VVS: Video-to-Video Retrieval with Irrelevant Frame Suppression

[1]Won Jo, [1]Geuntaek Lim, [1]Gwangjin Lee, [1]Hyunwoo Kim, [2]Byungsoo Ko, [1]Yukyung Choi
[1]Sejong University [2]Naver Vision

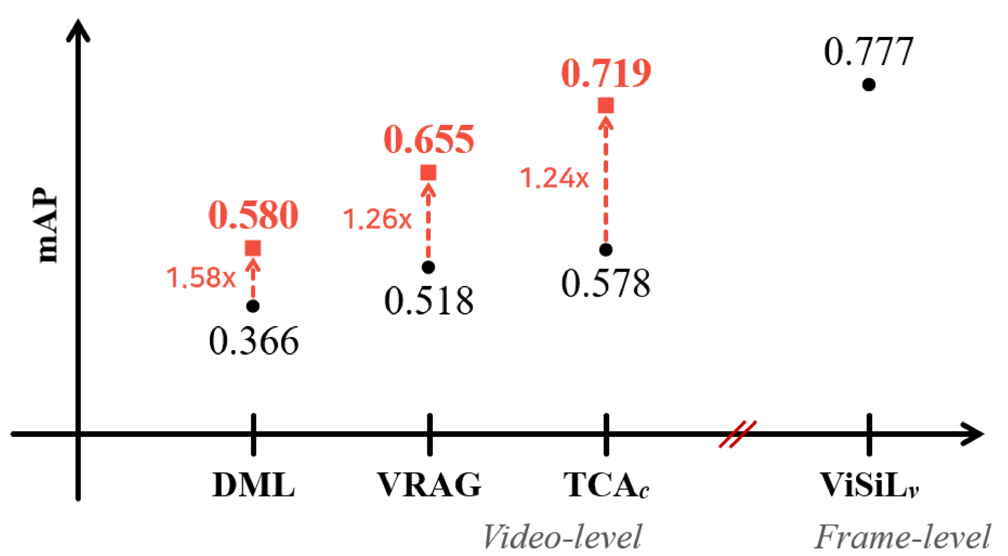
Key Questions

- Q1: Wouldn't achieving an accurate and efficient search be possible by mitigating information loss in the video-level approach?
- Q2: Should irrelevant frames (red boxes) be suppressed in the untrimmed videos?

Key Insights

If irrelevant frames are ideally suppressed, the video-level approaches can be more accurate (even comparable to the frame-level SOTA).

Temporal annotation (Jo *et al.*, IEEE Access, 2023)



