Supplementary Material: Self-Supervised Sparse Representation for Video Anomaly Detection

Jhih-Ciang Wu[⋆1,2], He-Yen Hsieh[⋆1], Ding-Jie Chen[1], Chiou-Shann Fuh[2], and Tyng-Luh Liu [⋆⋆1]

[1] Institute of Information Science, Academia Sinica, Taiwan
[2] National Taiwan University, Taiwan

This document provides more qualitative results and experimental setting details of our self-supervised sparse representation (S3R) model, which is designed to locate exact video segments of abnormal activities for simultaneously tackling two sorts of video anomaly detections, *i.e.*, oVAD (one-class) and wVAD (weakly-supervised) video anomaly detection tasks. We first present more details of dictionary construction. Then, we provide the statistics of the three datasets, including ShanghaiTech, UCF-Crime, and XD-Violence, for assessing oVAD and wVAD tasks. Finally, we visualize more prediction results of our model.

1 Dictionary Configurations

The dictionary learning aims to keep the representative atoms and filter out the redundant ones in the original set. We collect half of $|\tilde{\mathcal{X}}|$ atoms for all datasets to construct task-specific dictionaries D_T, where $|\tilde{\mathcal{X}}|$ is the number of normal training videos. The universal dictionary D_U contains 2000 atoms after the dictionary learning procedure.

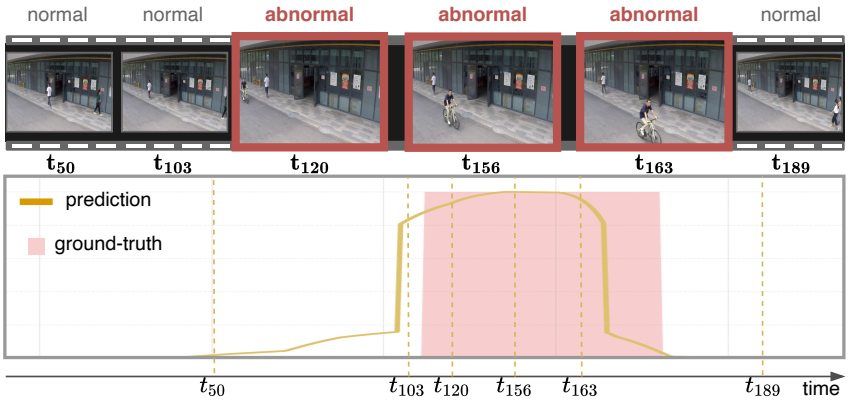
2 Reorganized Datasets

The proposed S3R model is a unified framework capable of solving both the oVAD and wVAD tasks. For fairly assessing our model compared to previous VAD methods, we follow [1,2] to reorganize video datasets, in which the dataset statistics for both tasks are summarized in Table 1.

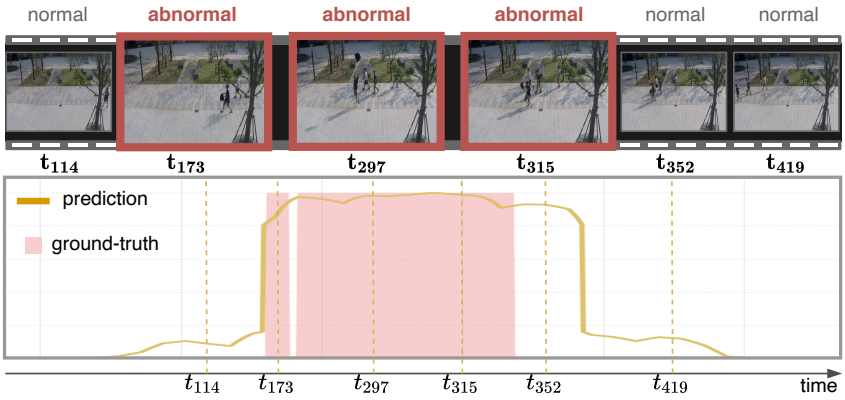
3 Qualitative Results

Fig. 1 to Fig. 3 shows the qualitative results on tackling the wVAD task using our S3R model, in which the results are assessed on the ShanghaiTech or the UCF-Crime datasets. For each subfigure, the top row shows the corresponding frames over the temporal dimension and highlights abnormal events using red boxes; the bottom chart shows the predicted probabilities in the y-axis using a solid brown line and displays the corresponding time over the x-axis. Note that the higher prediction scores indicate the higher probabilities of abnormal events the model believes, and we annotate the ground-truth by filling it in light-red.

