Self-Supervised Sparse Representation for Video Anomaly Detection

Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu
Institute of Information Science, Academia Sinica, Taiwan
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan

ABSTRACT

Video anomaly detection (VAD) aims at localizing unexpected actions or activities in a video sequence. Existing mainstream VAD techniques are based on either the one-class formulation, which assumes all training data are *normal*, or weakly-supervised, which requires only video-level normal/anomaly labels. To establish a unified approach to solving the two VAD settings, we introduce a ***self-supervised sparse representation*** (S3R) framework that models the concept of anomaly at feature level by exploring the synergy between dictionary-based representation and self-supervised learning. With the learned dictionary, S3R facilitates two coupled modules, ***en-Normal*** and ***de-Normal***, to reconstruct snippet-level features and filter out normal-event features. The self-supervised techniques also enable generating samples of pseudo normal/anomaly to train the anomaly detector. We demonstrate with extensive experiments that S3R achieves new state-of-the-art performances on popular benchmark datasets for both one-class and weakly-supervised VAD tasks.

ARCHITECTURE

ECCV TEL AVIV 2022

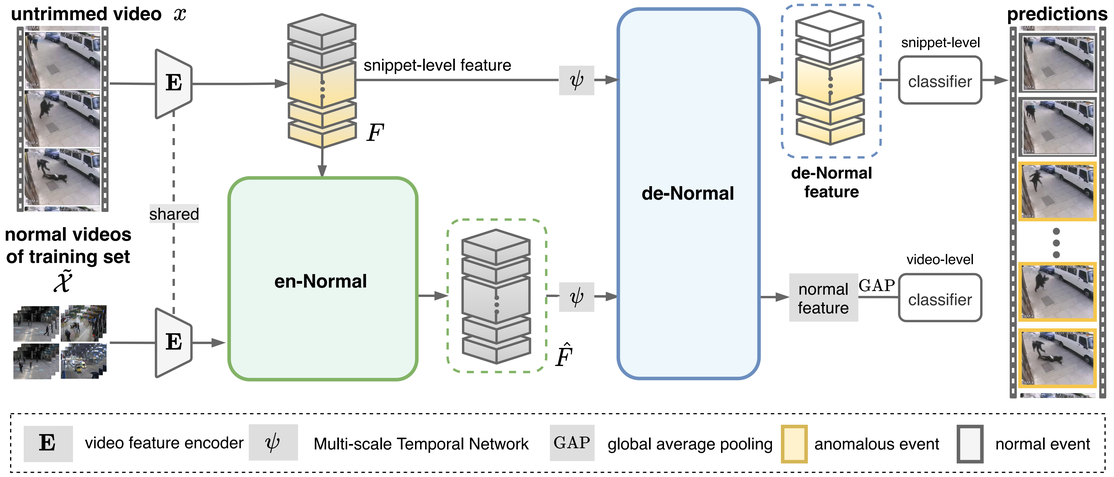

Figure: The proposed S3R framework couples dictionary learning with self-supervised techniques to model the concept of feature-level anomaly. S3R learns a normal-event dictionary for generating two opposite network modules, *i.e.*, en-Normal and de-Normal, to reconstruct snippet-level features and filter out the normal-event features.

EXPERIMENTS

ShanghaiTech

oVAD				wVAD			
Method	Feature	Year	AUC (%)	Method	Feature	Year	AUC (%)
Conv-AE	-	2016	60.85	Sultani *et al.*	I3D	2018	85.33
Stacked-RNN	-	2017	68.00	GCN-Anomaly	C3D	2019	76.44
Frame-Pred	-	2018	73.40	GCN-Anomaly	TSN_{flow}	2019	84.13
Mem-AE	-	2019	71.20	GCN-Anomaly	TSN	2019	84.44
MNAD	-	2020	70.50	AR-Net	$I3D_{flow}$	2020	82.34
VEC	-	2020	74.80	AR-Net	I3D	2020	85.38
STC Graph	-	2020	74.70	AR-Net	$I3D_{f2}$	2020	91.24
CAC	-	2020	79.30	MIST	C3D	2021	93.13
AMMC	-	2020	73.70	MIST	I3D	2021	94.83
HF2-VAD	-	2021	76.20	RTFM	C3D	2021	91.51
ROADMAP	-	2021	76.60	RTFM	I3D	2021	97.21
SVD-GAN	-	2021	78.42	MSL	C3D	2022	94.81
BDPN	-	2022	78.10	MSL	I3D	2022	96.08
S3R	I3D	2022	79.89	S3R	I3D	2022	**97.48**
S3R*	I3D	2022	**80.47**	S3R†	I3D	2022	97.47

