TEVAD: Improved video anomaly detection with captions

Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, David Khoo
Hyundai Motor Group Innovation Center

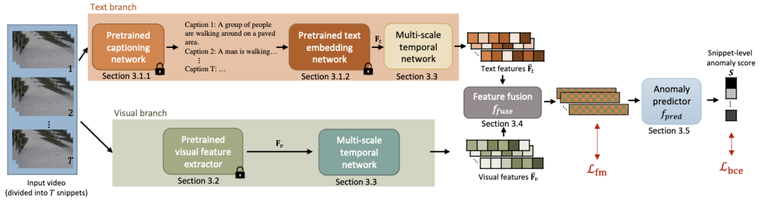
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Motivation

Previous methods do not consider the high-level semantic meanings of the videos making it difficult to detect certain abnormal events and generalize the models to complex scenarios. Moreover, the actual detection is done based on the anomaly scores generated by the models which are obscure to the front-end surveillance systems users.

Framework

TEVAD first splits the input video into T snippets and feed them into two individual branches. The text branch computes text features based on generated dense captions of snippets, while the visual branch extracts visual features. Both modality features go through a multi-scale temporal networks before being fused together and passed to a binary classifier that outputs anomaly scores for each video snippet which are then propagated to predict the frame level anomaly scores.



Experimental results

Experimental results show that our proposed framework achieves SOTA results on four benchmark datasets (Table 1-4). We perform an ablation study on different datasets to demonstrate the effectiveness of the main components in TEVAD and the results are shown in Table 5.

Visual	Text	Fusion	Ped2 (%)	Shanghai (%)	Crime (%)	Violence (%)
✓	×	×	83.81	94.17	83.1	76.94
✓	Vanilla	concat	93.17	97.85	83.18	77.91
✓	MTN	concat	96.71	97.86	**84.9**	79.3
✓	MTN	add	**98.69**	**98.1**	84.13	**79.76**
✓	MTN	product	94.12	97.2	83.83	78.49

Table 5. Ablation study results.

Type	Source	Method	AUC (%)
Unsup	CVPR'18	Liu *et al.* [32]	95.4
	WACV'22	FastAno [40]	96.3
	CVPR'21	SSMTL [16]	97.5
	CVPR'20	CL-VAD [11]	97.8
	TPAMI'21	Georgescu *et al.* [17]	98.7
Sup	CVPR'19	GCN-Anomaly [68]	93.2
	CVPR'18	Sultani *et al.* [48]	92.3
	ICCV'21	RTFM [50]	98.6
	–	**TEVAD**	**98.7**

Table 1. Frame-level AUC results on UCSD Ped2 dataset.

Type	Source	Method	AUC (%)
Unsup	ICCV'19	BODS [54]	68.3
	ICCV'19	GODS [54]	70.5
	Patter Recog'20	FSCN [60]	70.6
Sup	CVPR'18	Sultani *et al.* [48]	75.4
	CVPR'19	GCN-Anomaly [68]	82.1
	CVPR'21	MIST [13]	82.3
	CVPR'22	BN-SVP [45]	83.4
	ICCV'21	RTFM [50]	84.3
	IEEE Trans Multimedia'21	Chang *et al.* [8]	84.6
	TIP'21	Wu *et al.* [59]	84.9
	–	**TEVAD**	**84.9**

Table 3. Frame-level AUC results on UCF-Crime dataset.

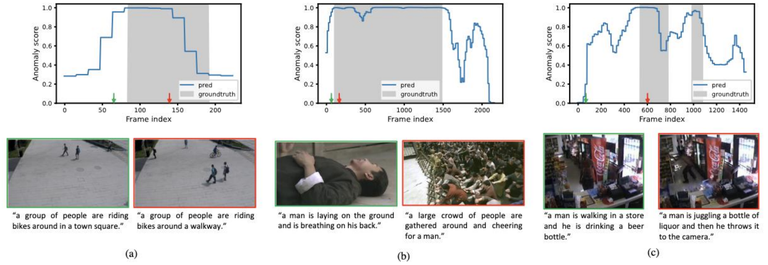
Type	Source	Method	AUC (%)
Unsup	CVPR'20	CL-VAD [11]	71.6
	TPAMI'21	Georgescu *et al.* [17]	82.7
	CVPR'22	SSPCAB [43]	83.6
	CVPR'22	SSMTL [1]	83.7
Sup	CVPR 2019	GCN-Anomaly [68]	84.4
	ICME'20	AR-Net [53]	91.2
	IEEE Trans Multimedia'21	Chang *et al.* [8]	92.3
	CVPR'21	MIST [13]	94.8
	CVPR'22	BN-SVP [45]	96.0
	ICCV'21	RTFM [50]	97.2
	TIP'21	Wu *et al.* [59]	97.5
	–	**TEVAD**	**98.1**

Table 2. Frame-level AUC results on ShanghaiTech dataset.

Type	Source	Method	AP (%)
Sup	arXiv'22	CSL-TAL [39]	71.7
	CVPR'18	Sultani *et al.* [48]	75.7
	TIP'21	Wu *et al.* [59]	75.9
	IEEE Trans Multimedia'21	Chang *et al.* [8]	76.9
	ICCV'21	RTFM [50]	77.8
	–	**TEVAD**	**79.8**

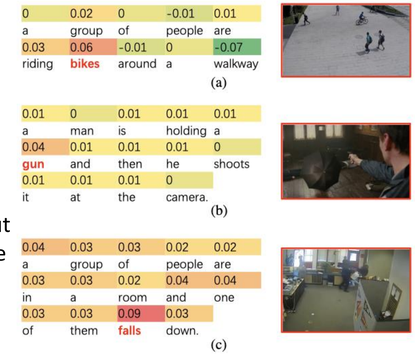
Table 4. Frame-level AP results on XD-Violence dataset.

We provide some **qualitative results** from (a) ShanghaiTech (riding a bike), (b) XD-Violence (riot) , and (c) UCF-Crime (vandalism) datasets. The top row shows predicted anomaly scores and the groundtruth labels. For frames labeled with green or red arrows, we also show the image frames and their associated generated captions in the bottom row.



We conduct additional analysis to demonstrate the explainability of incorporating captions for video anomaly detection tasks. During the inference phase, we iteratively mask each word in the caption of the snippet and calculate its anomaly score for each snippet of the video. Figure below shows the **explainability results** to understand the contribution of each word in captions of the snippets: (a) ShanghaiTech (riding a bike), (b) XD-Violence (shooting), and (c) UCF-Crime (arrest) datasets.

An image frame of the abnormal event from the snippet is also shown on the right of each caption. The score above each word in the caption is the difference between the anomaly score by masking this word and the original anomaly score without masking. Therefore, a higher score indicates a higher contribution to the predicted anomaly score.



Conclusions

Our contributions of this work are:

- We propose a framework, TEVAD, which exploits both visual and text features for video anomaly detection with different multi-modal fusion methods.
- We extend multi-scale temporal learning to text features to better capture the dependencies between snippet features.
- Our proposed framework outperforms the SOTA methods on four benchmark datasets and achieves improved robustness.
- We further conduct additional analysis to provide explainability for the anomalous videos identified through the use of a word-masking protocol.

Our codes are available at https://github.com/coranholmes/TEVAD