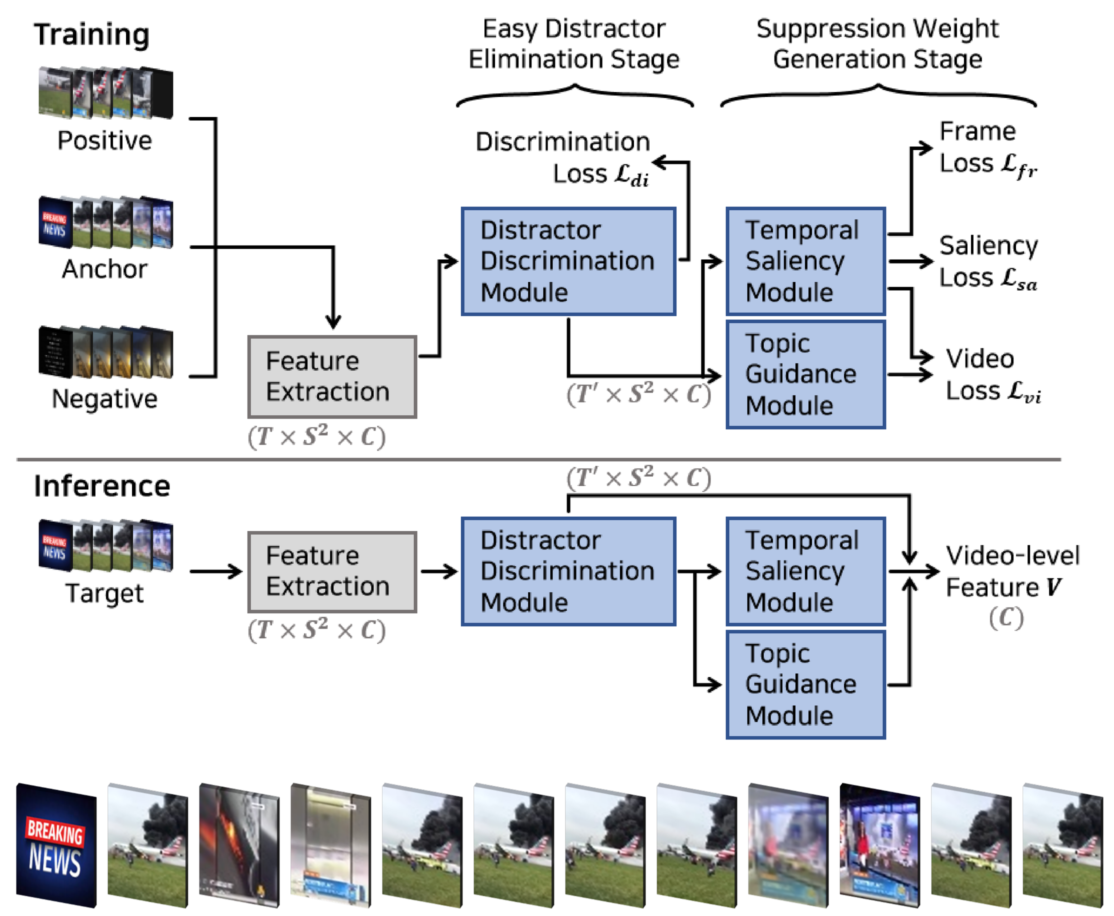
Contact Information

Paper
Project
Demo

Method

Proposed Framework

Understand irrelevant frames for describing a distinct video-level feature in an untrimmed video

