WACV
VIRTUAL JANUARY 5-9

ByteDance AI Lab

Temporal Context Aggregation for Video Retrieval with Contrastive Learning

Jie Shao*[1,3] , Xin Wen*[2,3] , Bingchen Zhao[2] , and Xiangyang Xue[1]

[1]School of Computer Science, Fudan University
[2]Department of Computer Science and Technology, Tongji University
[3]ByteDance AI Lab

WACV 2021

Content-Based Video Retrieval

- From Near-Duplicate Video Retrieval (NDVR) to Fine-grained Incident Video Retrieval (FIVR)
- Require higher-level video representation

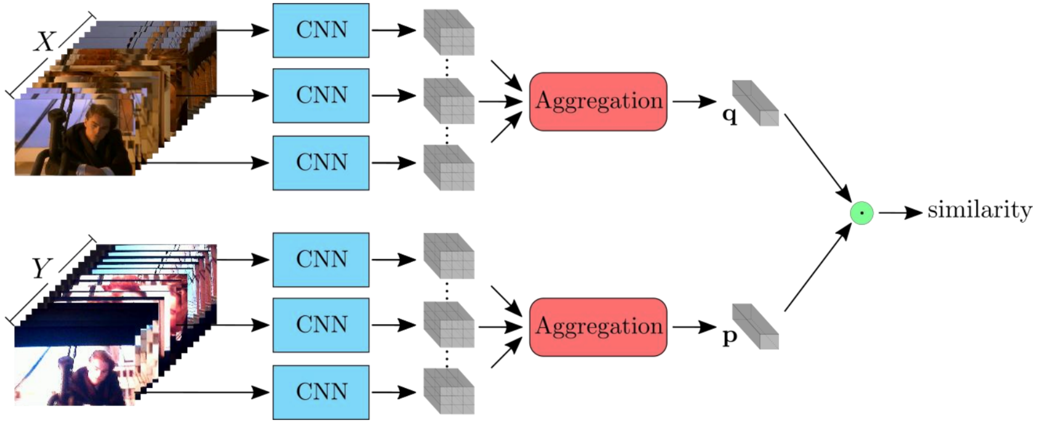
Query | Duplicate Scene | Complementary Scene | Incident Scene

Kordopatis-Zilos, Giorgos, et al. "FIVR: Fine-grained incident video retrieval." *IEEE Transactions on Multimedia* 21.10 (2019): 2638-2652.

To predict the similarity between video pairs

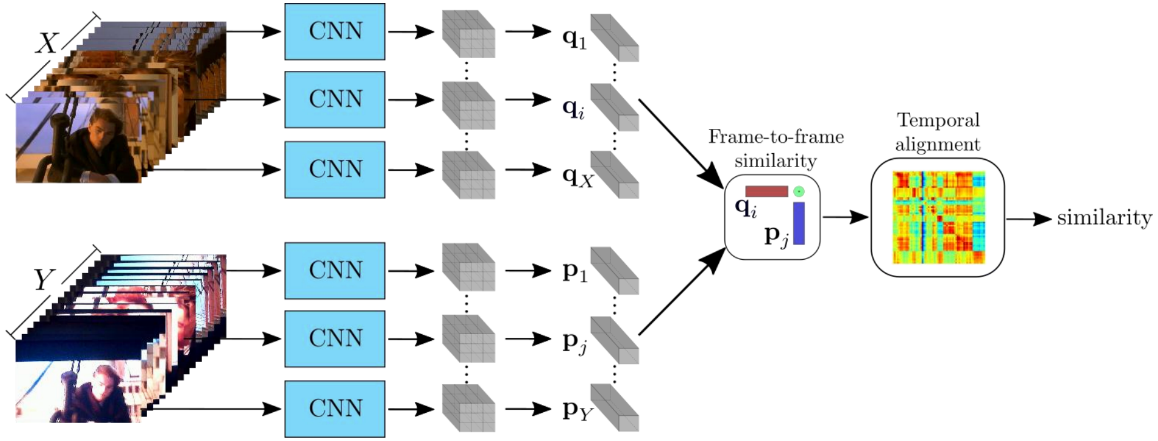
Video-level Methods

- Compute the similarity using video-level representations



Frame-level Methods

- Compute the similarity using frame-level representations



However, the frames of a video are commonly processed as *individual images* or *short clips*...

