM (V)	PLD (V)	**Ours** (V)	**Ours** (VA)
BJ	0.698	0.685	0.715	0.845	0.662	0.672	0.683	0.692	0.900	0.914	**0.925**
BP	0.713	0.714	0.746	0.988	0.674	0.682	0.701	0.722	0.970	0.988	**0.994**
ET	0.759	0.783	0.813	0.915	0.731	0.744	0.749	0.789	0.817	0.910	**0.917**
ERC	0.729	0.721	0.756	**1.000**	0.685	0.694	0.717	0.728	**1.000**	**1.000**	**1.000**
KP	0.729	0.742	0.772	0.961	0.701	0.705	0.714	0.745	**1.000**	0.972	0.986
MLB	0.721	0.687	0.727	0.935	0.668	0.677	0.714	0.693	**1.000**	**1.000**	**1.000**
NFL	0.693	0.724	0.737	**1.000**	0.671	0.681	0.681	0.727	0.970	**1.000**	**1.000**
NDC	0.738	0.751	0.782	0.953	0.698	0.704	0.722	0.759	0.958	0.959	**0.966**
SL	0.743	0.763	0.794	**0.889**	0.713	0.722	0.721	0.766	0.844	0.873	0.878
SF	0.681	0.674	0.709	0.789	0.642	0.648	0.653	0.683	**1.000**	0.929	0.940
Average	0.720	0.721	0.755	0.927	0.684	0.692	0.705	0.735	0.946	0.954	**0.961**

Table 3: Experimental results (top-5 mAP score) the CoSum dataset.

of the category, the highlight feature will have better transferability across categories, and the overall performance in all categories will be improved. From Table 4, the average highlight detection performance overall categories on three benchmarks can be significantly promoted when utilizing either a single visual modality or visual-audio modalities. The results sufficiently prove our assumption that learning the category-independent highlight feature can help improve the overall performance over different categories.

Except for the quantitive experiments to compare the performance of the highlight detector, we also conduct a qualitative comparison between MFDB and MTLB. We random sample 40 to 60 video clips from each category in Youtube Highlights Dataset. The t-SNE visualization results of their highlight features is shown in Figure 4.In Figure 4(a), the highlight features are still entangled with the category though they have a certain level of highlight discrimination. Within each category, the highlight clip features and the non-highlight clip features roughly distributes in two separate feature spaces. However, some highlight clip fea-

Figure 5: Visualization for Highlight Detection

Dataset	Visual		Visual-Audio	
	MTLB	MFDB	MTLB	MFDB
Youtube Highlights	0.623	**0.737** ↑	0.652	**0.749** ↑
TVSum	0.646	**0.806** ↑	0.673	**0.815** ↑
CoSum	0.865	**0.954** ↑	0.874	**0.961** ↑

Table 4: Ablation study of the necessity of MFDB. The average mAP is used in the YouTube Highlights, and the average top-5 mAP is used in the TVSum and CoSum.

tures of one category can also be close to the non-highlight clip features of another category. This is because the highlight features are not independent of the category if we have no added independence constraints to the highlight features and category features. In Figure 4(b), the distribution of the highlight clip features and the non-highlight clip features are split in the feature space, and the features with the same highlight label are clustered more closely despite the category differences. This result shows improved independence between the highlight features and the category information. In this way, the highlight discrimination and the transferability across categories can be improved. The qualitative comparison also proves the necessity of the MFDB module.

Dataset	Fusion Only	Fusion & Visual
Youtube Highlights	0.742	**0.749** ↑
TVSum	0.808	**0.815** ↑
CoSum	0.957	**0.961** ↑

Table 5: Ablation study of where to apply MFDB. The average mAP is used in the YouTube Highlights, and the average top-5 mAP is used in the TVSum and CoSum.

Where to Use MFDB In our framework, the MFDB module can be applied to fused and visual features. In our opinion, if the MFDB is utilized on both visual feature and fused feature hierarchically, the visual encoder may generate a better representation for the category and highlight classification than MFDB only on the fused feature. We conduct experiments to analyze and verify this assumption. As is demonstrated in Table 5, the highlight detection performance can be improved on all three benchmarks when adding the MFDB module to the visual encoder.

