Supplementary Material: 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
congqi.cao@nwpu.edu.cn zugexiaodui@mail.nwpu.edu.cn {peng.wang, ynzhang}@nwpu.edu.cn

Table 1. The anomaly classes in each scene of our NWPU Campus dataset.

Index	Scene	Anomalies	Index	Scene	Anomalies
1	D001	Climbing fence	23	D109	Loitering, Protest, Falling, Stealing
2	D002	Jaywalking	24	D111	Crossing lawn
3	D003	Dogs, Trucks	25	D121	Playing with water
4	D013	Cycling on square	26	D122	Playing with water
5	D014	Climbing fence	27	D124	Playing with water
6	D029	Battering, Group conflict	28	D127	Playing with water
7	D031	Group conflict, Battering, Climbing tree, Chasing, Littering	29	D129	Playing with water, Cycling on footpath
8	D035	Illegal parking, Jaywalking	30	D148	Forgetting backpack
9	D036	Climbing tree, Photographing in restricted area	31	D149	Stealing
10	D038	Jaywalking, Photographing in restricted area	32	D150	Stealing, Forgetting backpack
11	D042	Scuffle	33	D151	Stealing, Forgetting backpack
12	D043	Group conflict	34	D154	Suddenly stopping cycling in the middle of the road, Protest
13	D047	Suddenly stopping cycling in the middle of the road, U-turn, Driving on wrong side	35	D155	Protest
14	D048	Forgetting backpack	36	D158	Chasing, Loitering, Falling, Stealing
15	D054	U-turn	37	D164	Kicking trash can
16	D055	U-turn, Suddenly stopping cycling in the middle of the road, Driving on wrong side	38	D235	Cycling on square
17	D068	Climbing fence, Snatching bag	39	D236	Driving on wrong side
18	D076	Group conflict, Snatching bag	40	D248	Scuffle, Forgetting backpack
19	D077	Group conflict, Snatching bag	41	D268	Driving on wrong side
20	D092	Protest, Car crossing square	42	D273	Climbing fence
21	D094	Protest	43	D282	Loitering
22	D099	Kicking trash can			

1. NWPU Campus Dataset

The 28 anomaly classes in our NWPU Campus dataset are listed in Tab. 2. As can be seen, there are 4 classes of scene-dependent anomalies, *i.e.*, cycling on footpath, wrong turn, photographing in restricted area, and trucks. It should be noted that scene-dependent anomaly is different from location anomaly. Location anomaly means that whether an event is normal or not is determined by the location where

†Corresponding author

Table 2. The list of anomaly classes in our NWPU Campus dataset. ”s.d.” stands for a 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

D013 D129

Figure 1. The scenes ”D013” and ”D129” in our NWPU Campus dataset.

it occurs in a scene. For example, as shown in Fig. 1, in the scene ”D013” of our dataset, cycling on road is normal, while cycling on square is anomalous. Therefore, cycling on square is a location anomaly rather than a scene-dependent anomaly. As to scene-dependent anomaly, once a certain event occurs in the scene where it is not allowed, the event is considered to be abnormal regardless of its location. Hence, cycling on footpath in the scene ”D129” is a scene-dependent anomaly. In our dataset, we try to cover as many types of anomalies as possible to study the performances of algorithms for different anomalies, especially the scene-dependent anomaly.

The anomaly classes in each scene are shown in Tab. 1. We take photographing in restricted area as an example to demonstrate the scene-dependent anomaly in our dataset. In the scenes ”D036” and ”D038”, the training videos do not contain photographing, and thus photographing in the testing videos of these two scenes is anomalous. In contrast, although there are behaviors of photographing in the scene ”D055”, they are regarded as normal events since both training and testing videos include photographing.