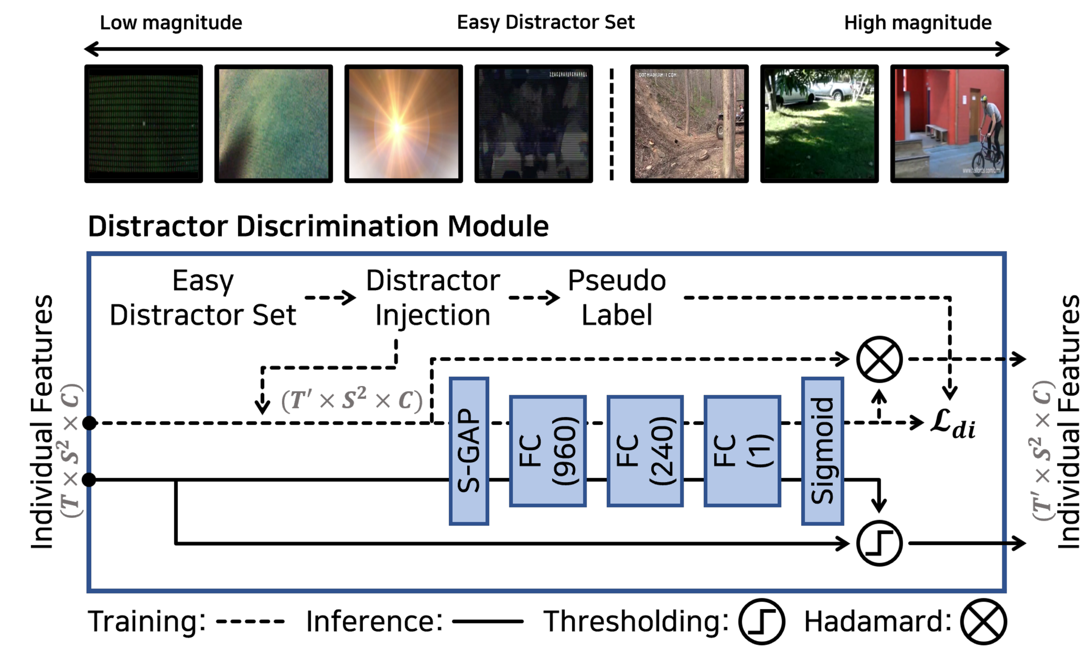


An example of the untrimmed video

Easy Distractor Eilmination Stage

Distractor Discrimination Module

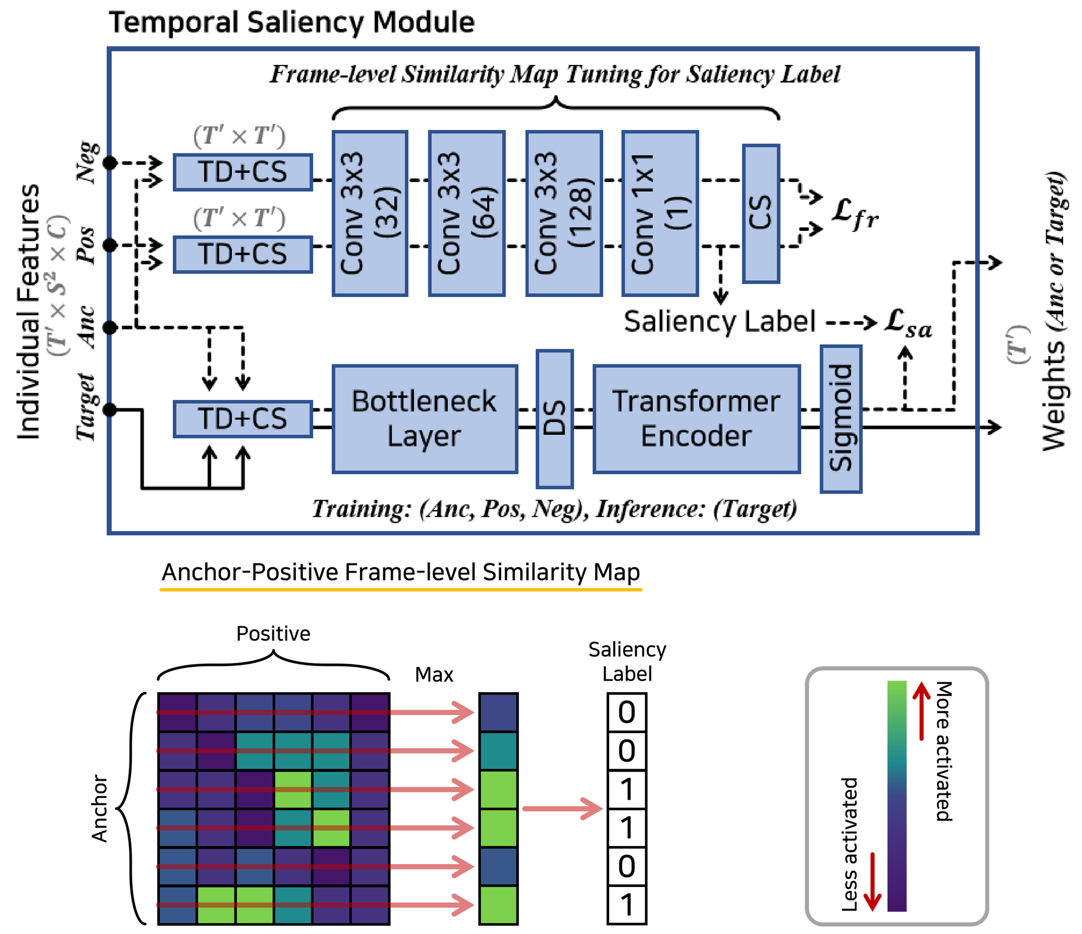
- Eliminate easy distractors, which are frames with little variation and few low-level characteristics (edges, corners, etc.)
- Train by generating pseudo-labels, leveraging the fact that the easy distractor's feature mainly exhibits a small magnitude due to having fewer elements activated from the backbone network



