JUNE 18-22, 2023

CVPR

VANCOUVER, CANADA

A New Comprehensive Benchmark for Semi-supervised Video Anomaly Detection and Anticipation

Congqi Cao, Yue Lu, Peng Wang, Yanning Zhang

ASGO, School of Computer Science, Northwestern Polytechnical University, China

Project Page: https://campusvad.github.io/

Paper Tag: THU-AM-372

西北工業大學

NORTHWESTERN POLYTECHNICAL UNIVERSITY

Quick Overview

NWPU Campus Dataset

Examples of the proposed dataset:

Walking (normal) | Climbing fence | Playing with water | Stealing | Stealing

Cycling (normal) | Cycling on footpath | Snatching bag | Protest | Group conflict

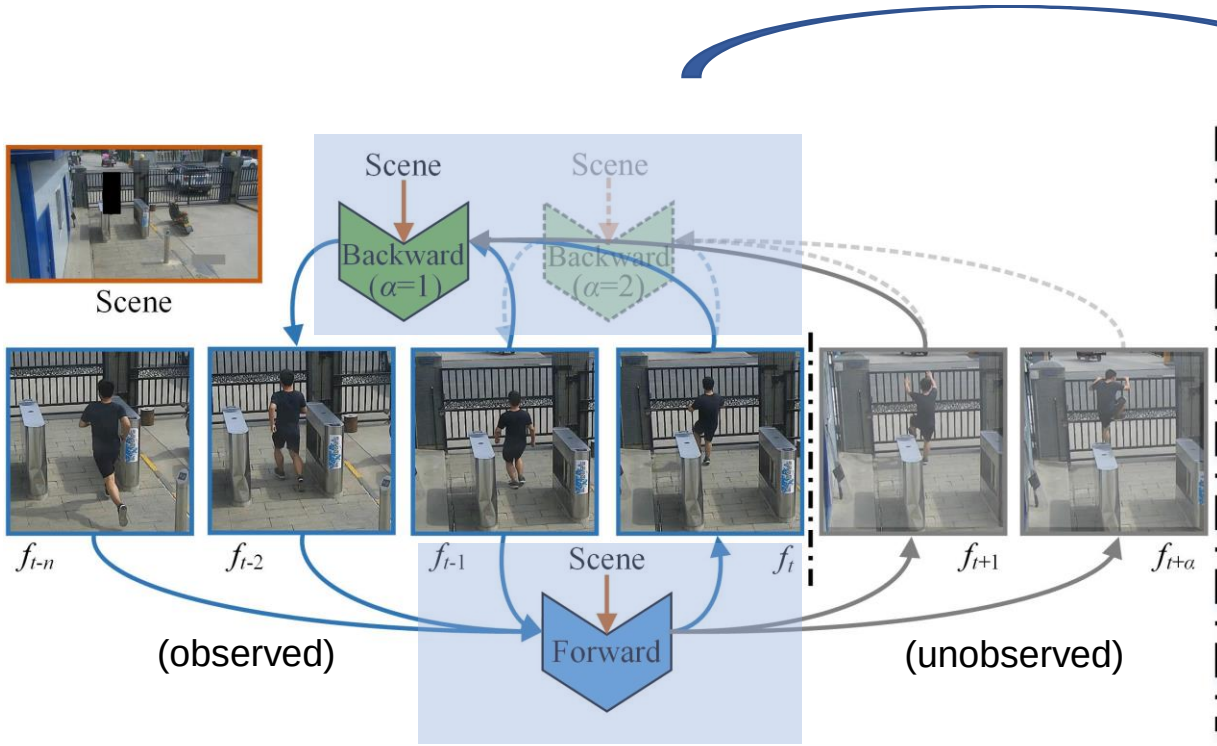
Driving (turning left, normal) | Wrong turn (turning left, abnormal) | Photographing in restricted area | Car crossing square | Dogs