2. Model details

We illustrate the details of our forward and backward networks which share the same structure in Fig. 2. The encoder of U-Net is based on ResNet-34 [4]. The U2-down and U3-down blocks output feature maps with spatial sizes of 64×64 and 32×32 pixels respectively. The scene image is fed into the scene encoder to generate a scene encoding of length N_s. We set N_s as the number of scenes in a dataset. To reduce computational complexity and avoid overfitting, we train the conditional variational auto-encoder (CVAE) to reconstruct feature maps in channel wise. That is, for the feature maps output by U2-down/U3-down, we take the feature map in each channel as an independent instance, and feed it into the CVAE after concatenated with the scene encoding. For each input feature map, the CVAE samples a latent variable vector of length 2 from posterior distribution by reparameterization technique [9], concatenates it with the scene encoding, and decodes it to reconstruct the feature map. The decoder of U-Net gradually increases the spatial size of feature maps by transposed convolution and finally outputs the predicted frames.

In our network, all layers adopt batch normalization [6] except the CVAEs which use layer normalization [1]. We use ReLU [13] activation function for all the convolutional layers and Leaky ReLU [12] for the transposed convolutional layers. The activation function for the linear layers in CVAEs is GELU [5], while it is softmax for the last linear layer in the scene encoder. We do not use normalization and non-linear activation in the last layer of the U-Net decoder.

We sample frames with a sampling rate of 2 on the ShanghaiTech, CUHK Avenue as well as IITB Corridor datasets. On our NWPU Campus dataset, we use the sampling rate of 12.5, which is consistent with the interval of anticipation times (*i.e.* 0.5s) in the VAA task. For non-integer frame positions, we use rounding to select its nearest frame. The scene image is resized to 480×480 pixels before fed into the scene encoder.

In the training phase, we first train the scene encoder with cross-entropy loss via scene classification. Then its parameters are frozen to train the whole network. We use Adam [8] optimizer to train our network for 80 epochs with a learning rate of 0.01, which is decayed to 0.001 at the 50th epoch. In the testing phase, the frame-level anomaly score is obtained from the maximum score of the objects on that frame. Following previous works [7, 10, 11], we adopt a Gaussian filter to smooth the frame-level anomaly scores.