Suppression Weight Generation Stage for Hard Distractor

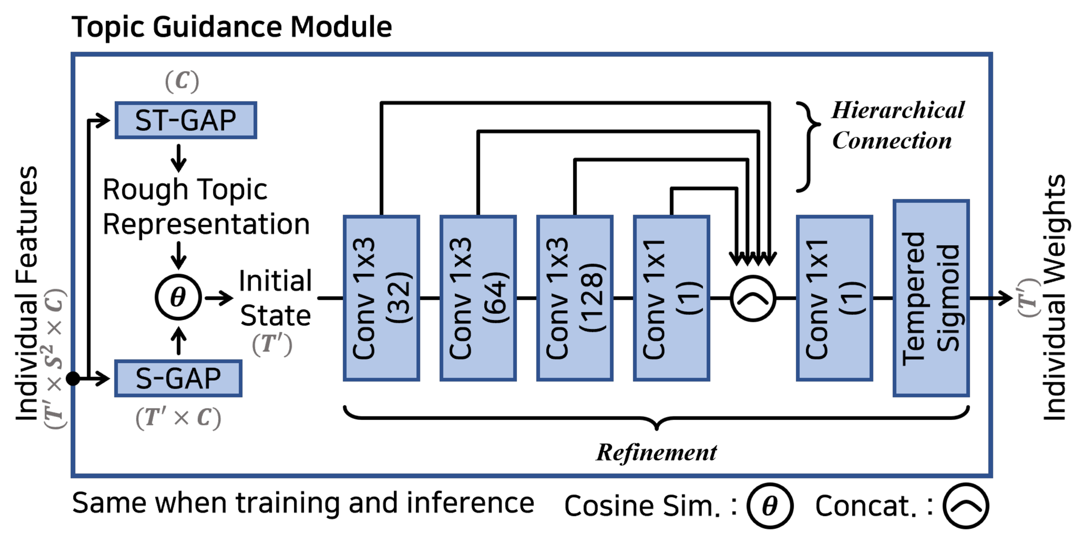
Temporal Saliency Module

- Suppress hard distractors by assessing frames based on the saliency signal
- Train by generating pseudo-labels, leveraging the fact that the highly activated part in the tuned frame-level similarity map represents frames with a strong correlation between a positive pair



Topical Guidance Module

- Suppress hard distractors by measuring the relevance of frames to the overall topic of a video
- Train with refining an initial state, leveraging the fact that the topic of a video is determined by the predominant content within it

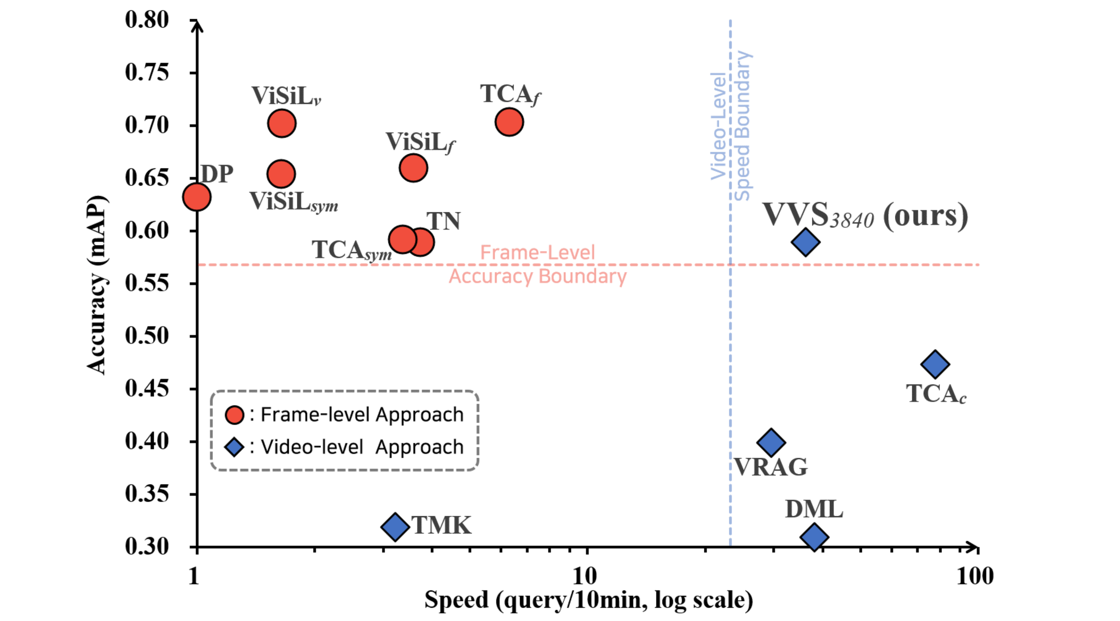


Conclusion

- In this paper, we demonstrate that suppression of irrelevant frames is essential in describing an untrimmed video with long and varied content as a video-level feature.
- Our method removes clearly identifiable frames and determines the extent to which the remaining frames should be suppressed, utilizing saliency information and topic relevance.

