2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-0129-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPR52729.2023.01953

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

Abstract

Semi-supervised video anomaly detection (VAD) is a critical task in the intelligent surveillance system. However, an essential type of anomaly in VAD named scene-dependent anomaly has not received the attention of researchers. Moreover, there is no research investigating anomaly anticipation, a more significant task for preventing the occurrence of anomalous events. To this end, we propose a new comprehensive dataset, NWPU Campus, containing 43 scenes, 28 classes of abnormal events, and 16 hours of videos. At present, it is the largest semi-supervised VAD dataset with the largest number of scenes and classes of anomalies, the longest duration, and the only one considering the scene-dependent anomaly. Meanwhile, it is also the first dataset proposed for video anomaly anticipation. We further propose a novel model capable of detecting and anticipating anomalous events simultaneously. Compared with 7 outstanding VAD algorithms in recent years, our method can cope with scene-dependent anomaly detection and anomaly anticipation both well, achieving state-of-the-art performance on ShanghaiTech, CUHK Avenue, IITB Corridor and the newly proposed NWPU Campus datasets consistently. Our dataset and code is available at: https://campusvad.github.io.

1. Introduction

Video anomaly detection (VAD) is widely applied in public safety and intelligent surveillance due to its ability to detect unexpected abnormal events in videos. Since anomalous events are characterized by unbounded categories and rare occurrence in practice, VAD is commonly set as a semi-supervised task, that is, there are only normal events without specific labels in the training set [1, 2]. The model trained only on the normal events needs to distinguish anomalous events from normal events in the testing phase.

Semi-supervised VAD has been studied for years. Especially in recent years, reconstruction-based and prediction-based methods [3–21] have made leaps and bounds in performance on existing datasets. For example, the frame-level AUCs (area under curve) on UCSD Ped1 and Ped2 datasets [22] have reached over 97% [2]. Despite the emergence of a few challenging datasets, researchers still overlook an important type of anomaly, *i.e.*, the scene-dependent anomaly [2]. Scene dependency refers to that an event is normal in one scene but abnormal in another. For example, playing football on the playground is a normal behavior, but playing on the road is abnormal. Note that single-scene datasets cannot contain any scene-dependent anomaly. Nevertheless, the existing multi-scene datasets (*e.g.*, ShanghaiTech [23], UBnormal [24]) also have not taken this type of anomaly into account. As a result, there is currently no algorithm for studying scene-dependent anomaly detection, limiting the comprehensive evaluation of VAD algorithms. In addition to detecting various types of anomalies, we argue that there is another task that also deserves the attention of researchers, which is to anticipate the occurrence of abnormal events in advance. If we can make an early warning before the anomalous event occurs based on the trend of the event, it is of great significance to prevent dangerous accidents and avoid loss of life and property. However, according to our investigation, there is no research on video anomaly anticipation, and no dataset or algorithm has been proposed for this field.

意思是在一个数据集中，同一个行为，在一个场景下正常，在另一个场景下异常？

就是单分类

In this paper, we work on semi-supervised video anomaly detection and anticipation. First and foremost, to address the issue that the VAD datasets lack scene-dependent anomalies and are not suitable for anomaly anticipation, we propose a new large-scale dataset, NWPU Campus. Compared with existing datasets, our proposed dataset mainly has the following three advantages. First, to the best of our knowledge, the NWPU Campus is the largest semi-supervised VAD dataset to date. It contains 43 scenes, whose number is 3 times that of ShanghaiTech, the real recorded dataset with the largest number of scenes among the existing datasets. The total video duration of the NWPU Campus is 16 hours, which is more than 3 times that

[†] Corresponding author

979-8-3503-0129-8/23/$31.00 ©2023 IEEE
DOI 10.1109/CVPR52729.2023.01953

Table 1. Comparisons of different semi-supervised VAD datasets. There are not any official training and testing splits in UMN. UBnormal has a validation set, which is not shown here. ”720p” means that the frame is 720 pixels high and 1280 or 1080 pixels wide. The frame resolutions of NWPU Campus are 1920×1080, 2048×1536, 704×576 and 1280×960 pixels. * represents the animated dataset.

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**	✓

of the existing largest semi-supervised VAD dataset IITB Corridor [29]. The quantitative comparison between the NWPU Campus and other datasets can be seen in Tab. 1. Second, the NWPU Campus has a variety of abnormal and normal events. In terms of anomalies, it contains 28 classes of anomalous events, which is more than any other dataset. Fig. 1 displays some examples from our dataset. More importantly, the NWPU Campus dataset contains scene-dependent anomalous events, which are missing in other datasets. As an example, the behavior of a vehicle turning left is anomalous in the scene where left turns are prohibited, while it is normal in other unrestricted scenes. Along with the diversity of anomalous events, the normal events in our dataset are diverse as well. Unlike other datasets, we do not only take walking and standing as normal behaviors. In our dataset, regular walking, cycling, driving and other daily behaviors that obey rules are also considered as normal events. Third, in addition to being served as a video anomaly detection benchmark, the NWPU Campus is the first dataset proposed for video anomaly anticipation (VAA). The existing datasets do not deliberately consider the anomalous events applicable to anticipation. In contrast, we take into account the complete process of the events in the data collection phase so that the occurrence of abnormal events is predictable. For instance, before the vehicle turns left (the scene-dependent anomalous event as mentioned before), the movement trend of it can be observed, and hence the algorithm could make an early warning. As a comparison, it is considered to be abnormal when a vehicle simply appears in the ShanghaiTech dataset, which is unpredictable and therefore not suitable for anomaly anticipation.