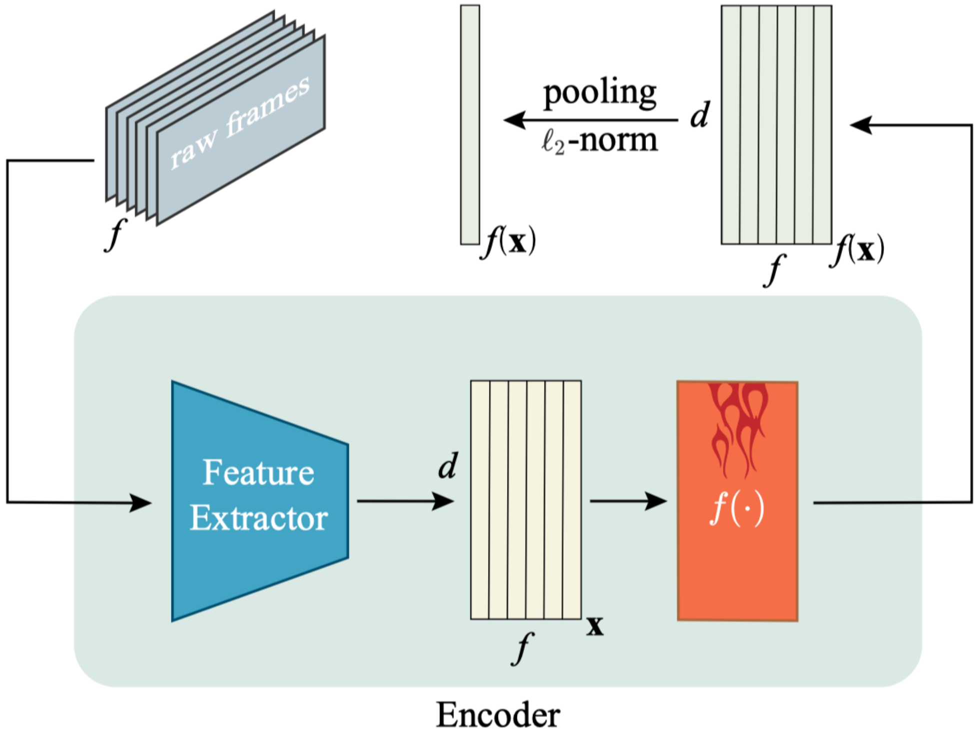
Kordopatis-Zilos, Giorgos, et al. "Visil: Fine-grained spatio-temporal video similarity learning." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.

Without long-range semantic dependencies...

Potentially unnecessary visual data may dominate the video representation, and mislead the model to retrieve negative samples sharing similar scenes.