Results

Comparison with Other Approaches



Quantitative Results & Ablation Results

	Approach	FIVR-200K DSVR	FIVR-200K CSVR	FIVR-200K ISVR
frame	TN	0.724	0.699	0.589
	DP	0.775	0.740	0.632
	TCA$_f$	0.877	0.830	**0.703**
	ViSiL$_v$	**0.892**	**0.841**	0.702
video	DML	0.398	0.378	0.309
	VRAG	0.484	0.470	0.399
	TCA$_c$	0.570	0.553	0.473
	VVS (ours)	**0.711**	**0.689**	**0.590**

	DDM	TSM	TGM	FIVR-5K DSVR	FIVR-5K CSVR	FIVR-5K ISVR
(a)				0.692	0.700	0.651
(b)	✓			0.715	0.725	0.672
(c)		✓		0.702	0.710	0.661
(d)			✓	0.716	0.724	0.677
(e)		✓	✓	0.719	0.726	0.680
(f)	✓	✓		0.724	0.732	0.683
(g)	✓		✓	0.738	0.746	0.698
(h)	✓	✓	✓	**0.744**	**0.752**	**0.705**

Qualitative Results

Gray region: eliminated by DDM
Orange line: suppressed by TSM and TGM

Untrimmed Video Task Versatility

Approach	Reference	SumMe F-score	SumMe *Rank.*	TVSum F-score	TVSum *Rank.*	Average *Rank.*	Data splits
Random summary	-	40.2	19	54.4	16	17.5	-
SUM-FCN$_{unsup}$	(Rochan, Ye, and Wang 2018)	41.5	17	52.7	17	17	M Rand
DR-DSN	(Zhou, Qiao, and Xiang 2018)	41.4	18	57.6	13	15.5	5 Rand
EDSN	(Gonuguntla et al. 2019)	42.6	15	57.3	14	14.5	5 Rand
RSGN$_{unsup}$	(Zhao et al. 2021)	42.3	16	58.0	12	14	5 Rand
UnpairedVSN	(Rochan and Wang 2019)	47.5	12	55.6	15	13.5	5 Rand
PCDL	(Zhao, Li, and Lu 2019)	42.7	14	58.4	10	12	5 FCV
ACGAN	(He et al. 2019)	46.0	13	58.5	9	11	5 FCV
SUM-Ind_{LU}	(Yaliniz and Ikizler-Cinbis 2021)	46.0	13	58.7	8	10.5	5 Rand†
ERA	(Wu, Lin, and Silva 2021)	48.8	9	58.0	12	10.5	5 Rand†
SUM-GAN-sl	(Apostolidis et al. 2019)	47.8	11	58.4	10	10.5	5 Rand†
SUM-GAN-AAE	(Apostolidis et al. 2020b)	48.9	8	58.3	11	9.5	5 Rand†
MCSF$_{late}$	(Kanafani et al. 2021)	47.9	10	59.1	6	8	5 Rand†
SUM-GDA$_{unsup}$	(Li et al. 2021)	50.0	7	59.6	5	6	5 FCV
CSNet+GL+RPE	(Jung et al. 2020)	50.2	6	59.1	6	6	5 FCV
CSNet	(Jung et al. 2019)	51.3	2	58.8	7	4.5	5 FCV
DSR-RL-GRU	(Phaphuangwittayakul et al. 2021)	50.3	5	60.2	4	4.5	5 Rand†
AC-SUM-GAN	(Apostolidis et al. 2020a)	50.8	4	60.6	3	3.5	5 Rand†
CA-SUM	(Apostolidis et al. 2022)	51.1	3	61.4	2	2.5	5 Rand†
VVS$_{3840}$ (ours)	-	**51.7**	**1**	**61.5**	**1**	**1**	5 Rand†

Video summarization benchmark