Besides comprehensive benchmarks, there is currently a lack of algorithms for scene-dependent anomaly detection and video anomaly anticipation. Therefore, in this work, we further propose a novel forward-backward frame prediction model that can detect anomalies and simultaneously anticipate whether an anomalous event is likely to occur in the future. Moreover, it has the ability to handle scene-dependent anomalies through the proposed scene-conditioned auto-encoder. As a result, our method achieves state-of-the-art performance on ShanghaiTech [23], CUHK Avenue [27], IITB Corridor [29], and our NWPU Campus datasets.

In summary, our contribution is threefold:

- We propose a new dataset NWPU Campus, which is the largest and most complex semi-supervised video anomaly detection benchmark to date. It makes up for the lack of scene-dependent anomalies in the current research field.
- We propose a new video anomaly anticipation task to anticipate the occurrence of anomalous events in advance, and the NWPU Campus is also the first dataset proposed for anomaly anticipation, filling the research gap in this area.
- We propose a novel method to detect and anticipate anomalous events simultaneously, and it can cope with scene-dependent anomalies. Comparisons with 7 state-of-the-art VAD methods on the NWPU Campus, ShanghaiTech, CUHK Avenue and IITB Corridor datasets demonstrate the superiority of our method.

2. Related Work

2.1. Video Anomaly Detection Datasets

We focus on semi-supervised video anomaly detection in this paper, so the weakly-supervised video anomaly detection datasets such as UCF-Crime [30] and XD-Violence

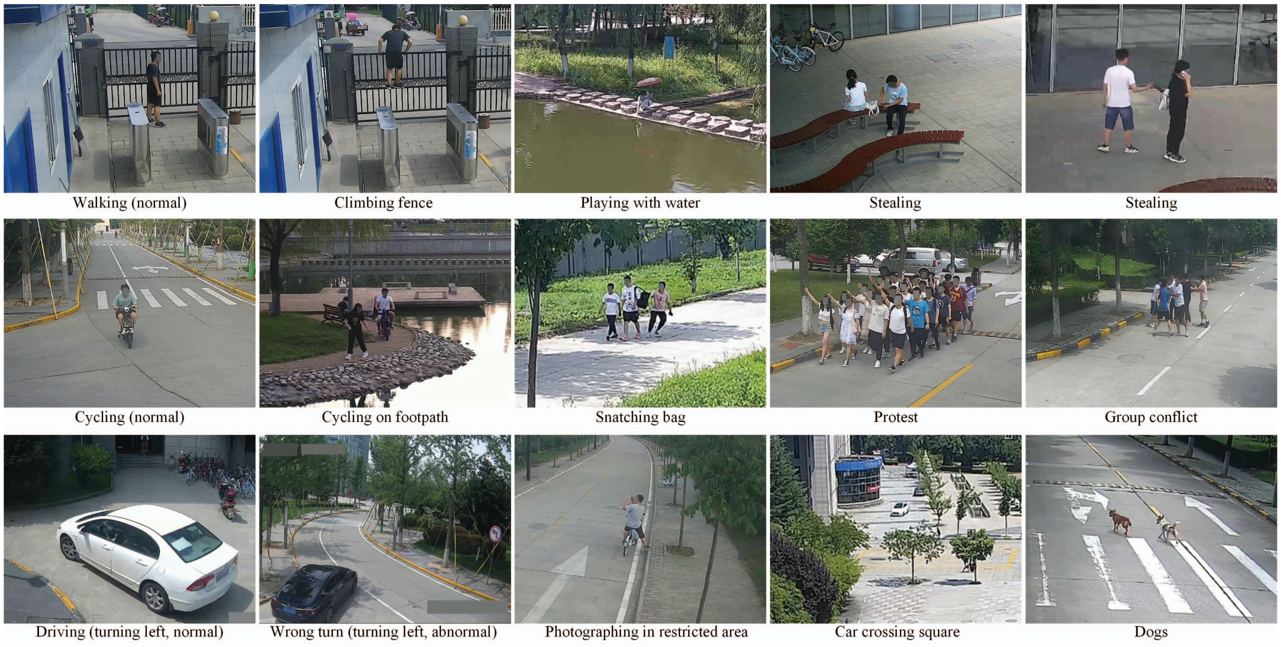

Figure 1. Samples from the proposed NWPU Campus dataset. The samples in the first column are normal events, while the others are different types of anomalous events.

[31] will not be discussed. The commonly used semi-supervised VAD datasets include USCD Ped1 & Ped2 [22], Subway Entrance & Exit [25], UMN [26], CUHK Avenue [27], ShanghaiTech [23], Street Scene [28], IITB Corridor [29] and UBnormal [24].

The UCSD Ped1 & Ped2 [22] datasets each contain a camera overlooking a pedestrian walkway, in which most of the anomalies are intrusions of other objects, such as bicycles, cars and skateboards. Therefore, the anomalies can be readily detected through static images, resulting in the saturation of performance (97.4% [32] on Ped1 and 99.2% [33] on Ped2 in frame-level AUC). The Subway Entrance & Exit [25] datasets include two indoor scenes of the subway entrance and exit. The abnormal events are only related to people, including jumping through turnstiles, wrong direction, *etc*. The UMN [26] contains three outdoor scenes, and the only type of anomalous event is the crowd dispersing suddenly. There are not any official training and testing splits in it. The CUHK Avenue [27] contains a camera looking at the side of a building with pedestrian walkways by it, and the abnormal behaviors include running, throwing bags, child skipping, *etc*. The ShanghaiTech [23] includes a total of 13 outdoor scenes on the campus, and quite a few of the anomalous events are related to objects, such as bicycles, cars, skateboards and strollers, even though it seems normal for these objects to appear in real life. The anomalous events in ShanghaiTech are generic across scenes and this dataset does not contain scene-dependent anomalies. The Street Scene [28] contains a camera looking down on a scene of a two-lane street with bike lanes and pedestrian sidewalks. Compared with previous datasets, it includes location anomalies, such as cars parked illegally and cars outside a car lane. The IITB Corridor [29] is the largest single-scene semi-supervised VAD dataset as far as we know. The scene consists of a corridor where the normal activities are walking and standing, and the anomalous behaviors are performed by volunteers, including chasing, fighting, playing with ball, *etc*. The UBnormal [24] is generated by animations, containing a total of 22 types of abnormal events in 29 virtual scenarios. However, there is a distribution gap between animated videos and real recorded videos. Acsintoae *et al*. [24] have to use an additional model (CycleGAN [34]) to reduce the distribution gap.

