Probing Visual-Audio Representation for Video Highlight Detection via HPCL

Shuaicheng Li, Feng Zhang*, Kunlin Yang, Lingbo Liu, Shinan Liu, Jun Hou, Shuai Yi

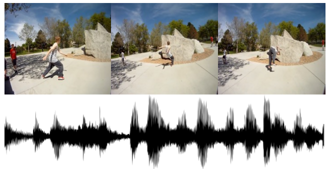
商汤 sensetime

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

Whether the multimodal signal is synchronized ?

We perform highlight prediction by judging the synchronization relationship between multimodal signals.

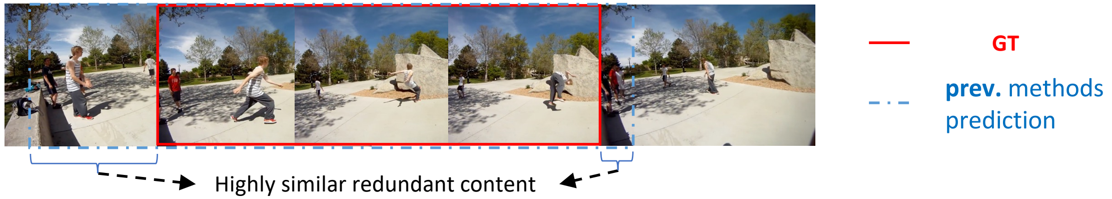
(Synced actual audio)



(Post editing BGM)

we explore the dependencies between within-modality features and exclude the unrelated clues to facilitate the specialized characteristic of inter-segment alignment.

Difficulty locating highlights accurately due to video frame redundancy



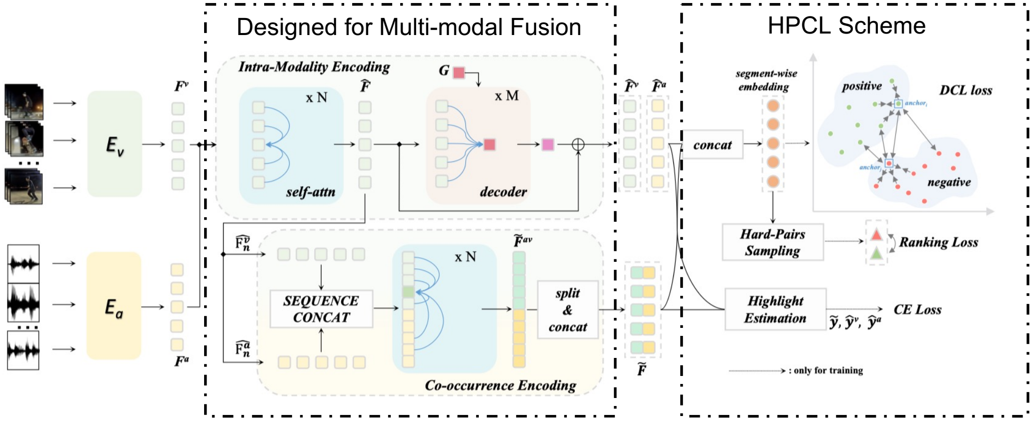
We use the following two schemes,

1） hard-pairs mining,

2） hard-pairs guided contrastive learning scheme

to achieve more accurate predictions.

Architecture of our HPCL



$$score = \lambda_1 \tilde{y} + \lambda_2 \hat{y}^a + \lambda_3 \hat{y}^v$$

Experiments

Highlight detection on YouTube Highlight dataset in terms of mAP.

Category	Uni-Modality					Multi-Modality				
	RRAE [41]	LIM-s [39]	Video2GIF [13]	LSVM [32]	SL[40]	MN[17]	Joint-VA [1]	TCG [43]	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**

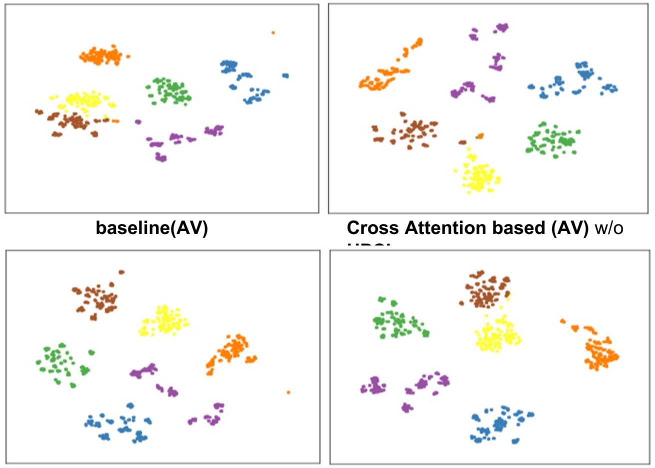
Results are also reported on TVSum dataset

Importance of our feature encoding module and HPCL scheme

Architecture Variants	YouTube Highlight	TVSum
V Only	0.659	0.763
A Only	0.651	0.752
AV	0.675	0.784
Cross Attention based (**AV**)	0.697	0.789
Ours (**AV**)	**0.747**	**0.801**

Learning Scheme	YouTube Highlight	TVSum
CE loss (baseline)	0.702	0.766
CE loss + HPCL	0.733	0.792
CE loss + HPCL + rank loss	**0.747**	**0.801**

Visual qualitative analysis



baseline(AV)

Cross Attention based (AV) w/o

Cross Attention based (AV) w/ **HPCL**

Ours