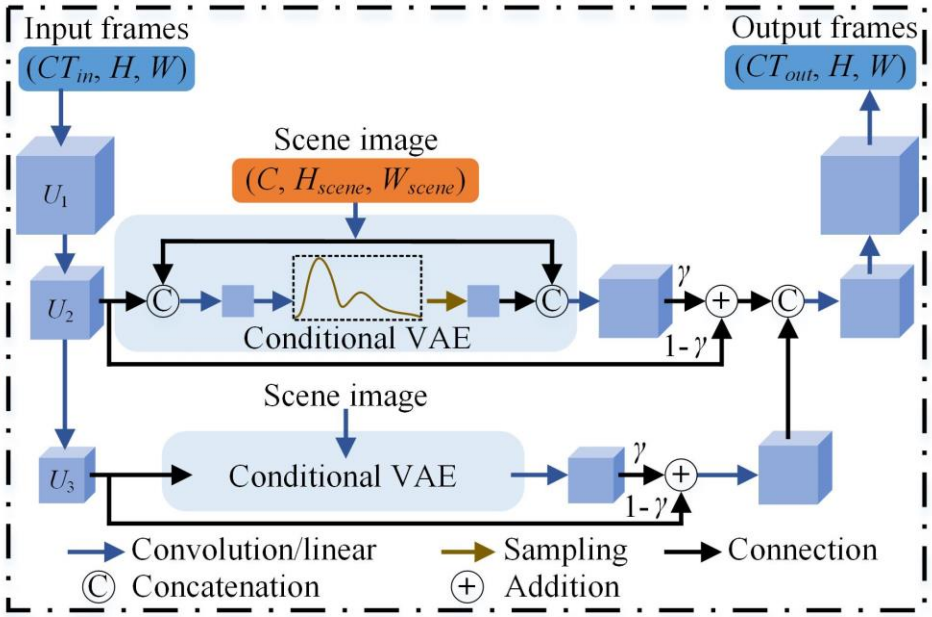
Comparison with other datasets:

Dataset	Year	# Frames			# Abnormal event classes	Resolution	#Scenes	Scene dependency
		Total	Training	Testing				
Subway Entrance [25]	2008	86,535	18,000	68,535	5	512×384	1	✗
Subway Exit [25]	2008	38,940	4,500	34,440	3	512×384	1	✗
UMN [26]	2009	7,741	-	-	3	320×240	3	✗
USCD Ped1 [22]	2010	14,000	6,800	7,200	5	238×158	1	✗
USCD Ped2 [22]	2010	4,560	2,550	2,010	5	360×240	1	✗
CUHK Avenue [27]	2013	30,652	15,328	15,324	5	640×360	1	✗
ShanghaiTech [23]	2017	317,398	274,515	42,883	11	856×480	13	✗
Street Scene [28]	2020	203,257	56,847	146,410	17	1280×720	1	✗
IITB Corridor [29]	2020	483,566	301,999	181,567	10	1920×1080	1	✗
UBnormal [24] *	2022	236,902	116,087	92,640	22	720p	29	✗
NWPU Campus	(ours)	**1,466,073**	**1,082,014**	**384,059**	**28**	multiple	**43**	✓

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA
NORTHWESTERN POLYTECHNICAL UNIVERSITY

Forward-backward Scene-conditioned Auto-encoder





Structure of the forward/backward network

Background

Video Anomaly Detection (VAD)

Video:

(From the ShanghaiTech dataset)

Anomaly Probability: 0.1 0.2 0.9 0.1 Time

Training set: only contains **normal** events
Testing set: contain both **normal and abnormal** events

Motivation

Widely Used VAD Datasets

Training examples:

Testing examples:

UCSD Ped2 | CUHK Avenue | ShanghaiTech | IITB Corridor

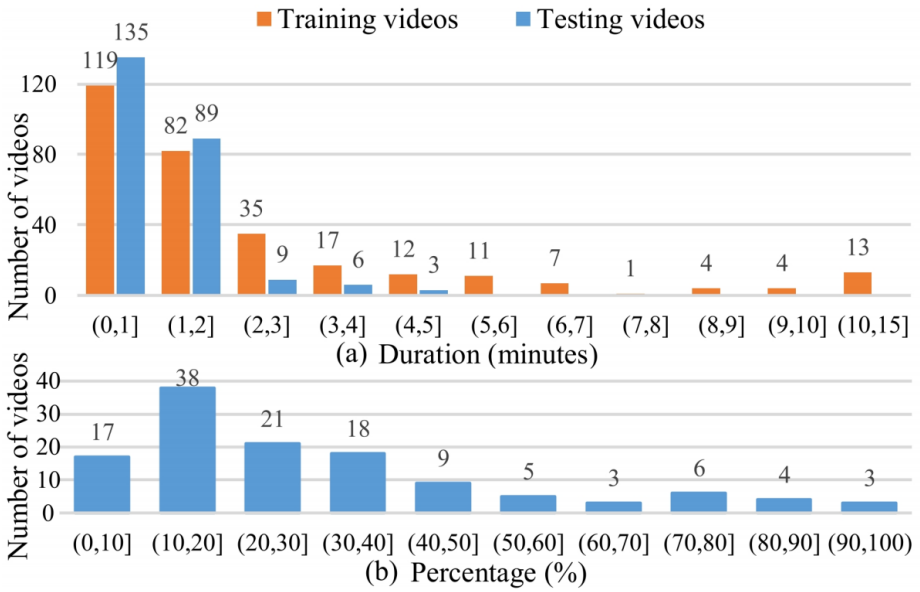
Our Proposed Dataset

NWPU Campus Dataset

Frame count and duration of the NWPU Campus dataset.

NWPU Campus (25 FPS)		
1,466,073 (16.29h)		
Training frames	Testing frames	
1,082,014 (12.02h)	384,059 (4.27h)	
Normal	Normal	Abnormal
1,082,014 (12.02h)	318,793(3.54h)	65,266(0.73h)

The distributions of training and testing videos (a) and abnormal testing videos (b).



The list of anomaly classes. “s.d.” stands for the scene-dependent anomaly.

Climbing fence	Car crossing square	Cycling on footpath (s.d.)	Kicking trash can
Jaywalking	Snatching bag	Crossing lawn	Wrong turn (s.d.)
Cycling on square	Chasing	Loitering	Scuffle
Littering	Forgetting backpack	U-turn	Battering
Driving on wrong side	Falling	Suddenly stopping cycling in the middle of the road	Group conflict
Climbing tree	Stealing	Illegal parking	Trucks (s.d.)
Protest	Playing with water	Photographing in restricted area (s.d.)	Dogs

Several samples from the proposed NWPU Campus dataset.

Our Proposed Dataset

NWPU Campus Dataset

Comparison with the existing semi-supervised VAD datasets.

Dataset	Year	# Frames			# Abnormal event classes	Resolution	#Scenes	Scene dependency
		Total	Training	Testing				
Subway Entrance [25]	2008	86,535	18,000	68,535	5	512×384	1	✗
Subway Exit [25]	2008	38,940	4,500	34,440	3	512×384	1	✗
UMN [26]	2009	7,741	-	-	3	320×240	3	✗
USCD Ped1 [22]	2010	14,000	6,800	7,200	5	238×158	1	✗
USCD Ped2 [22]	2010	4,560	2,550	2,010	5	360×240	1	✗
CUHK Avenue [27]	2013	30,652	15,328	15,324	5	640×360	1	✗
ShanghaiTech [23]	2017	317,398	274,515	42,883	11	856×480	13	✗
Street Scene [28]	2020	203,257	56,847	146,410	17	1280×720	1	✗
IITB Corridor [29]	2020	483,566	301,999	181,567	10	1920×1080	1	✗
UBnormal [24] *	2022	236,902	116,087	92,640	22	720p	29	✗
NWPU Campus	(ours)	**1,466,073**	**1,082,014**	**384,059**	**28**	multiple	**43**	✓