It should be noted that all of the above datasets do not take scene-dependent anomalies and anomaly anticipation into account. Therefore, a benchmark for the comprehensive evaluation of anomaly detection and anticipation is pressingly needed in the current research stage. The proposed NWPU Campus dataset has the features of large scale, multiple scenarios, and diverse as well as extensive events. It is committed to meeting the requirement for a new comprehensive benchmark.

2.2. Video Anomaly Detection Methods

Prevalent semi-supervised VAD methods mainly contain distance-based [35–37], reconstruction-based [4, 5, 38, 39]

Table 2. 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)

and prediction-based [3, 6–21] methods. Especially, the prediction-based methods have attracted wide attention in recent years. They usually predict the current (*i.e.*, the last observed) frame via previous frames [3, 6–8, 10–18, 20, 21], and compute anomaly score based on the error between the predicted frame and the observable groundtruth frame. To discriminate between abnormal and normal motion, some methods [10, 21] use optical flow as the condition of conditional VAE to enhance frame prediction. Combining with memory modules [4, 6, 10, 11, 40] that can explicitly utilize normal patterns is also an improvement trend of this kind of methods. Besides predicting the current frame, there are a few prediction-based methods completing the middle frame with bidirectional frame prediction [9, 19], which requires the observation of groundtruth frames in both directions.

Different from those prediction-based models, our forward-backward prediction model does not need to observe future frames during inference. It can estimate the prediction error of future frames whose groundtruth frames are unavailable, making it able to anticipate anomalies.

这里怎么做呢？有点儿意思。

3. Proposed Dataset

3.1. Dataset Collection

We set up cameras at 43 outdoor locations on the campus to record the activities of pedestrians and vehicles. As anomalous events rarely occur in real life, there are a total of more than 30 volunteers performing a part of normal and abnormal events. In our dataset, the classes of normal events include regular walking, cycling, driving and other daily behaviors that obey rules. The types of anomalies consist of single-person anomalies (*e.g.*, climbing fence, playing with water), interaction anomalies (*e.g.*, stealing, snatching bag), group anomalies (*e.g.*, protest, group conflict), scene-dependent anomalies (*e.g.*, cycling on footpath, wrong turn, photographing in restricted area), location anomalies (*e.g.*, car crossing square, crossing lawn), appearance anomalies (*e.g.*, dogs, trucks) and trajectory anomalies (*e.g.*, jaywalking, u-turn). Some normal and abnormal samples are shown in Fig. 1. There are different manifestations for each kind of anomalous event in our dataset. For instance, stealing may occur when two people are sitting next to each other or when one person is following another. Additionally, to avoid algorithms detecting anomalies according to specific performers, the volunteers also perform normal behaviors that are similar to the anomalous behavior if possible. For example, the normal behavior served as a contrast to climbing fence is merely walking up to the fence and then leaving.

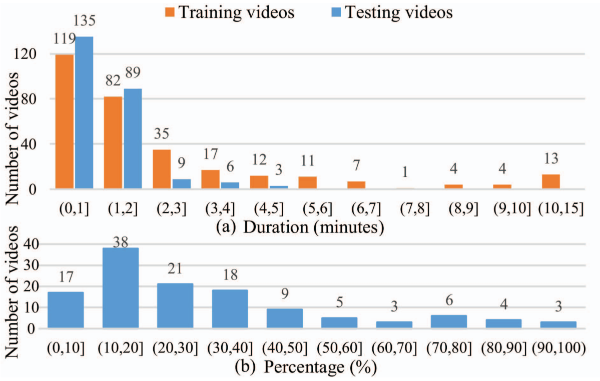

Figure 2. The distributions of training and testing videos according to duration (a), and abnormal testing videos according to the percentage of abnormal frames in each video (b).

Finally, we collect 16 hours of videos from these 43 scenes, including 305 training videos and 242 testing videos. In the training data, there are only normal events that come from real events (without volunteers) and performed events (with volunteers), while the testing data contains both normal events and anomalous events. In the testing set, there are a total of 28 classes of abnormal events, most of which are performed by volunteers and some actually occur. All the anomaly classes and the anomaly classes in each scene are provided in the supplementary material. We annotate frame-level labels for the testing videos to indicate the presence or absence of anomalous events in each frame. According to the setting of semi-supervised VAD, algorithms only need to distinguish abnormality from normality. Thus, the specific classes of the abnormal events are not annotated. It should be noted that not all the testing videos contain anomalies, since there is no guarantee that an anomalous event will certainly happen in a video in practical applications. To protect the privacy of volunteers and pedestrians, all the faces in our dataset are blurred.

3.2. Dataset Statistics

The statistics of frame count and duration of our NWPU Campus dataset are shown in Tab. 2. The entire dataset lasts 16.29 hours, involving 4.27 hours of the testing set. Fig. 2(a) shows the duration distribution of the 305 training videos and 242 testing videos. The average duration of the training videos is 2.37 minutes, and that of the testing videos is 1.05 minutes. There are 124 videos in the testing set that contain anomalous events, and Fig. 2(b) presents the percentage of abnormal frames in the abnormal videos.

