# TEVAD: Improved video anomaly detection with captions Supplementary Materials

Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, David Aik-Aun Khoo
Hyundai Motor Group Innovation Center in Singapore
2 Bulim Link, Singapore 649674
weiling.chen,kengteck.ma,zijian.yew, david.khoo@hmgics.com

## A. Multi-scale Temporal Network

We extend Multi-scale Temporal Network (MTN) [8] to process text features to learn the long and short range temporal dependencies between snippet text features.

Pre-computed text features $\mathbf{F_{txt}}$ (*i.e.* sentence embeddings of the video snippets), where $\mathbf{F_{txt}} \in \mathbb{R}^{d_{\text{txt}}}$, are fed into the three pyramid dilated convolution (PDC) layers respectively given as $\mathbf{F}^{(P_i)} = f_{conv}(\mathbf{F_{txt}};\theta)$ for $i \in \{1,2,3\}$, where $\mathbf{F}^{(P_i)} \in \mathbb{R}^{d_{\text{txt}}/4}$ and $f_{conv}$ is a 1D convolution function. $\theta$ comprises the weights for all convolution functions described in this section. The three feature vectors are then concatenated and processed by another 1D convolution layer as $\mathbf{F}^{(PDC)} = f_{conv}(\mathbf{F}^{(P_1)}|\mathbf{F}^{(P_2)}|\mathbf{F}^{(P_3)};\theta)$.

Meanwhile, $\mathbf{F_{txt}}$ goes through a convolution layer and is fed into the non-local block (NLB). A convolution function is applied separately for three times to obtain $\mathbf{F}^{(c_i)} = f_{conv}(f_{conv}(\mathbf{F_{txt}});\theta)$ for $i \in \{1,2,3\}$ and finally produce $\mathbf{F}^{(NLB)} = f_{conv}((\mathbf{F}^{(c_1)})(\mathbf{F}^{(c_2)})^T(\mathbf{F}^{(c_3)});\theta)$, where $\mathbf{F}^{(PDC)} \in \mathbb{R}^{d_{\text{txt}}/4}$.

The outputs from the two blocks are concatenated and added to the original features to produce the final output of text MTN is given as $\bar{\mathbf{F}}_{\mathbf{txt}} = f_{conv}(\mathbf{F}^{(PDC)}|\mathbf{F}^{(NLB)};\theta) + \mathbf{F}$, where $\bar{\mathbf{F}}_{\mathbf{txt}} \in \mathbb{R}^{d_{\text{txt}}}$. Both visual and text features go through the similar process so that TEVAD is able to learn the temporal dependencies between video snippets in both modalities.

## B. Datasets

We have carried out our experiments on four benchmark datasets, namely UCSD Ped2 [11], ShanghaiTech [4], UCF-Crime [7], and XD-Violence [10].

**UCSD Ped2**[1]**:** This is a small-scale dataset proposed in 2013 consisting of 16 normal videos and 12 abnormal videos. The videos in this dataset were collected from campus CCTV. We follow the setting in [3, 8, 12] and randomly select 6 abnormal videos and 4 normal videos to form the training set and the rest as the test set to carry out the experiments.

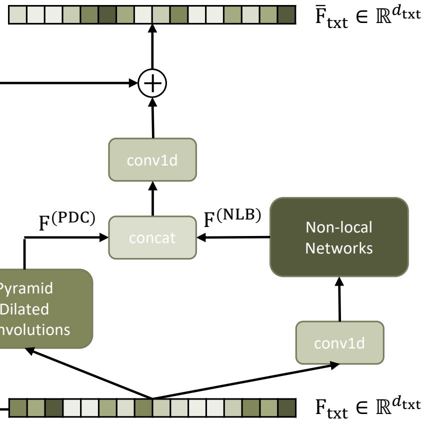

$\bar{\text{F}}_{\text{txt}} \in \mathbb{R}^{d_{\text{txt}}}$
conv1d
$\text{F}^{(\text{PDC})}$
concat
$\text{F}^{(\text{NLB})}$
Non-local Networks
Pyramid Dilated Convolutions
conv1d
$\text{F}_{\text{txt}} \in \mathbb{R}^{d_{\text{txt}}}$

Figure 1. An overview of MTN.