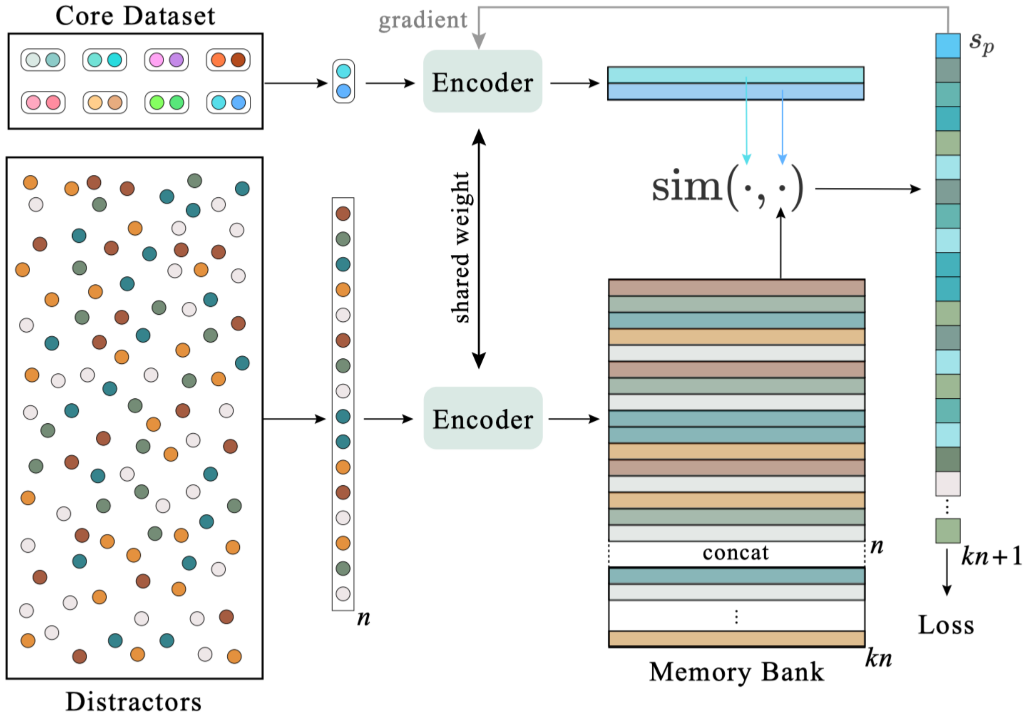
Our motivation

- Incorporating temporal contextual information with the self-attention mechanism (Transformer)
- Output both frame-level descriptor and video-level descriptor



Then, how to train it?

- Supervised contrastive learning with memory bank
- Utilize large quantities of negative samples in the *distractor* subset
- Norm + softmax loss = automatic hard sample mining



$$\frac{\partial \mathcal{L}_{\text{softmax}}}{\partial \mathbf{w}_a} = \frac{\partial \mathbf{z}_a}{\partial \mathbf{w}_a} \cdot \frac{\partial \mathcal{L}_{\text{softmax}}}{\partial \mathbf{z}_a}$$

$$= \frac{1}{\|\mathbf{w}_a\|} \left(\mathbf{I} - \mathbf{z}_a \mathbf{z}_a^\top\right) \left[(\sigma(\mathbf{s})_p - 1)\, \mathbf{z}_p + \sum_{j=1}^{N-1} \sigma(\mathbf{s})_n^j \mathbf{z}_n^j \right]$$

$$\propto \overbrace{(1 - \sigma(\mathbf{s})_p)[(\mathbf{z}_a^\top \mathbf{z}_p)\mathbf{z}_a - \mathbf{z}_p]}^{\text{positive}} + \underbrace{\sum_{j=1}^{N-1} \sigma(\mathbf{s})_n^j [\mathbf{z}_n^j - (\mathbf{z}_a^\top \mathbf{z}_n^j)\mathbf{z}_a]}_{\text{negatives}},$$

Ablations

Model	DSVR	CSVR	ISVR
NetVLAD	0.513	0.494	0.412
LSTM	0.505	0.483	0.400
GRU	0.515	0.495	0.415
Transformer	**0.551**	**0.532**	**0.454**

(a) **Model** (mAP on FIVR-5K)

Feature	DSVR	CSVR	ISVR
iMAC	0.547	0.526	0.447
L_3-iRMAC	**0.570**	**0.553**	**0.473**

(b) **Feature** (mAP on FIVR-200K)

Loss	τ/γ	DSVR	CSVR	ISVR
InfoNCE	0.07	0.493	0.473	0.394
InfoNCE	$1/256$	0.566	0.548	0.468
Circle	256	**0.570**	**0.553**	**0.473**

(c) **Loss function** (mAP on FIVR-200K)

Method	Bank Size	DSVR	CSVR	ISVR
triplet	-	0.510	0.509	0.455
ours	256	0.605	0.615	0.575
ours	4096	0.609	**0.617**	**0.578**
ours	65536	**0.611**	**0.617**	0.574

(d) **Bank size** (mAP on FIVR-5K)

Momentum	DSVR	CSVR	ISVR
0 (bank)	**0.609**	**0.617**	**0.578**
0.1	0.606	0.612	0.569
0.9	0.605	0.611	0.568
0.99	0.602	0.606	0.561
0.999	0.581	0.577	0.520

(e) **Momentum** (mAP on FIVR-5K)

Similarity Measure	DSVR	CSVR	ISVR
cosine	0.609	0.617	0.578
chamfer	**0.844**	**0.834**	**0.763**
symm. chamfer	0.763	0.766	0.711
chamfer+comparator	0.726	0.735	0.701

(f) **Similarity Measure** (mAP on FIVR-5K)

Table 1: **Ablations on FIVR about:** (a): Temporal context aggregation methods; (b): Frame feature representations; (c): Loss functions for contrastive learning ($\gamma = 1/\tau$); (d) Size of the memory bank; (e) Momentum parameter of the queue of MoCo [17], degenerate to memory bank when set to 0; (f) Similarity measures (video-level and frame-level), comparator: the comparator network used in ViSiL$_v$ [31], with original parameters retained.