In order to highlight the traits of our dataset, we com-

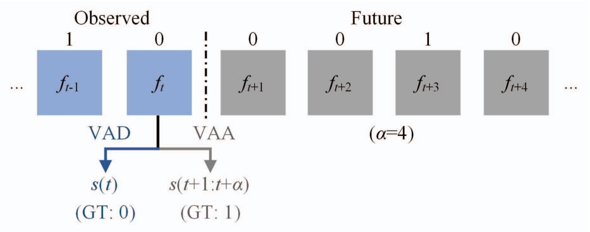

Figure 3. Illustration of video anomaly detection (VAD) and anticipation (VAA). f_t is the frame at time t. ”0” represents normality and ”1” represents abnormality. $s()$ denotes the anomaly score. α is the anticipation time. ”GT” stands for groundtruth.

prehensively compare NWPU Campus with other widely-used datasets for semi-supervised video anomaly detection, as shown in Tab. 1. It can be concluded that the proposed NWPU Campus dataset has three outstanding traits. First, it is the largest semi-supervised video anomaly detection dataset, which is over three times larger than the existing largest dataset (*i.e.*, IITB Corridor). Second, the scenes and anomaly classes of our dataset are diverse and complex. It is a real recorded dataset with the largest number of abnormal event classes and scenes by far. Although the UBnormal dataset also has multiple scenarios, it is a virtual dataset generated by animation rather than real recordings. Third, our dataset takes into account the scene-dependent anomalous events, which is an important type of anomaly not included in other multi-scene datasets. Besides the above three advantages, the NWPU Campus is also the first dataset proposed for video anomaly anticipation, which will be introduced in detail in the next section.

4. Proposed Method

4.1. Problem Formulation

Video anomaly detection (VAD) aims to detect whether an anomaly is occurring at the current moment. As to anomaly anticipation, considering that it is difficult and inessential to anticipate the exact time of the occurrence of an abnormal event, we define video anomaly anticipation (VAA) to anticipate whether an anomaly will occur in a future period of time, which is meaningful and useful for early warnings of anomalous events. We illustrate the VAD and VAA tasks in Fig. 3.

Suppose the current time step is t. For VAD, an algorithm can compute an anomaly score $s(t)$ for the current frame f_t based on the observed frames $f_{t-n}, \cdots, f_t$, where n represents the observed duration. In Fig. 3, f_t is a normal frame, and therefore the anomaly score $s(t)$ is expected to be as low as possible. For VAA, at the current moment t, we anticipate whether an anomaly will occur at any future frame in the period of $[t+1, t+\alpha]$ that has not been observed, where $\alpha \geq 1$ is the anticipation time. We use the score $s(t+1:t+\alpha)$ to represent the anticipated probability of an anomaly occurring during $t+1$ to $t+\alpha$ frames. In Fig. 3 where $\alpha = 4$ is taken as an example, since f_{t+3} is abnormal, the groundtruth of $s(t+1:t+\alpha)$ is 1, denoting there will be an anomaly in frames $f_{t+1}, \cdots, f_{t+4}$. We expect that the anomaly score $s(t+1:t+\alpha)$ to be as high as possible, which is contrary to $s(t)$.

As can be seen, the groundtruth is different for VAD and VAA. For VAD, we denote the frame-level labels of a video as $G_0 = \{g_t\}_{t=1}^{T}$, where $g_t \in \{0,1\}$ indicates the frame f_t is normal (0) or abnormal (1), and T is the length of the video. Based on G_0, the frame-level labels for VAA where the anticipation time is α can be calculated by:

$$G_\alpha = \{\max(\{g_{t+i}\}_{i=1}^{\alpha})\}_{t=1}^{T-\alpha}, \qquad (1)$$

where $\max()$ denotes the maximum value in a set.

Note that the action anticipation models (*e.g.* [41–43]) are not applicable to semi-supervised VAA, since there are no anomaly data and labels to train them in a supervised manner. Therefore, we propose a novel method for semi-supervised VAD and VAA in the next section.

4.2. Forward-backward Scene-conditioned Auto-encoder

Our model is based on the prevalent frame prediction model. However, future groundtruth frames are not visible in VAA, and hence the prediction error cannot be calculated. To address this issue, we propose to estimate the prediction error of future frames by forward-backward prediction, and the proposed model is shown in Fig. 4.

Moreover, we propose to employ a scene-conditioned auto-encoder to handle the scene-dependent anomalies. Specifically, we take the encoding of scene image as the condition of conditional variational auto-encoder (CVAE), train it to generate image features related to the scene, and finally decode the features into the predicted frames.

4.2.1 Forward-backward Frame Prediction

As shown in Fig. 4, our model includes a forward and a backward frame prediction networks. The forward network predicts multiple future frames in one shot based on the observed frames, and the backward network reversely predicts an observed frame based on the future frames generated by the forward network and a part of the observed frames. Our motivation is that, if the future frame is anomalous in forward prediction, the predicted image will be inaccurate. When we use the inaccurate image as a part of the input for the backward frame prediction model, the output frame will also have a large error with the groundtruth frame, which is available since it has been observed. Therefore, we can anticipate the future anomalies through the error of forward-backward frame prediction.