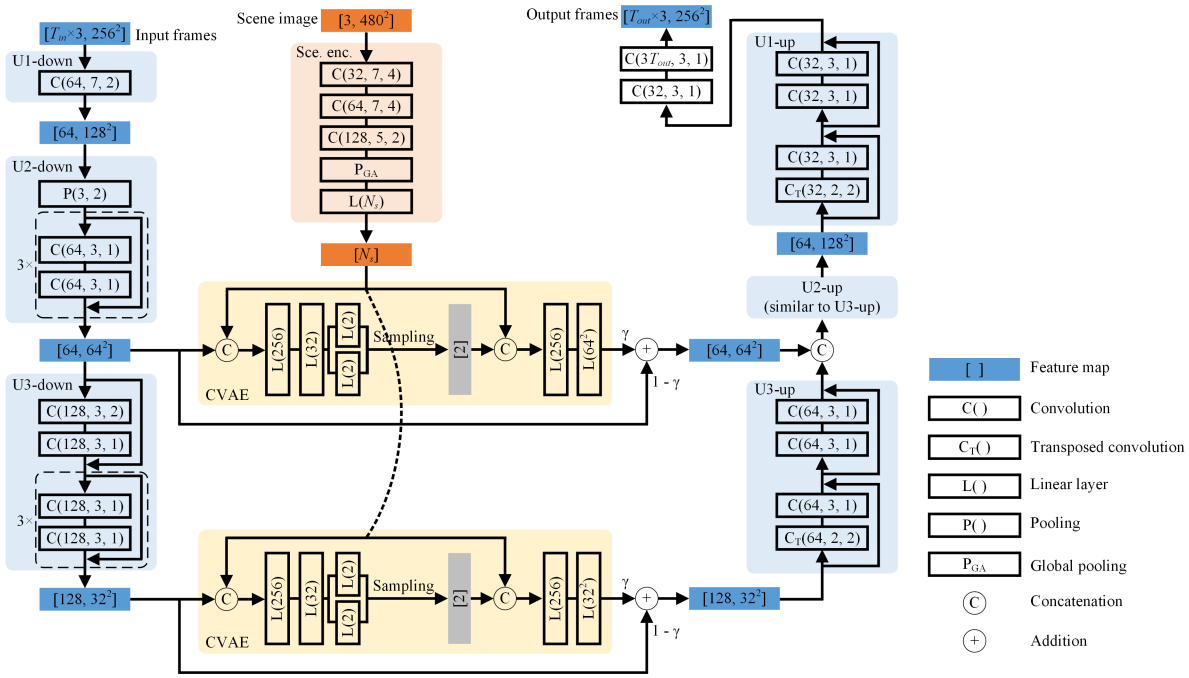

Figure 2. The specific structure of our network. For a feature map or image with shape [$channel$, $height$, $width$], we denote its dimensions as [$channel$, $height \times width$] for convenience, where $height$ and $width$ are always equal. [N_s] represents a vector of length N_s. Convolutional layers / transposed convolutional layers are represented by $\mathrm{C}/\mathrm{C_T}(output\ channel, kernel\ size, stride)$. $\mathrm{L}(output\ channel)$ represents linear layers. $\mathrm{P}(kernel\ size, stride)$ represents the max pooling layer. $\mathrm{P_{GA}}$ denotes the global average pooling layer. The normalization and non-linear activation layers, as well as the 1×1 convolutional layers that used in certain residual connections are not displayed in the figure. "Sce. enc." denotes the scene encoder.

Table 3. Training videos and testing videos of the ShanghaiTech-sd dataset. The number before the underscore represents the scene number.

Training videos (35)			Testing videos (20)	
01_0016	06_004	10_010	01_0014	06_0155
01_0029	06_005	10_011	01_0027	10_0037
01_0063	06_007	12_002	01_0051	10_0074
01_0073	06_008	12_003	01_0052	12_0142
01_0076	06_009	12_004	01_0053	12_0148
01_0129	06_014	12_005	01_0138	12_0151
01_0131	10_001	12_006	01_0139	12_0154
01_0134	10_002	12_007	01_0163	12_0173
01_0177	10_006	12_008	06_0147	12_0174
06_001	10_007	12_009	06_0150	12_0175
06_002	10_008	12_015		
06_003	10_009			

3. Study on Scene-dependent Anomalies

3.1. ShanghaiTech-sd Dataset

The training videos and testing videos of the ShanghaiTech-sd datasets reorganized by us are shown in Tab. 3. In the training videos, the scene "01" contains "cycling" events picked from the testing set of the original ShanghaiTech dataset, while cycling is not included in other scenes. All the scenes in the testing set contain cycling. However, cycling in the scene "01" is a normal behavior, while it is an abnormal behavior in other scenes.

3.2. Visualization

To qualitatively study the effect of our scene-conditioned VAE in terms of scene-dependent anomalies, we visualize the score curves of our method in Fig. 3. Note that since the range of anomaly scores varies from method to method, we first scale the scores to the range of $[0, 1]$ for each method.

In the "12_0175" video of ShanghaiTech-sd and the "D036_08" video of NWPU Campus, cycling and photographing are anomalous events, respectively. Since the training sets contain cycling and photographing in other scenes, the powerful representation capacity of CNNs allows our model without scene-conditioned VAEs (*i.e.* γ=0) to predict frames accurately, resulting in low anomaly scores. In contrast, using scene-conditioned VAEs (*i.e.* γ=1) can increase the frame prediction error for these two behaviors in the corresponding scene, which generates significantly higher anomaly scores and thus identifies scene-dependent anomalies. We find that this is the main way