* represents the animated dataset

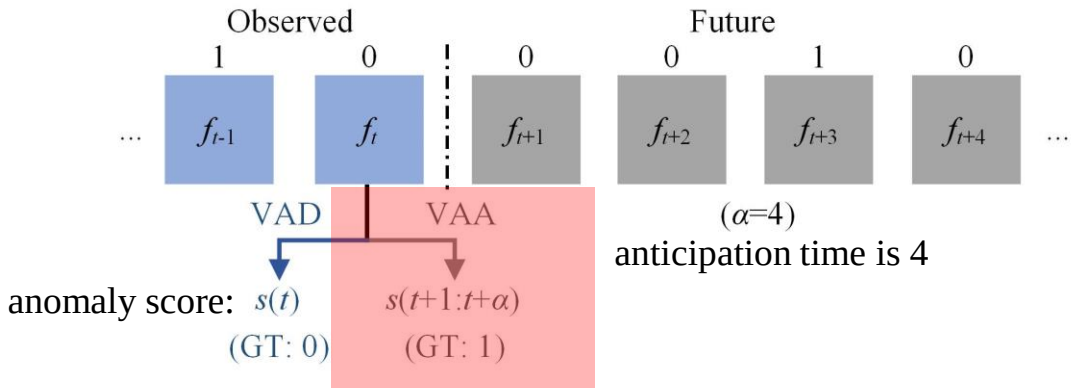
NWPU Campus:

- the largest dataset for video anomaly detection with the longest **duration**, the largest **number of classes of anomalies** and **scenes**
- the only one containing **scene-dependent anomalies**
- the first one proposed for semi-supervised **video anomaly anticipation**

Problem Formulation

Video Anomaly Anticipation (VAA)

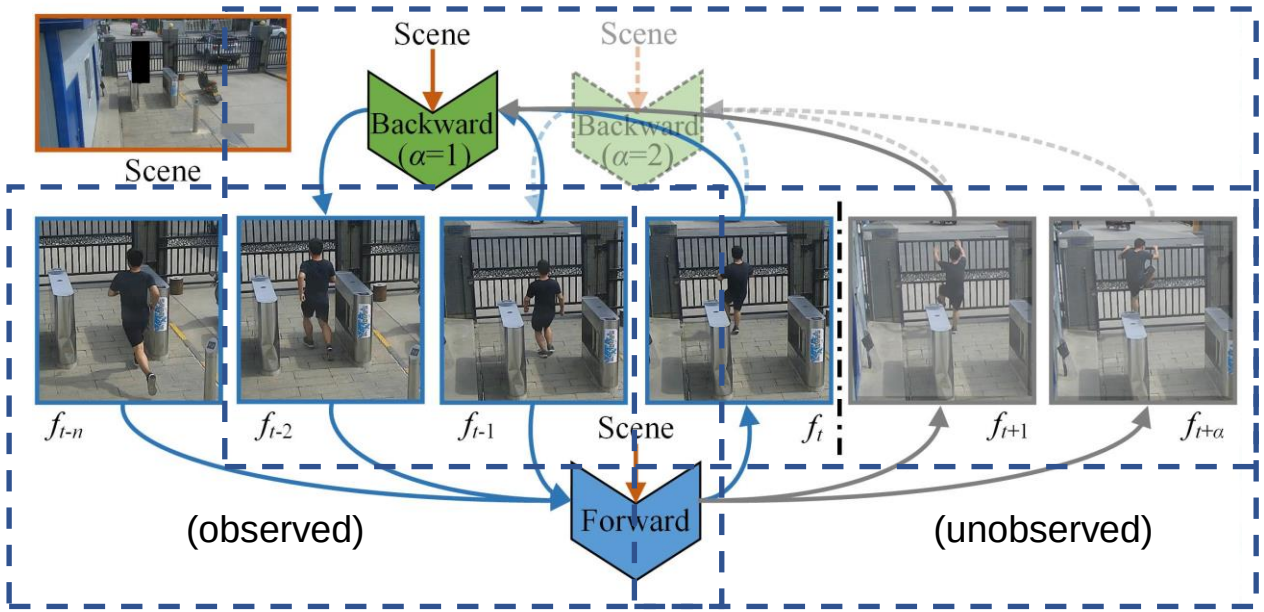
To anticipate whether an anomaly will occur in a future **period** of time.