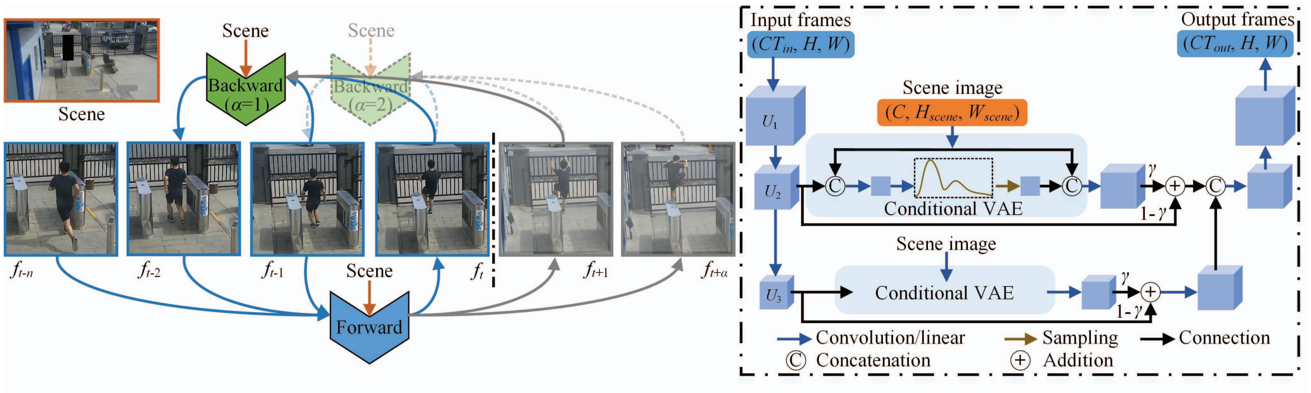

Figure 4. The proposed forward-backward scene-conditioned auto-encoder. It consists of a forward and a backward frame prediction networks. Each network has the same U-Net architecture with conditional VAEs that take the scene image as the input. t, n and α respectively represent the current time, the observation time and the anticipation time. C, T, H and W respectively represent the channel, temporal length, height and width of the input frames. U_i denotes the i-th level of the U-Net. γ is a weight in scalar. Best viewed in color.

At the current time step t, the forward network takes the observed frames $f_{t-n}, \cdots, f_{t-1}$ as the input, and outputs the predicted frames $\hat{f}_t, \cdots, \hat{f}_{t+\alpha}$. We compute the mean square error (MSE) loss and L1 loss between every predicted frame $\hat{f}_{t+i}$ ($i \in [0, \alpha]$) and its groundtruth frame to train the forward network:

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

where λ_{L1} is the weight of L1 loss.

For training the backward network that anticipates the anomaly score of the i-th ($i \in [1, \alpha]$) future frame, we feed the predicted future frames $\hat{f}_{t+i}, \cdots, \hat{f}_{t+1}$ and the real future frames $f_{t+i}, \cdots, f_{t+1}$ respectively along with the observed frames $f_t, \cdots, f_{t+i+1-n}$ into it. In this way, our backward network can make use of the observed information to make more accurate short-term anomaly anticipation. The output predicted frames of the two forms of inputs are denoted as $\hat{f}^{(1)}_{t+i-n}$ and $\hat{f}^{(2)}_{t+i-n}$, respectively, which share the same groundtruth frame f_{t+i-n}. We calculate the average MSE and L1 losses between $\hat{f}^{(1)}_{t+i-n}$ and f_{t+i-n}, as well as $\hat{f}^{(2)}_{t+i-n}$ and f_{t+i-n} to train the backward network:

$$L_b = \frac{1}{2}(L_f(\hat{f}^{(1)}_{t+i-n}, f_{t+i-n}) + L_f(\hat{f}^{(2)}_{t+i-n}, f_{t+i-n})). \tag{3}$$

During inference, only the predicted forward future frames $\hat{f}_{t+i}, \cdots, \hat{f}_{t+1}$ and the observed frames $f_t, \cdots, f_{t+i+1-n}$ are required for backward prediction. For different time steps $t+1, \cdots, t+\alpha$, the backward networks share the same weights.

4.2.2 Scene-conditioned VAE

Both the forward and backward networks are three-level U-Nets [44] of the same architecture, containing CVAEs that guide the encoding of input frames to be associated with scenes. The input frames are merged in time and channel dimensions and fed into the encoder of a 2D convolutional network, which outputs three feature maps of different shapes. The feature maps at U_2 and U_3 levels are fed into the CVAEs to generate new feature maps conditioned on the scene image. Then the scene-conditioned feature maps are added to the input of CVAEs with a weight $\gamma \in [0, 1]$. Finally, the predicted frames are generated through subsequent decoding convolutional layers.

A CAVE takes as input the feature maps of the frames and the encoding of the scene image. Note that the frames only focus on the local regions of detected objects, while the objects in the scene image are masked out and only the background is retained. The scene image is encoded by convolutional layers, concatenated with the frame feature maps and fed into the encoder of CVAE to generate the parameters of a posterior distribution. We use the reparameterization technique [45] to sample latent variables from the posterior distribution, and feed them into the CVAE decoder after concatenated with the scene encoding to generate scene-conditioned feature maps. We assume that the prior distribution is a standard Gaussian distribution and calculate the Kullback-Leibler (KL) divergence between it and the posterior distribution as the loss:

$$L_{KL}(\mathcal{N}(\hat{\mu}, \hat{\sigma}^2) \| \mathcal{N}(0, 1)) = -\frac{1}{2}(\log \hat{\sigma}^2 - \hat{\mu}^2 - \hat{\sigma}^2 + 1), \tag{4}$$

where $\hat{\mu}$ and $\hat{\sigma}^2$ are the mean and variance of the posterior Gaussian distribution. In testing stage, if the input feature maps do not match the scene, they will be reconstructed by the CVAE with large errors, thereby identifying scene-dependent anomalies.