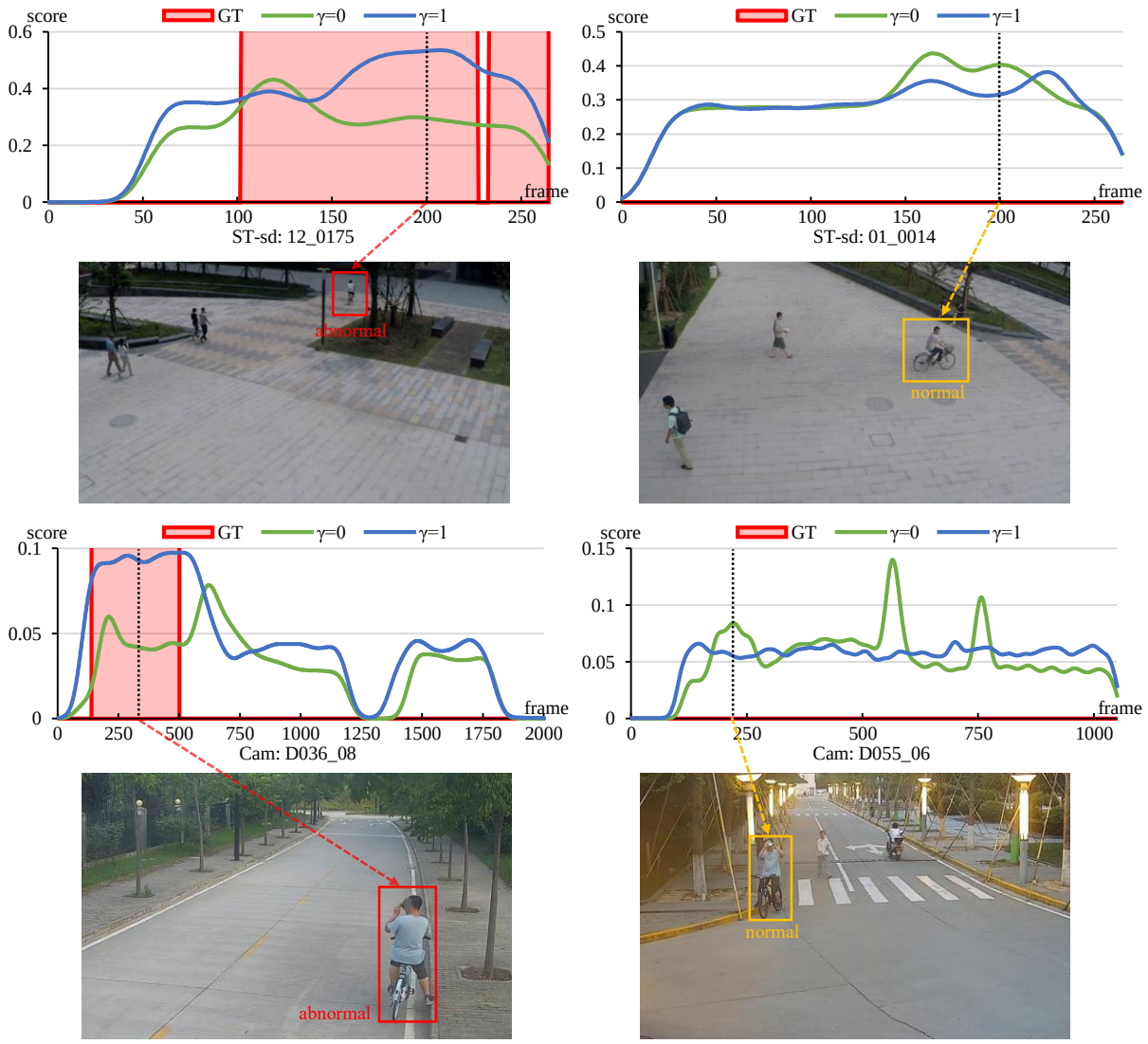

Figure 3. Visualization of score curves on ShanghaiTech-sd and NWPU Campus datasets. γ is the hyper-parameter in our model to control whether to use the proposed scene-conditioned VAE. "GT" stands for the groundtruth. A higher score represents a higher probability of anomaly.

in which our proposed scene-conditioned VAE works. Additionally, the scene-conditioned VAE can also reduce the frame prediction error in the case of normal events. It can be seen that in the "01_0014" video of ShanghaiTech-sd and the "D055_06" video of NWPU Campus where cycling and photographing are respectively normal events, the model with scene-conditioned VAEs outputs lower abnormal scores, which can reduce false alerts. Overall, the proposed scene-conditioned VAE is able to reconstruct the events not appear in the training set of the scene with large errors, so as to detect scene-dependent anomalies.