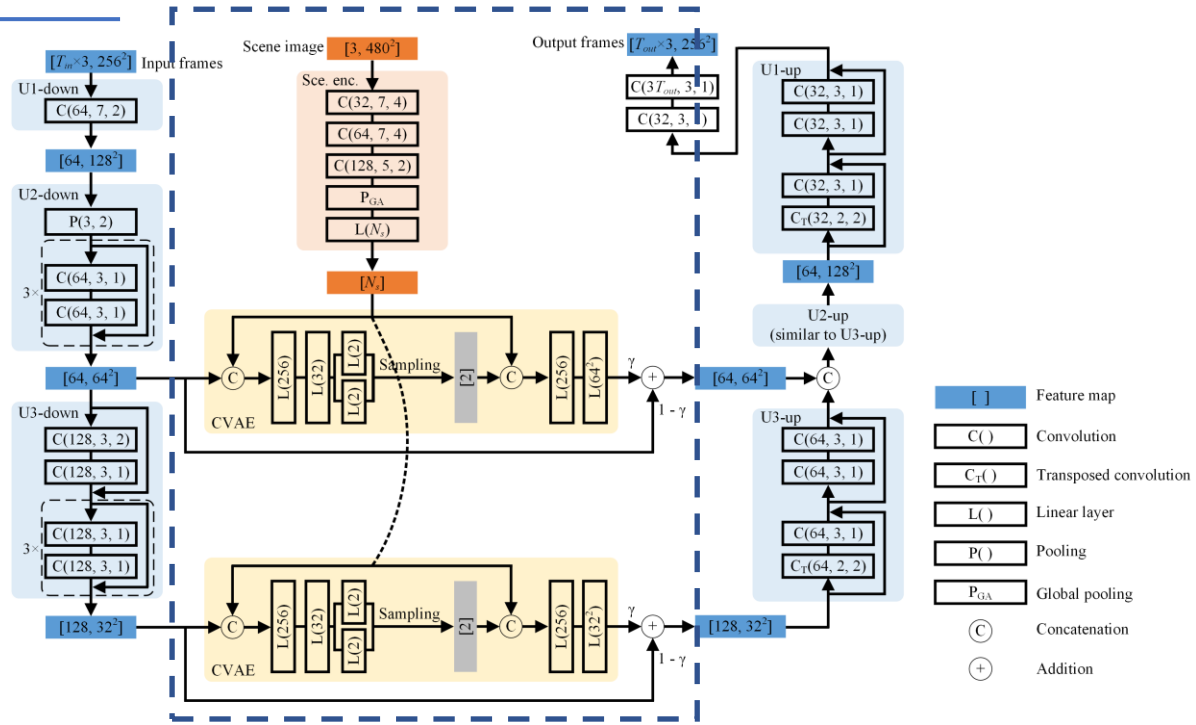


- No need to distinguish the exact boundary between normality and anomaly
- Be robust to potential boundary ambiguity

Method

Forward-backward Scene-conditioned Auto-encoder





Structure of the forward/backward network

Training:

Forward loss: $L_f(f, \hat{f}) = \|f - \hat{f}\|_2^2 + \lambda_{L1}|f - \hat{f}|$

Backward loss: $L_b = \frac{1}{2}(L_f(\hat{f}^{(1)}_{t+i-n}, f_{t+i-n}) + L_f(\hat{f}^{(2)}_{t+i-n}, f_{t+i-n}))$

CVAE loss: $L_{KL}(\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)\|\mathcal{N}(0,1)) = -\frac{1}{2}(\log\hat{\sigma}^2 - \hat{\mu}^2 - \hat{\sigma}^2 + 1)$

Inference:

Anomaly Detection: $s(t) = L_f(f_t, \hat{f}_t)$

Anomaly Anticipation: $s(t+1:t+\alpha) = \max(\{L_f(f_{t+i-n}, \hat{f}_{t+i-n})\}_{i=1}^{\alpha})$