⋆ Both authors contributed equally to this work.
⋆⋆ Corresponding author



(a) Testing video id: "06_0153"

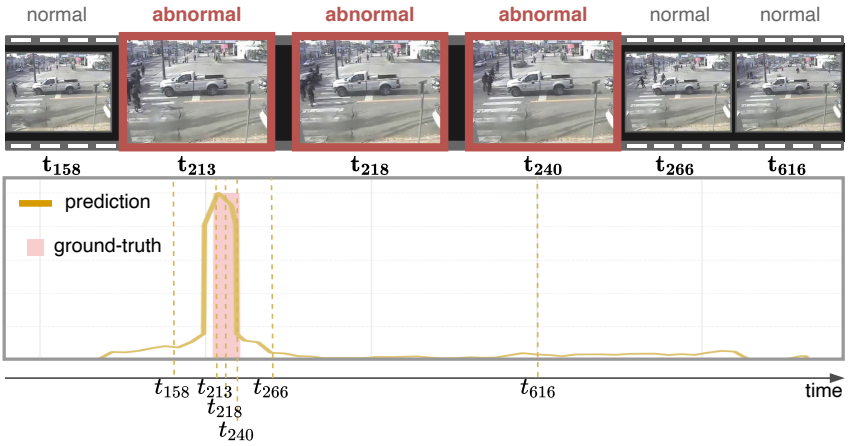


(b) Testing video id: "04_0046"

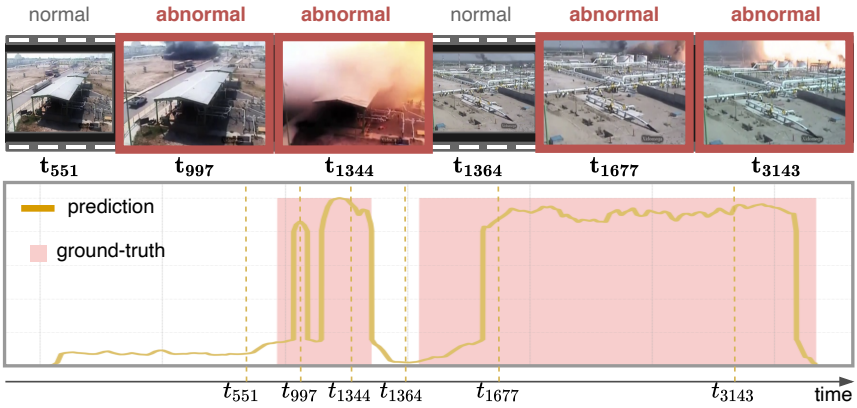
Fig. 1: Qualitative result on wVAD task using the ShanghaiTech dataset. Our S3R model correctly detects various anomalous events, such as *riding the bike* in (a) and *throwing objects* in (b). Note that (b) shows the case that our S3R model continuously predicts the anomalous activity over the temporal dimension compared with the discrete ground-truth label.