**ShanghaiTech**[2]**:** This is a medium-scale dataset released in 2018 including 437 videos with 307 normal videos and 130 anomaly videos. Similar to UCSD Ped2, the videos were collected from campus surveillance systems and covers 13 background scenes. The original dataset was designed for unsupervised anomaly detection tasks where only the normal data are available during training. To perform experiments on this dataset, we follow the setting in [8, 9, 12] and split the dataset into a training set made up of 238 videos and the remaining 199 videos as test set.

**UCF-Crime**[3]**:** This dataset was released in 2018 and includes a total of 1900 videos with the duration of 128 hours. This dataset was collected from real-world surveillance systems and contains 13 crime related abnormal events including abuse, arrest, arson, assault, accident, burglary, explosion, fighting, robbery, shooting, stealing, shoplifting, and

[1] UCSD-Pedestrian dataset: http://www.svcl.ucsd.edu/projects/anomaly/dataset.html

[2] ShanghaiTech dataset: https://svip-lab.github.io/dataset/campus_dataset.html

[3] UCF-Crime dataset: https://webpages.charlotte.edu/cchen62/dataset.html

vandalism. The dataset is split into training and test set with 1,610 and 290 videos respectively.

**XD-Violence**[4]**:** This dataset was published in 2020 and contains 4,754 videos, of which 3,954 are assigned for training while the rest are for testing. The dataset has a total duration of more than 217 hours. As suggested by its name, this dataset covers 6 violence related classes, namely abuse, road accident, explosion, fighting, riot and shooting. Different from UCF-Crime, the dataset further includes scenes collected from movies. Another feature differentiating this dataset from other video anomaly detection datasets is that it is an audiovisual dataset. However, we exploit visual features only in our experiments.

## C. Limitations

To generate the captions of snippets, we employ a sliding window strategy and compute the caption for a consecutive 64 frames for every 16 frame. In this way, the text features contain the information not only from the current snippet but the following three snippets while the visual features only contain the information from current snippet. This inconsistency may result in the predicted starting and ending frames of anomalous events slightly earlier than the ground truths (See Figure 3.(a) in main paper). Nevertheless, this does not affect the overall performance significantly.

In addition, since the video anomaly detection datasets do not contain the necessary captions to train the caption generation models, we use the pre-trained models trained on other video captioning datasets. This results in the inaccuracy of captions in some cases. However, this issue can be resolved if some captions are provided during training.

## D. Societal Impact

While anomaly detection technology in video surveillance can be misused, the potential societal benefits far outweighs such risks [1,2,5,6]. The model we have proposed in this work can be used to reduce the personal and property losses in many real-world scenarios including healthcare, manufacturing, public safety, *etc*.

In particular, our proposed method does not require any personal identifiable information to be collected and processed. By carefully introducing automated processes into the current systems, privacy protection can be enhanced as there will be less need for humans to monitor day-to-day activities in public spaces. Only video snippets flagged as anomaly by our proposed algorithm are manually reviewed.

## E. Additional Qualitative Results

We present additional qualitative results from different benchmark datasets in Figures 2 and 3. Similar to what we have presented in Section 4.6 of main paper, our proposed TEVAD can predict anomaly scores effectively. The captions related to day-to-day activities (ShanghaiTech and UCSD Ped2 datasets) are more accurate compared to those related to rarer abnormal events related to crime or violence. Nevertheless, these rarer abnormal events are still reflected with semantically similar words like "fencing movements", "war", "hitting", *etc*.

## References

[1] S Anoopa and A Salim. Survey on anomaly detection in surveillance videos. *Materials Today: Proceedings*, 2022. 2

[2] Paola Cocca, Filippo Marciano, and Marco Alberti. Video surveillance systems to enhance occupational safety: A case study. *Safety Science*, 84:140–148, 2016. 2

[3] Chengkun He, Jie Shao, and Jiayu Sun. An anomaly-introduced learning method for abnormal event detection. *Multimedia Tools and Applications*, 77(22):29573–29588, 2018. 1

[4] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018. 1

[5] Rashmiranjan Nayak, Umesh Chandra Pati, and Santos Kumar Das. A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing*, 106:104078, 2021. 2

[6] Vijeta Sharma, Manjari Gupta, Ajai Kumar, and Deepti Mishra. Video processing using deep learning techniques: A systematic literature review. *IEEE Access*, 2021. 2