Experiments

VAD Performance Benchmarking

Method	Year	ST	Ave	Cor	Cam
FFP [3]	CVPR 18	72.8	84.9	64.7	-
MemAE [4]	ICCV 19	71.2	83.3	-	61.9
MPED-RNN [46]	CVPR19	73.4	-	64.3	-
MTP [29]	WACV 20	76.0	82.9	67.1	-
VEC-AM [13]	ACM MM 20	74.8	89.6	-	-
CDDA [36]	ECCV 20	73.3	86.0	-	-
BMAN [9]	TIP 20	76.2	90.0	-	-
Ada-Net [7]	TMM 20	70.0	89.2	-	-
MNAD [6]	CVPR 20	70.5	88.5	-	62.5
OG-Net [39]	CVPR 20	-	-	-	62.5
CT-D2GAN [47]	ACM MM 21	77.7	85.9	-	-
ROADMAP [17]	TNNLS 21	76.6	88.3	-	-
MESDnet [18]	TMM 21	73.2	86.3	-	-
AMMC-Net [40]	AAAI 21	73.7	86.6	-	64.5
MPN [11]	CVPR 21	73.8	89.5	-	64.4
HF^2-VAD [10]	ICCV 21	76.2	**91.1**	-	63.7
SSAGAN [48]	TNNLS 22	74.3	88.8	-	-
DLAN-AC [49]	ECCV 22	74.7	89.9	-	-
LLSH [35]	TCSVT 22	77.6	87.4	73.5	62.2
VABD [21]	TIP 22	78.2	86.6	72.2	-
Ours	-	**79.2**	86.8	**73.6**	**68.2**

- Our method achieves leading performance
- Our dataset is more challenging

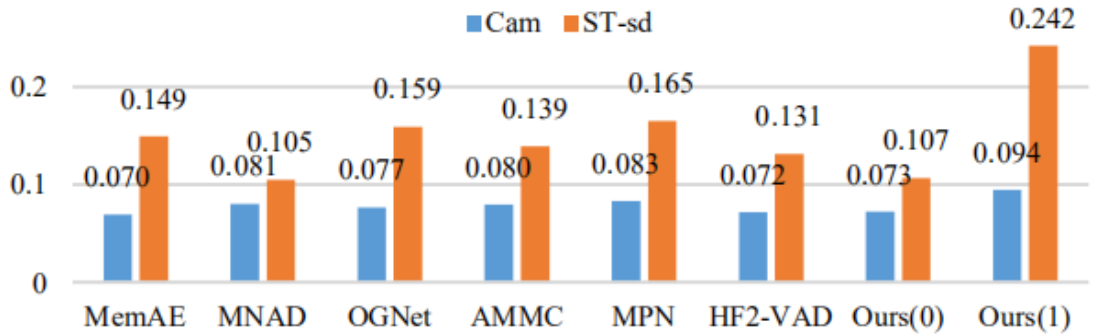
Experiments

Study on Scene-dependent Anomalies

Results on scene-dependent anomalous datasets.

Method	Cam	ST-sd (reorganized)
MemAE [4]	61.9	67.4
MNAD [6]	62.5	68.2
OG-Net [39]	62.5	69.6
AMMC-Net [40]	64.5	64.9
MPN [11]	64.4	76.9
HF^2-VAD [10]	63.7	70.8
Ours (γ=0)	65.8	70.4
Ours (γ=1)	**68.2**	**82.7**

Score gaps of different methods (the higher, the better).



Scene-conditioned VAE: +2.4% and +12.3%

The score gap is much higher than other methods.

Experiments

Video Anomaly Anticipation

Results for video anomaly anticipation with different anticipation times.

α_t	0.5s	1.0s	1.5s	2.0s	2.5s	3.0s
Chance	50.0	50.0	50.0	50.0	50.0	50.0
Human	-	-	-	-	-	90.4
Ours (f-only)	65.2	64.6	64.2	63.6	63.1	62.5
Ours (f+b)	65.8	65.3	64.9	64.6	64.2	64.0

Forward-backward prediction is helpful for anomaly anticipation.

Conclusion

- We propose **a large-scale and challenging dataset** NWPU Campus for semi-supervised video anomaly detection and anticipation.

- We propose **a forward-backward scene-conditioned model** for VAD and VAA as well as handling scene-dependent anomalies.

- The proposed model achieves **leading performance on VAD**. It can also cope with **scene-dependent anomalies** and **anomaly anticipation**.

Project Page: https://campusvad.github.io/