References

1. Sun, C., Jia, Y., Hu, Y., Wu, Y.: Scene-aware context reasoning for unsupervised abnormal event detection in videos. In: ACMMM. pp. 184–192 (2020)
2. Zhong, J.X., Li, N., Kong, W., Liu, S., Li, T.H., Li, G.: Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In: CVPR. pp. 1237–1246 (2019)

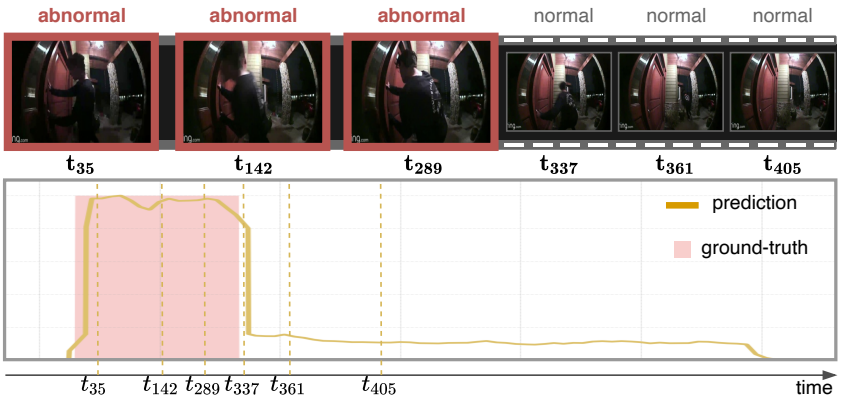


(a) Testing video id: "RoadAccidents009_x264"

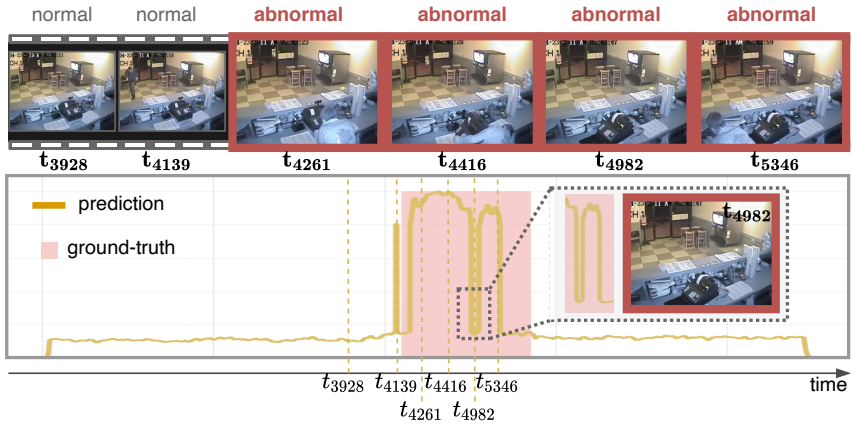


(b) Testing video id: "Explosion033_x264"

Fig. 2: Qualitative result on wVAD task using the UCF-Crime dataset. Our S3R model correctly detects the anomalous events such as *road accident* in (a) and *explosion* in (b). This figure indicates that our model can detect anomalous events even the frames are from different viewpoints in a video.



(a) Testing video id: "Burglary033_x264"



(b) Testing video id: "Burglary061_x264"

Fig. 3: Qualitative result on wVAD task using the UCF-Crime dataset. Our S3R model can (a) detect *stealing* accurately and (b) decrease the anomaly score when the burglary activity disappear in the surveillance camera temporarily. Note that the ground-truth label at t_{4982} is set to 1; however, the corresponding frame shows no abnormal activity, which correctly matched our prediction.

Table 1: The statistics of datasets. The one-class (named as unsupervised in previous works) setting follows [1] to collect all normal videos as the training set and keeps the same testing splits in UCF-Crime and XD-Violence datasets.

Supervision	Types	**ShanghaiTech**			**UCF-Crime**			**XD-Violence**		
		Training	Testing	Total	Training	Testing	Total	Training	Testing	Total
one-class	Normal Videos	330	0	330	800	150	950	2049	300	2349
	Anomalous Videos	0	107	107	0	140	140	0	500	500
	Total	330	107	437	800	290	1090	2049	800	2849
weakly-supervised	Normal Videos	175	155	330	800	150	950	2049	300	2349
	Anomalous Videos	63	44	107	810	140	950	1905	500	2405
	Total	238	199	437	1610	290	1900	3954	800	4754