Evaluation

Method		FIVR-200K			EVVE
		DSVR	CSVR	ISVR	
Video-level	DML [33]	0.398	0.378	0.309	-
	HC [52]	0.265	0.247	0.193	-
	LAMV+QE [4]	-	-	-	0.587
	TCA$_c$	**0.570**	**0.553**	**0.473**	**0.598**
Frame-level	DP [9]	0.775	0.740	0.632	-
	TN [54]	0.724	0.699	0.589	-
	ViSiL$_f$ [31]	0.843	0.797	0.660	0.597
	ViSiL$_{sym}$ [31]	0.833	0.792	0.654	0.616
	ViSiL$_v$ [31]	**0.892**	**0.841**	0.702	0.623
	TCA$_f$	0.877	0.830	**0.703**	0.603
	TCA$_{sym}$	0.728	0.698	0.592	**0.630**

Table 3: **mAP on FIVR-200K and EVVE.** The proposed approach achieves the best trade-off between performance and efficiency with both video-level and frame-level features against state-of-the-art methods.

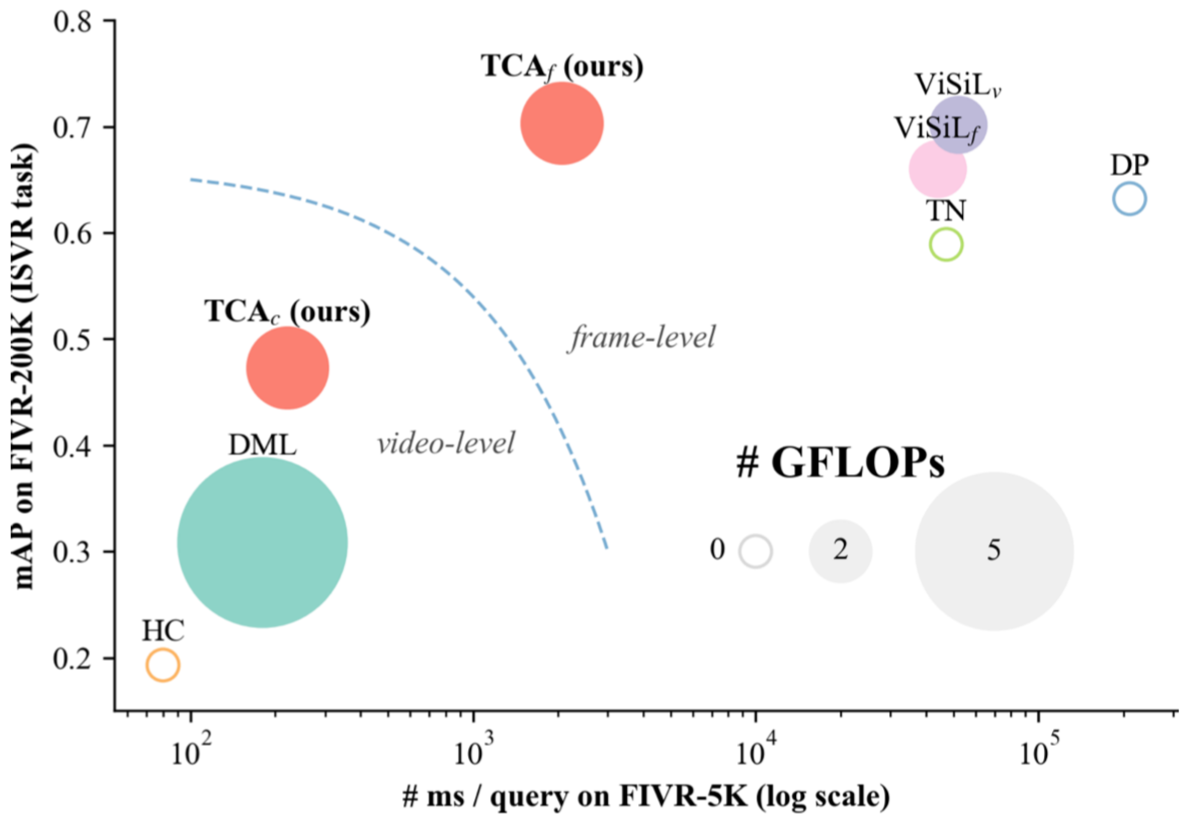
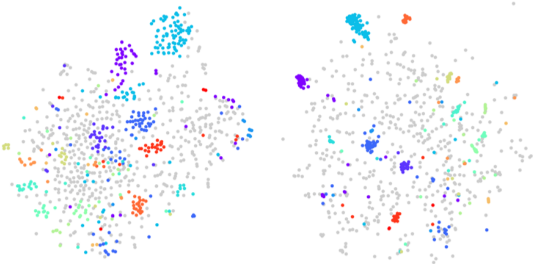