Finally, the total loss is the sum of the losses of forward prediction, backward prediction, and KL divergence with the weight of λ_{KL}. We minimize the total loss to jointly train the whole model.

4.2.3 Anomaly Score

During inference, we calculate the error between the predicted forward frame $\hat{f}_t$ and its groundtruth frame f_t by Eq. (2) as the anomaly score for VAD:

$$s(t) = L_f(f_t, \hat{f}_t). \tag{5}$$

For VAA with the anticipation time of α, we first estimate the anomaly score of f_{t+i} ($i \in [1, \alpha]$) through forward-backward prediction. Then, the maximum error in the period of $[t+1, t+\alpha]$ is taken as the anticipation anomaly score:

$$s(t+1 : t+\alpha) = \max(\{L_f(f_{t+i-n}, \hat{f}_{t+i-n})\}_{i=1}^{\alpha}). \tag{6}$$

Consequently, we can detect and anticipate anomalies simultaneously.

5. Experiments

5.1. Experimental Setup

Datasets. We experiment on the ShanghaiTech [23], CUHK Avenue [27], IITB Corridor [29] and our proposed NWPU Campus datasets, which are described in Tab. 1 and the Related Work section. Our dataset is available at: (it will be released after the double-blind review). For convenience, we abbreviate the above datasets to ”ST”, ”Ave”, ”Cor”, and ”Cam” respectively in the following tables.

Evaluation Metric. We use the area under the curve (AUC) of receiver operating characteristic (ROC) to evaluate the performance for both VAD and VAA. Note that we concatenate all the frames in a dataset and then compute the overall frame-level AUC, which is widely adopted.

Implementation Details. The input frames of our model are the regions of 256×256 pixels centered on objects that detected by the pre-trained ByteTrack [50] implemented by MMTracking [51]. For the forward and backward networks, they both take T_{in}=8 frames as the input, while they output T_{out}=7 and T_{out}=1 frames, respectively. The 1st frame output by the forward network is used for anomaly detection, and the 2nd to 7th frames are fed into the backward network for anomaly anticipations of different anticipation times. We design the encoder of U-Net based on ResNet [52] and the decoder are multiple convolutional layers. The network for scene encoding is a classification model to classify scenes, which is firstly trained with known scene information, and then frozen during training the entire model. The weights γ, λ_{L1} and λ_{KL} are 1, 1 and 0.1 by default. We adopt the maximum local error [53] to focus on the errors in local regions. Please refer to the supplementary material for a detailed description of our model.

Table 3. Comparison of different methods on the ShanghaiTech, CUHK Avenue, IITB Corridor and NWPU Campus datasets in AUC (%) metric. The best result on each dataset is shown in bold.

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**

5.2. VAD Performance Benchmarking

The comparison between our method and other existing methods on the ShanghaiTech, CUHK Avenue, IITB Corridor and NWPU Campus datasets is shown in Tab. 3. We reproduce a total of 7 recent reconstruction-based [39], distance-based [35] and prediction-based [4, 6, 10, 11, 40] methods on our NWPU Campus dataset using their official codes. For a fair comparison, the self-supervised learning based methods [54–56] are excluded, and we use the same detected objects as the inputs for the reproduced methods. The γ in our model is set to 0 for those datasets without scene-dependent anomalies. As can be seen in Tab. 3, our method outperforms the others on the NWPU Campus, IITB Corridor and ShanghaiTech datasets, all of which contain over 10 classes of abnormal events. The superior performance demonstrates the advantage of our method for complex and large-scale VAD. We find that the relatively low performance on the CUHK Avenue is mainly due to the inaccurate object tracking of the tracking algorithm, which is caused by the low resolution of this dataset. The performance for VAD on the NWPU Campus is lower than that on other datasets because our dataset contains various types of

Table 4. AUCs (%) of different methods on scene-dependent anomalous datasets. The ShanghaiTech-sd dataset used in this table is reorganized by us. The best results are shown in bold.

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**

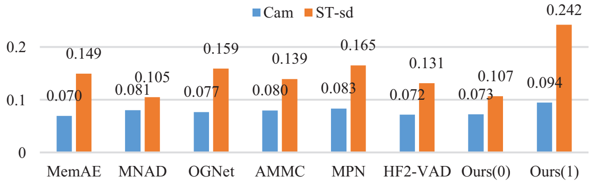

Figure 5. Score gaps of different methods. "Ours (0)" and "Ours (1)" denote our methods with $\gamma = 0$ and $\gamma = 1$, respectively. A higher value means better.

anomalies, and each anomaly has multiple manifestations, making it much more challenging than other datasets.

5.3. Study on Scene-dependent Anomalies

In addition to our NWPU Campus dataset, we also reorganize a new dataset named ShanghaiTech-sd using a part of the videos from the ShanghaiTech dataset to specifically study scene-dependent anomaly detection. ShanghaiTech-sd contains 4 scenes where "cycling" is set as a scene-dependent anomaly. The performances of different methods are shown in Tab. 4. It can be seen that the proposed scene-conditioned VAE (*i.e.* γ=1) makes a significant improvement, with increases of 2.4% and 12.3% on the NWPU Campus and ShanghaiTech-sd, respectively, surpassing other methods by a margin. We analyze the score gaps between normal and abnormal scores of those methods, as can be seen in Fig. 5. In particular, the score gap of our method with γ=1 is obviously higher than that with γ=0 and other methods, suggesting that the proposed scene-conditioned VAE can distinguish scene-dependent anomalies. We provide the details of the ShanghaiTech-sd dataset and more analysis in the supplementary material.