[7] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018. 1

[8] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4975–4986, 2021. 1

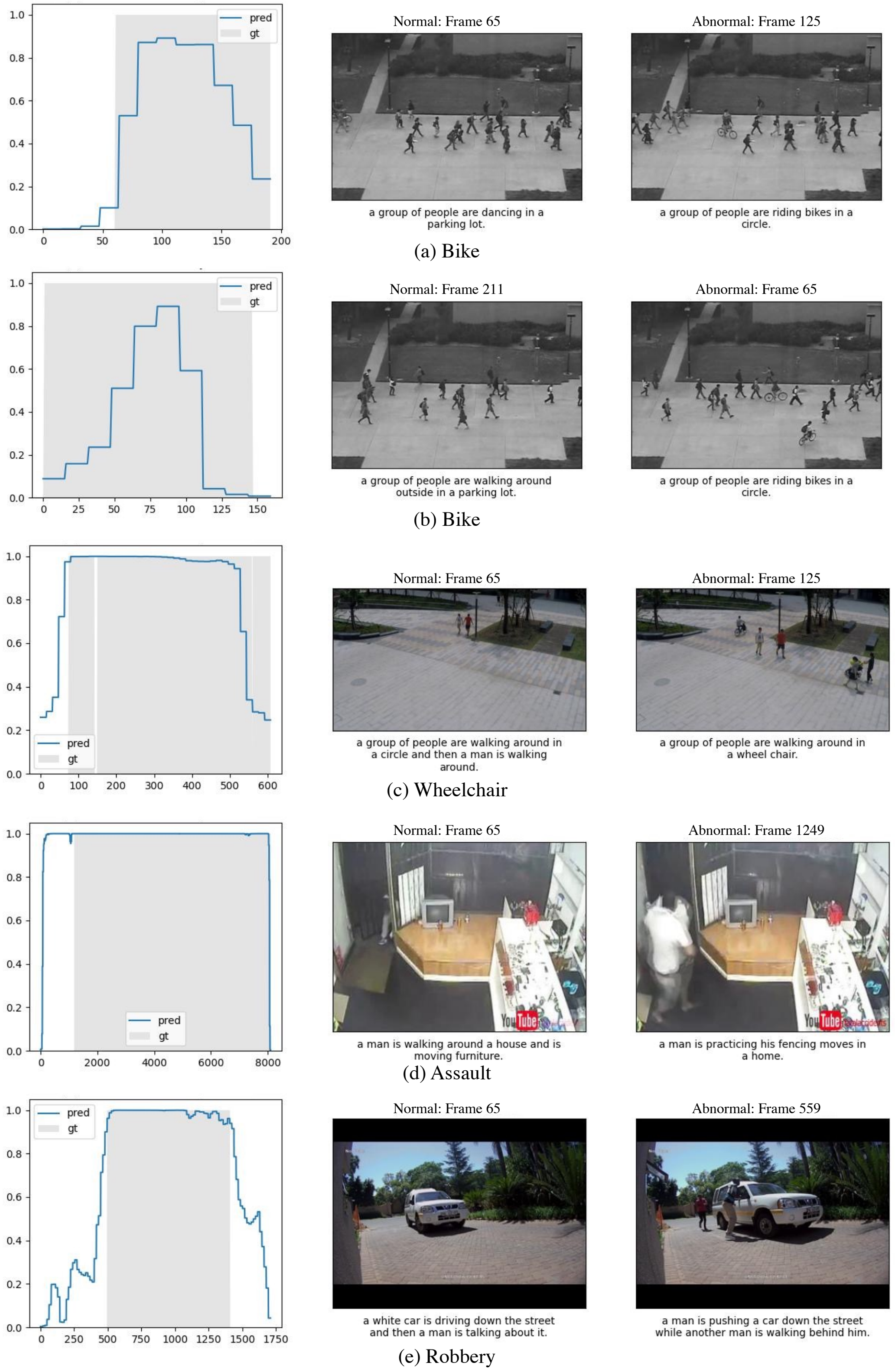
[9] Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, 30:3513–3527, 2021. 1

[10] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*, pages 322–339. Springer, 2020. 1

[11] Dan Xu, Rui Song, Xinyu Wu, Nannan Li, Wei Feng, and Huihuan Qian. Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts. *Neurocomputing*, 143:144–152, 2014. 1