Figure 2: Video Retrieval performance comparison on ISVR task of FIVR [30] in terms of mAP, inference time, and computational cost of the model (ISVR is the most complete and hard task of FIVR). The proposed approach achieves the best trade-off between performance and efficiency with both video-level and frame-level features against state-of-the-art methods. (*Best viewed in color*)

Qualitative Results



(a) DML [33] (b) Ours (TCA_c)

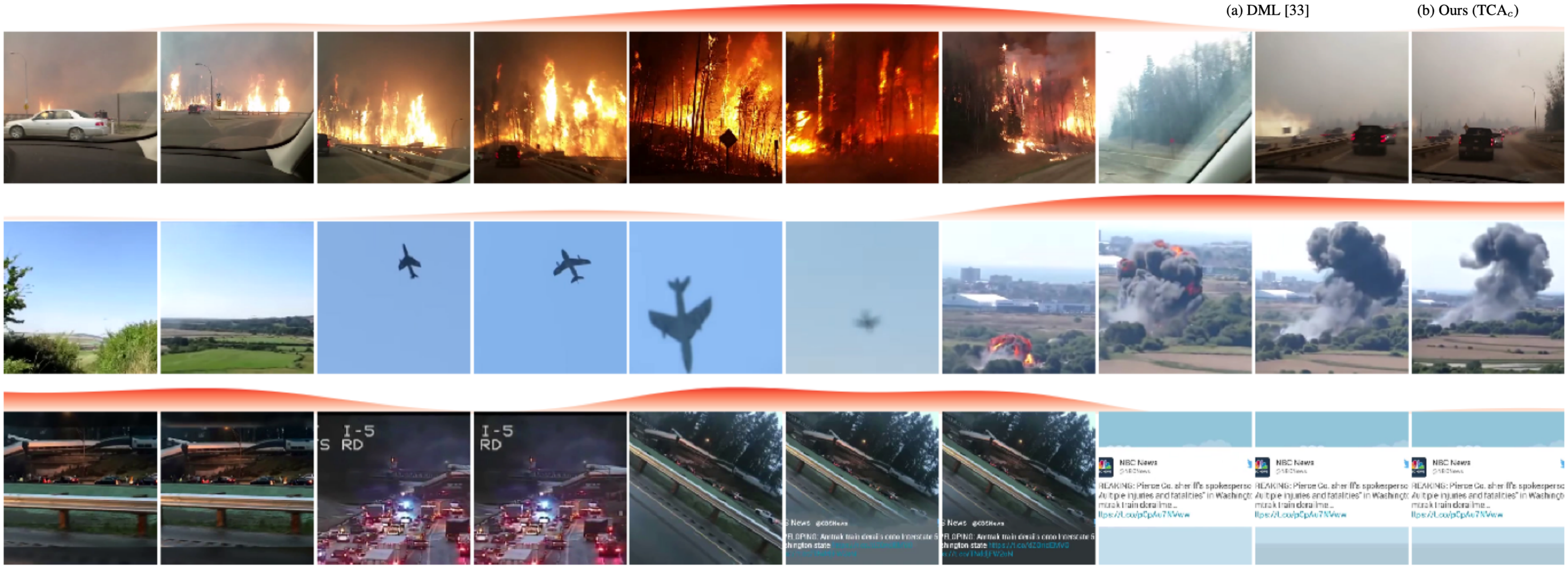


Figure 5: **Visualization of average attention weight (response) of example videos in FIVR.** The weights are normalized and interpolated for better visualization, and darker color indicates higher average response of the corresponding frame. Each case tends to focus on salient and informative frames: video #1 focuses on key segments about the fire; video #2 has a higher focus on the explosion segment; and video #3 selectively ignores the meaningless ending.

WACV
VIRTUAL JANUARY 5-9

ByteDance AI Lab

Thank you!

- Code will be available soon: https://arxiv.org/abs/2008.01334
- Contact this guy for any question: https://wen-xin.info (Xin Wen)
- This guy is looking for a summer research position in Computer Vision: http://info.zhaobc.me (Bingchen Zhao)