5.4. Video Anomaly Anticipation

We conduct experiments on the NWPU Campus dataset for VAA with different anticipation times, as shown in Tab. 5. We report the results of stochastic anticipations ("Chance") and human beings ("Human"). Four volunteers not involved in the construction of the dataset participate in the evaluation of anomaly anticipation. Since humans cannot perceive time precisely, the volunteers only anticipate whether an anomalous event will occur in 3 seconds or not. The result of "Human" is the average performance of all the volunteers. For the forward-only model (*i.e.*, f-only), we calculate the maximum error between the predicted future frames in α_t seconds and the current frame, which is then taken as the anticipated anomaly score. The forward-backward model (*i.e.*, f+b) computes anomaly scores as mentioned in Sec. 4.2.3. It can be seen that our forward-backward prediction method is more effective than the forward-only method. However, there is still much room for improvement compared with the performance of humans, which demonstrates that the proposed dataset and VAA task are extremely challenging for algorithms.

Table 5. AUCs (%) for video anomaly anticipation with different anticipation times (*i.e.* α_t seconds) on the NWPU Campus dataset. "f" and "b" denote forward and backward predictions.

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

6. Conclusion

In this work, we propose a new comprehensive dataset NWPU Campus, which is the largest one in semi-supervised VAD, the only one considering scene-dependent anomalies, and the first one proposed for video anomaly anticipation (VAA). We define VAA to anticipate whether an anomaly will occur in a future period of time, which is of great significance for early warning of anomalous events. Moreover, we propose a forward-backward scene-conditioned model for VAD and VAA as well as handling scene-dependent anomalies. In the future, our research will focus not only on the short-term VAA, but also on long-term anticipation.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Project No. U19B2037, 61906155, 62206221), the Key R&D Project in Shaanxi Province (Project No. 2023-YBGY-240), the Young Talent Fund of Association for Science and Technology in Shaanxi, China (Project No. 20220117), and the National Key R&D Program of China (No. 2020AAA0106900).

References

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15:1–15:58, 2009. 1

[2] Bharathkumar Ramachandra, Michael J. Jones, and Ranga Raju Vatsavai. A Survey of Single-Scene Video Anomaly Detection. *IEEE TPAMI*, 44(5):2293–2312, 2022. 1

[3] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future Frame Prediction for Anomaly Detection - A New Baseline. In *CVPR*, pages 6536–6545, 2018. 1, 4, 7

[4] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton Van Den Hengel. Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection. In *ICCV*, pages 1705–1714, 2019. 1, 3, 4, 7, 8

[5] Trong Nguyen Nguyen and Jean Meunier. Anomaly Detection in Video Sequence With Appearance-Motion Correspondence. In *ICCV*, pages 1273–1283, 2019. 1, 3

[6] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning Memory-Guided Normality for Anomaly Detection. In *CVPR*, pages 14360–14369, 2020. 1, 4, 7, 8

[7] Hao Song, Che Sun, Xinxiao Wu, Mei Chen, and Yunde Jia. Learning Normal Patterns via Adversarial Attention-Based Autoencoder for Abnormal Event Detection in Videos. *IEEE TMM*, 22(8):2138–2148, 2020. 1, 4, 7

[8] Yu Zhang, Xiushan Nie, Rundong He, Meng Chen, and Yilong Yin. Normality Learning in Multispace for Video Anomaly Detection. *IEEE TCSVT*, pages 1–1, 2020. 1, 4

[9] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. BMAN: Bidirectional Multi-Scale Aggregation Networks for Abnormal Event Detection. *IEEE TIP*, 29:2395–2408, 2020. 1, 4, 7

[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. 1, 4, 7, 8

[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. 1, 4, 7, 8

[12] Joey Tianyi Zhou, Le Zhang, Zhiwen Fang, Jiawei Du, Xi Peng, and Yang Xiao. Attention-Driven Loss for Anomaly Detection in Video Surveillance. *IEEE TCSVT*, 30(12):4639–4647, 2020. 1, 4

[13] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze Test Helps: Effective Video Anomaly Detection via Learning to Complete Video Events. In *ACM MM*, pages 583–591, 2020. 1, 4, 7

[14] Jongmin Yu, Younkwan Lee, Kin Choong Yow, Moongu Jeon, and Witold Pedrycz. Abnormal Event Detection and Localization via Adversarial Event Prediction. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2021. 1, 4

[15] Weixin Luo, Wen Liu, Dongze Lian, and Shenghua Gao. Future Frame Prediction Network for Video Anomaly Detection. *IEEE TPAMI*, pages 1–1, 2021. 1, 4

[16] Dongyue Chen, Lingyi Yue, Xingya Chang, Ming Xu, and Tong Jia. NM-GAN: Noise-modulated generative adversarial network for video anomaly detection. *PR*, 116:107969, 2021. 1, 4

[17] Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu Wang, and Qi Qi. Robust Unsupervised Video Anomaly Detection by Multipath Frame Prediction. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2021. 1, 4, 7

[18] Zhiwen Fang, Joey Tianyi Zhou, Yang Xiao, Yanan Li, and Feng Yang. Multi-Encoder Towards Effective Anomaly Detection in Videos. *IEEE TMM*, 23:4106–4116, 2021. 1, 4, 7

[19] Zhiwen Fang, Jiafei Liang, Joey Tianyi Zhou, Yang Xiao, and Feng Yang. Anomaly Detection With Bidirectional Consistency in Videos. *IEEE Transactions on Neural Networks and Learning Systems*, 33(3):1079–1092, 2022. 1, 4

[20] Sijia Zhang, Maoguo Gong, Yu Xie, A. K. Qin, Hao Li, Yuan Gao, and Yew-Soon Ong. Influence-aware Attention Networks for Anomaly Detection in Surveillance Videos. *IEEE TCSVT*, pages 1–1, 2022. 1, 4