[12] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1237–1246, 2019. 1

[4] XD-Violence dataset: https://roc-ng.github.io/XD-Violence/



(a) Bike

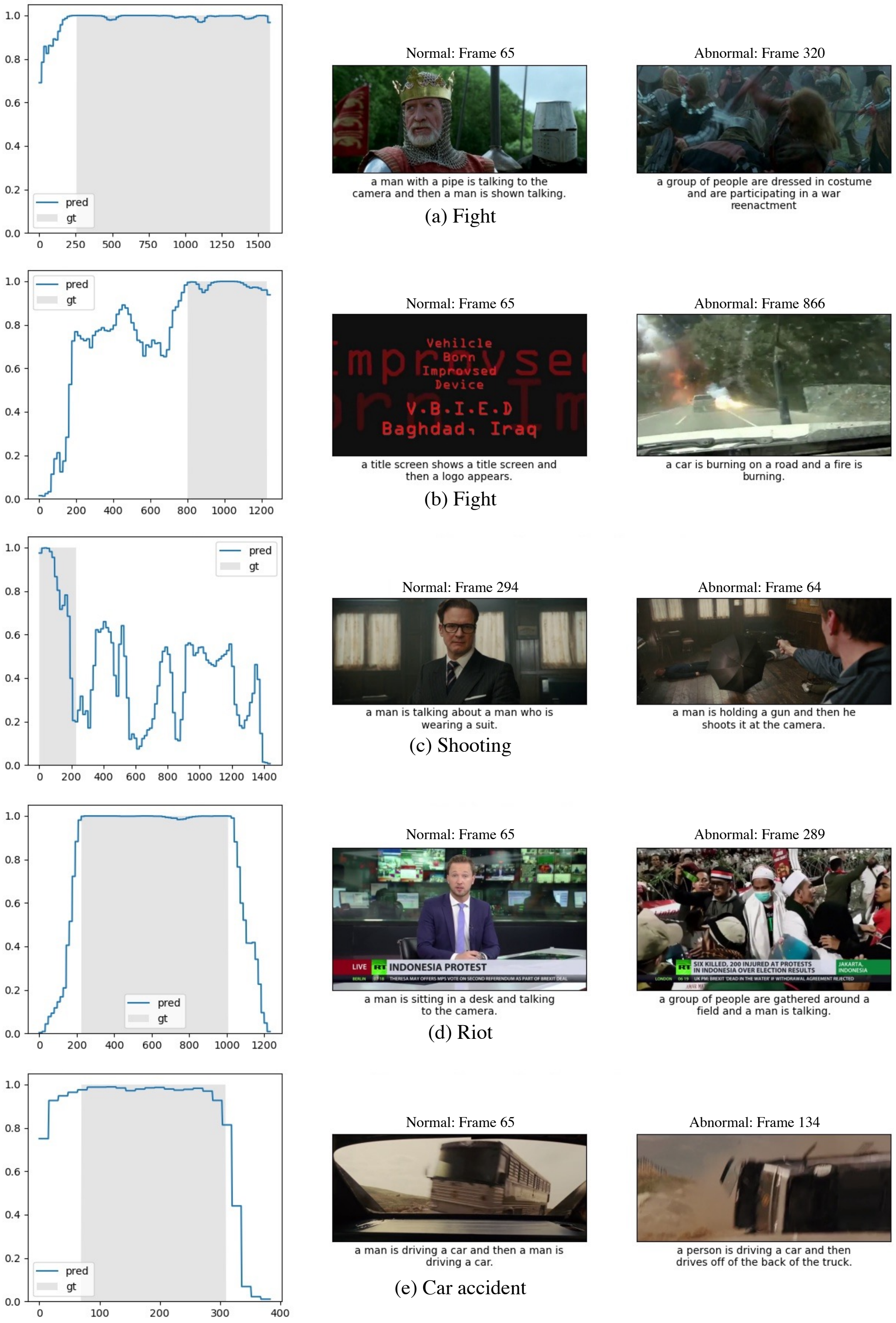
(b) Bike

(c) Wheelchair

(d) Assault

(e) Robbery

Figure 2. Additional qualitative results. (a) and (b) are from Ped2 dataset. (c) is from ShanghaiTech. (d) and (e) are from Crime dataset. For each row, the first column shows predicted anomaly scores and the groundtruth labels. The next two columns show the image frames from a normal and abnormal snippets with their associated generated captions in the bottom.



(a) Fight

(b) Fight

(c) Shooting

(d) Riot

(e) Car accident

Figure 3. Additional qualitative results on Violence dataset.