3.3. Number of Input Frames

The number of input and output frames usually has an effect to the model. In our experiments, we just follow the common setting in action recognition [3, 15, 16] and use 8 frames as the input for our model. We make an analysis of the number of input frames, as shown in Fig. 4. It can be seen that the performance of our model is not over sensitive to this hyperparameter. When T_{in}=12 and T_{out}=11, our model obtains higher results on ShanghaiTech (79.8%) and NWPU Campus (68.4%).

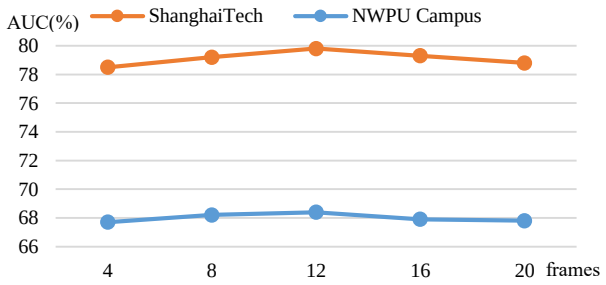

Figure 4. Effect of the number of frames.

3.4. Computational Complexity

The MACs of our forward and backward networks are 11.2G and 10.9G, respectively. Note that we only need the forward network for VAD. The MACs of representative methods MNAD [14], AMMC-Net [2], MPN [11] and HF^2-VAD [10] are 57.5G, 93.9G, 55.0G and 1.8G, respectively. Since HF^2-VAD resizes the input images to a smaller scale, it has lower MACs. However, compared to other models with the same input resolution, our model is much more lightweight.

3.5. Discussion of Ethics

All the participants are informed and consent to the release of the dataset. Besides, we have masked the personal information such as faces and license plates in the dataset. To reduce bias, we consider anomalous events and their manifestations as comprehensive as possible to represent the diversity of the real situation. Meanwhile, face masking also reduces the appearance bias against minority groups.

The positive societal impact is that the proposed dataset can provide data and test-bed for detecting and anticipating harmful behaviors, thus protecting people's lives and property. The possible negative social impact is that the performed anomalous behaviors could be imitated. However, we have explicitly warn against imitating those behaviors when releasing the dataset.

References

[1] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. *CoRR*, abs/1607.06450, 2016. 2

[2] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. Appearance-Motion Memory Consistency Network for Video Anomaly Detection. In *AAAI*, pages 938–946, 2021. 5

[3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 6201–6210, 2019. 4

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[5] Dan Hendrycks and Kevin Gimpel. Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units. *CoRR*, abs/1606.08415, 2016. 2

[6] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*, volume 37, pages 448–456, 2015. 2

[7] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-Centric Auto-Encoders and Dummy Anomalies for Abnormal Event Detection in Video. In *CVPR*, pages 7834–7843, 2019. 2

[8] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 2

[9] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014. 2

[10] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *ICCV*, pages 13588–13597, 2021. 2, 5

[11] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *CVPR*, pages 15425–15434, 2021. 2, 5

[12] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, volume 30, page 3, 2013. 2

[13] Vinod Nair and Geoffrey E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *ICML*, pages 807–814, 2010. 2

[14] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning Memory-Guided Normality for Anomaly Detection. In *CVPR*, pages 14360–14369, 2020. 5

[15] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2740–2755, 2019. 4

[16] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *European Conference on Computer Vision*, 2018. 4