[21] Jing Li, Qingwang Huang, Ying-Jun Du, Xiantong Zhen, Shengyong Chen, and Ling Shao. Variational Abnormal Behavior Detection With Motion Consistency. *IEEE TIP*, 31:275–286, 2022. 1, 4, 7

[22] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, pages 1975–1981, 2010. 1, 2, 3

[23] Weixin Luo, Wen Liu, and Shenghua Gao. A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework. In *ICCV*, pages 341–349, 2017. 1, 2, 3, 7

[24] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. UBnormal: New Benchmark for Supervised Open-Set Video Anomaly Detection. In *CVPR*, pages 20111–20121, 2022. 1, 2, 3

[25] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and David Reinitz. Robust Real-Time Unusual Event Detection using Multiple Fixed-Location Monitors. *IEEE TPAMI*, 30(3):555–560, 2008. 2, 3

[26] University of Minnesota. Unusual crowd activity dataset of university of minnesota. `http://mha.cs.umn.edu/proj_events.shtml#crowd`. 2, 3

[27] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal Event Detection at 150 FPS in MATLAB. In *ICCV*, pages 2720–2727, 2013. 2, 3, 7

[28] Bharathkumar Ramachandra and Michael Jones. Street Scene: A new dataset and evaluation protocol for video anomaly detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020. 2, 3

[29] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-timescale Trajectory Prediction for Abnormal Human Activity Detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2615–2623, 2020. 2, 3, 7

[30] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-World Anomaly Detection in Surveillance Videos. In *CVPR*, pages 6479–6488, 2018. 2

[31] 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 *ECCV*, volume 12375, pages 322–339, 2020. 3

[32] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In *ICIP*, pages 1577–1581, 2017. 3

[33] Hung Vu, Tu Dinh Nguyen, Trung Le, Wei Luo, and Dinh Q. Phung. Robust Anomaly Detection in Videos Using Multilevel Representations. In *AAAI*, pages 5216–5223, 2019. 3

[34] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *ICCV*, pages 2242–2251, 2017. 3

[35] Yue Lu, Congqi Cao, Yifan Zhang, and Yanning Zhang. Learnable Locality-Sensitive Hashing for Video Anomaly Detection. *IEEE TCSVT*, pages 1–1, 2022. 3, 7

[36] Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan. Clustering Driven Deep Autoencoder for Video Anomaly Detection. In *ECCV*, volume 12360, pages 329–345. 2020. 3, 7

[37] Peng Wu, Jing Liu, and Fang Shen. A Deep One-Class Neural Network for Anomalous Event Detection in Complex Scenes. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2609–2622, 2020. 3

[38] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *CVPR*, 2016. 3

[39] Muhammad Zaigham Zaheer, Jin-Ha Lee, Marcella Astrid, and Seung-Ik Lee. Old Is Gold: Redefining the Adversarially Learned One-Class Classifier Training Paradigm. In *CVPR*, pages 14171–14181, 2020. 3, 7, 8

[40] 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. 4, 7, 8

[41] Rohit Girdhar and Kristen Grauman. Anticipative Video Transformer. In *ICCV*, pages 13485–13495, 2021. 5

[42] Dayoung Gong, Joonseok Lee, Manjin Kim, Seong Jong Ha, and Minsu Cho. Future Transformer for Long-term Action Anticipation. In *CVPR*, pages 3042–3051, 2022. 5

[43] Tianshan Liu and Kin-Man Lam. A Hybrid Egocentric Activity Anticipation Framework via Memory-Augmented Recurrent and One-shot Representation Forecasting. In *CVPR*, pages 13894–13903, 2022. 5

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015. 6

[45] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014. 6

[46] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos. In *CVPR*, pages 11988–11996, 2019. 7

[47] Xinyang Feng, Dongjin Song, Yuncong Chen, Zhengzhang Chen, Jingchao Ni, and Haifeng Chen. Convolutional Transformer based Dual Discriminator Generative Adversarial Networks for Video Anomaly Detection. In *ACM MM*, pages 5546–5554, 2021. 7

[48] Chao Huang, Jie Wen, Yong Xu, Qiuping Jiang, Jian Yang, Yaowei Wang, and David Zhang. Self-Supervised Attentive Generative Adversarial Networks for Video Anomaly Detection. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2022. 7

[49] Zhiwei Yang, Peng Wu, Jing Liu, and Xiaotao Liu. Dynamic Local Aggregation Network with Adaptive Clusterer for Anomaly Detection. In *ECCV*, pages 404–421, 2022. 7

[50] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object Tracking by Associating Every Detection Box. In *ECCV*, volume 13682, pages 1–21, 2022. 7

[51] MMTracking Contributors. MMTracking: OpenMMLab video perception toolbox and benchmark. https://github.com/open-mmlab/mmtracking, 2020. 7

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[53] Congqi Cao, Yue Lu, and Yanning Zhang. Context Recovery and Knowledge Retrieval: A Novel Two-Stream Framework for Video Anomaly Detection. *CoRR*, abs/2209.02899, 2022. 7

[54] Ziming Wang, Yuexian Zou, and Zeming Zhang. Cluster Attention Contrast for Video Anomaly Detection. In *ACM MM*, pages 2463–2471, 2020. 7

[55] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly Detection in Video via Self-Supervised and Multi-Task Learning. In *CVPR*, pages 12742–12752, 2021. 7

[56] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video Anomaly Detection by Solving Decoupled Spatio-Temporal Jigsaw Puzzles. In *ECCV*, volume 13670, pages 494–511, 2022. 7