UCF-Crime

oVAD				wVAD			
Method	Feature	Year	AUC (%)	Method	Feature	Year	AUC (%)
SVM Baseline	-		50.00	Sultani *et al.*	I3D	2018	77.92
Conv-AE	-	2016	50.60	GCN-Anomaly	TSN	2019	82.12
S-SVDD	-	2018	58.50	MIST	I3D	2021	82.30
Lu *et al.*	C3D	2013	65.51	Wu *et al.*	I3D	2020	82.44
BODS	I3D	2019	68.26	RTFM	I3D	2021	84.30
GODS	I3D	2019	70.46	Chang *et al.*	I3D	2021	84.62
STC Graph	RPN	2020	72.70	MSL	I3D	2022	85.30
S3R	I3D	2022	77.15	S3R	I3D	2022	**85.99**
S3R*	I3D	2022	**79.58**	S3R†	I3D	2022	85.00

XD-Violence

oVAD				wVAD			
Method	Feature	Year	AUC (%)	Method	Feature	Year	AP (%)
-	-	-	-	Sultani *et al.*	I3D	2018	75.68
SVM Baseline	-	-	50.78	Wu *et al.*	I3D	2020	75.41
OCSVM	-	1999	27.25	RTFM	I3D	2021	77.81
Conv-AE	-	2016	30.77	MSL	I3D	2022	78.28
S3R	I3D	2022	51.64	S3R	I3D	2022	**80.26**
S3R*	I3D	2022	**53.52**	S3R†	I3D	2022	79.54

Table: Comparison of frame-level AUC performance for oVAD ("o" for one-class) and wVAD ("w" for weakly-supervised) on three benchmarks. We present the current SOTA with the corresponding feature and published year.

TWO MODULES

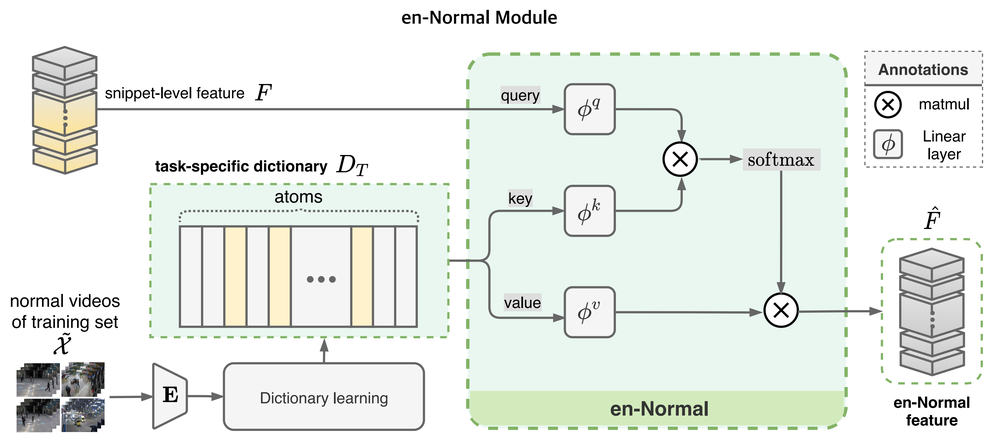

Figure: The pipeline for en-Normal. This module takes the snippet-level feature F and task-specific dictionary D_T to reconstruct feature $\hat{F}$ via an attention mechanism.

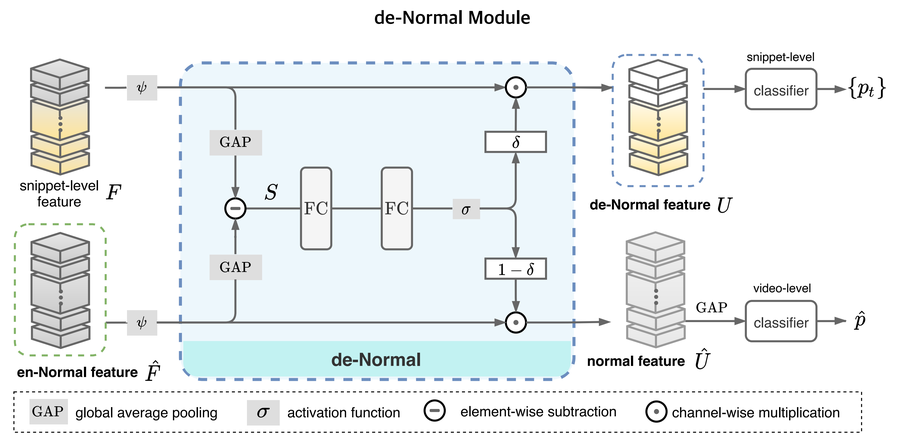

Figure: The illustration of the de-Normal module. This module takes the channel-wise difference between F and $\hat{F}$ to form the cross-video semantics S. Then, the channel scale δ is derived to depress S for describing normal events.

VISUALIZATION