Ablation Study of Visual-Audio Alignment

Dataset	w/o $\mathcal{L}_{Align}$	with $\mathcal{L}_{Align}$
Youtube Highlights	0.740	**0.749** ↑
TVSum	0.811	**0.815** ↑
CoSum	0.958	**0.961** ↑

Table 6: Ablation study of the Visual-Audio Alignment. The average mAP is used in the YouTube Highlights, and the average top-5 mAP is used in the TVSum and CoSum.

We conducted an ablation study to explore the effect of the Visual-Audio Alignment process, and the experiments are illustrated in Table 6. From this table, we can verify that aligning the visual and audio features before fusing them can promote the learning of highlight detection. Instead of fusing the features from two different modalities and source domain, the visual-audio alignment help to pull the feature spaces of visual feature and audio feature to be close to each other first.

Conclusion

In this work, we first attempt to learn the highlight video detector with good transferability via learning the category-independent highlight representation, which can make full use of the annotated video segments of all categories. To implement this idea, we propose the Multi-task Feature Decomposition Branch, which disentangles the features into highlight and category features, which are nearly independent. Besides, we propose to pull the feature spaces of different modalities to be close before fusion, which benefits the multimodal representation learning of the video.

References

Badamdorj, T.; Rochan, M.; Wang, Y.; and Cheng, L. 2021. Joint Visual and Audio Learning for Video Highlight Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 8107–8117.

Bao, J.; Chen, D.; Wen, F.; Li, H.; and Hua, G. 2018. Towards Open-Set Identity Preserving Face Synthesis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6713–6722.

Cai, S.; Zuo, W.; Davis, L. S.; and Zhang, L. 2018. Weakly-Supervised Video Summarization Using Variational Encoder-Decoder and Web Prior. In *ECCV*.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4724–4733.

Chu, W.-S.; Song, Y.; and Jaimes, A. 2015. Video co-summarization: Video summarization by visual co-occurrence. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3584–3592.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR*, abs/2010.11929.

Dou, Z.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Zhang, P.; Yuan, L.; Peng, N.; Liu, Z.; and Zeng, M. 2022. An Empirical Study of Training End-to-End Vision-and-Language Transformers. In *2022 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.*, 17(1): 2096–2030.

Gao, Z.; Liu, J.; Chen, S.; Chang, D.; Zhang, H.; and Yuan, J. 2021. CLIP2TV: An Empirical Study on Transformer-based Methods for Video-Text Retrieval. *arxiv*, abs/2111.05610.

Garcia del Molino, A.; and Gygli, M. 2018. PHD-GIFs: Personalized Highlight Detection for Automatic GIF Creation. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, 600–608. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356657.

Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780.

Gong, B.; Chao, W.-L.; Grauman, K.; and Sha, F. 2014. Diverse Sequential Subset Selection for Supervised Video Summarization. In *Advances in Neural Information Processing Systems*, volume 27.

Gygli, M.; Grabner, H.; and Van Gool, L. 2015. Video summarization by learning submodular mixtures of objectives. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3090–3098.

Gygli, M.; Song, Y.; and Cao, L. 2016. Video2GIF: Automatic Generation of Animated GIFs from Video. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1001–1009.

He, R.; Wu, X.; Sun, Z.; and Tan, T. 2019. Wasserstein CNN: Learning Invariant Features for NIR-VIS Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7): 1761–1773.

Hong, F.-T.; Huang, X.; Li, W.-H.; and Zheng, W.-S. 2020. MINI-Net: Multiple Instance Ranking Network for Video Highlight Detection. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII*, 345–360. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-58600-3.

Hou, X.; Li, Y.; and Wang, S. 2021. Disentangled Representation for Age-Invariant Face Recognition: A Mutual Information Minimization Perspective. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3672–3681.

Jiao, Y.; Li, Z.; Huang, S.; Yang, X.; Liu, B.; and Zhang, T. 2018. Three-Dimensional Attention-Based Deep Ranking Model for Video Highlight Detection. *IEEE Transactions on Multimedia*, 20(10): 2693–2705.

Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; and Plumbley, M. D. 2020. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2880–2894.

Li, J.; Selvaraju, R. R.; Gotmare, A. D.; Joty, S. R.; Xiong, C.; and Hoi, S. C. H. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Advances in Neural Information Processing Systems(NIPS)*.

Mahasseni, B.; Lam, M.; and Todorovic, S. 2017. Unsupervised Video Summarization with Adversarial LSTM Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2982–2991.

Pan, S. J.; and Yang, Q. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10): 1345–1359.

Panda, R.; Das, A.; Wu, Z.; Ernst, J.; and Roy-Chowdhury, A. K. 2017a. Weakly Supervised Summarization of Web Videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 3677–3686.

Panda, R.; Das, A.; Wu, Z.; Ernst, J.; and Roy-Chowdhury, A. K. 2017b. Weakly Supervised Summarization of Web Videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 3677–3686.

Potapov, D.; Douze, M.; Harchaoui, Z.; and Schmid, C. 2014a. Category-Specific Video Summarization. In *ECCV*.

Potapov, D.; Douze, M.; Harchaoui, Z.; and Schmid, C. 2014b. Category-specific video summarization. In *ECCV*.

Rochan, M.; Krishna Reddy, M. K.; Ye, L.; and Wang, Y. 2020. Adaptive Video Highlight Detection by Learning from

User History. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI*, 261–278. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-58588-4.

Singh, K. K.; Ojha, U.; and Lee, Y. J. 2019. FineGAN: Unsupervised Hierarchical Disentanglement for Fine-Grained Object Generation and Discovery. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6483–6492.

Song, Y.; Vallmitjana, J.; Stent, A.; and Jaimes, A. 2015. TVSum: Summarizing web videos using titles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5179–5187.

Sun, M.; Farhadi, A.; Chen, T.-H.; and Seitz, S. 2016. Ranking Highlights in Personal Videos by Analyzing Edited Videos. *IEEE Transactions on Image Processing*, 25(11): 5145–5157.

Tran, L.; Yin, X.; and Liu, X. 2017. Disentangled Representation Learning GAN for Pose-Invariant Face Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1283–1292.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems(NIPS)*.

Wang, L.; Liu, D.; Puri, R.; and Metaxas, D. N. 2020. Learning Trailer Moments in Full-Length Movies with Co-Contrastive Attention. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII*, 300–316. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-58522-8.

Wei, F.; Wang, B.; Ge, T.; Jiang, Y.; Li, W.; and Duan, L. 2022. Learning Pixel-Level Distinctions for Video Highlight Detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiong, B.; Kalantidis, Y.; Ghadiyaram, D.; and Grauman, K. 2019. Less Is More: Learning Highlight Detection From Video Duration. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1258–1267.

Xu, M.; Wang, H.; Ni, B.; Zhu, R.; Sun, Z.; and Wang, C. 2021. Cross-category Video Highlight Detection via Set-based Learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7950–7959.

Yang, H.; Wang, B.; Lin, S.; Wipf, D.; Guo, M.; and Guo, B. 2015a. Unsupervised Extraction of Video Highlights via Robust Recurrent Auto-Encoders. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 4633–4641.

Yang, H.; Wang, B.; Lin, S.; Wipf, D. P.; Guo, M.; and Guo, B. 2015b. Unsupervised Extraction of Video Highlights via Robust Recurrent Auto-Encoders. *2015 IEEE International Conference on Computer Vision (ICCV)*, 4633–4641.

Yao, T.; Mei, T.; and Rui, Y. 2016. Highlight Detection with Pairwise Deep Ranking for First-Person Video Summarization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 982–990.

Zhang, K.; Chao, W.-L.; Sha, F.; and Grauman, K. 2016. Video Summarization with Long Short-Term Memory. In *ECCV*, volume 9911, 766–782.

Zhang, Z.; Tran, L.; Yin, X.; Atoum, Y.; Liu, X.; Wan, J.; and Wang, N. 2019. Gait Recognition via Disentangled Representation Learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4705–4714.