Comprehensive Regularization in a Bi-directional Predictive Network for Video Anomaly Detection

Chengwei Chen[1], Yuan Xie*[1], Shaohui Lin[1], Angela Yao[2], Guannan Jiang[3], Wei Zhang[3], Yanyun Qu[4], Ruizhi Qiao[5], Bo Ren[5], Lizhuang Ma*[1]

[1] East China Normal University, Shanghai, China
[2] National University of Singapore, Singapore
[3] Contemporary Amperex Technology Co., Limited (CATL), Fujian, China
[4] Xiamen University, Fujian, China
[5] Tencent Youtu Lab, Shenzhen, China

52184501028@stu.ecnu.edu.cn, {yxie,shlin}@cs.ecnu.edu.cn, ayao@comp.nus.edu.sg, {jianggn,zhangwei}@catl.com,
yyqu@xmu.edu.cn, {ruizhiqiao,timren}@tencent.com, lzma@cs.ecnu.edu.cn

Abstract

Video anomaly detection aims to automatically identify unusual objects or behaviours by learning from normal videos. Previous methods tend to use simplistic reconstruction or prediction constraints, which leads to the insufficiency of learned representations for normal data. As such, we propose a novel bi-directional architecture with three consistency constraints to comprehensively regularize the prediction task from pixel-wise, cross-modal, and temporal-sequence levels. First, *predictive consistency* is proposed to consider the symmetry property of motion and appearance in forwards and backwards time, which ensures the highly realistic appearance and motion predictions at the pixel-wise level. Second, *association consistency* considers the relevance between different modalities and uses one modality to regularize the prediction of another one. Finally, *temporal consistency* utilizes the relationship of the video sequence and ensures that the predictive network generates temporally consistent frames. During inference, the pattern of abnormal frames is unpredictable and will therefore cause higher prediction errors. Experiments show that our method outperforms advanced anomaly detectors and achieves state-of-the-art results on UCSD Ped2, CUHK Avenue, and ShanghaiTech datasets.

Introduction

Video anomaly detection (VAD) is critical for video surveillance systems. A key challenge in developing machine learning methods for VAD is that very few or even no samples of abnormal data are available for learning. This makes it a one-class classification problem (Perera and Patel 2019), in which one must learn a distribution based only on normal distances. VAD methods learn the normal distribution implicitly within a model; anomalies are then detected by the model's inability to either reconstruct (Zhou et al. 2019; Gong et al. 2019; Nguyen and Meunier 2019) or predict (Liu et al. 2018; Lu et al. 2019; Zhou et al. 2019) some data samples. During inference, normal samples are assumed to have low reconstruction or prediction errors, while samples with high reconstruction or prediction errors are anomalies.

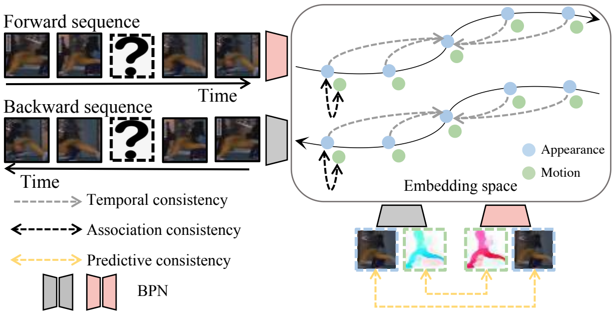


Figure 1: Our bi-directional VAD framework predicts missing patches forwards and backwards in time for video anomaly detection. The multi-granularity consistency constraints regularize the prediction of patches from pixel-wise, cross-modality, and temporal-sequence levels. High prediction errors suggest the presence of an abnormal event. For better presentation, we set the number of input sequence frames to 5 in all figures.

Naturally, the *"normal"* model must be sufficiently expressive for such an assumption to hold. Learning such a model for a medium as rich and high-dimensional as video can be especially challenging. The ability to capture the intrinsic properties of the video, such as the appearance, dynamic information, and temporal sequencing all play an important role. In our method, we employ a patch-wise prediction approach to do VAD through the bi-directional architecture. Some previous VAD works predict entire frames (Liu et al. 2018; Lu et al. 2019); we follow (Yu et al. 2020) and predict only patches associated with video events based on detected objects, which avoids the interference of irrelevant background. The abnormal event is detected by the prediction error of appearance and optical flow in forward and backward directions. As shown in Fig. 1, this architecture enables comprehensive consistency constraints to regularize the prediction task from pixel-wise, cross-modal, and temporal-sequence levels.

In contrast to the simple reconstruction or prediction con-

*corresponding author

straints of previous models, we propose multi-granular consistency constraints based on the video's inherent characteristics. Firstly, in our predictive consistency, the backwards treatment presents the possibility to enforce forward-backward constraints in motion and appearance prediction. This consistency is based on the symmetry property of motion and appearance forwards and backwards in time. The appearance in the forwards direction should be consistent with the backwards predicted appearance at the corresponding pixel. Similarly, the optical flow in the forwards prediction should be the inverse of the backwards prediction. However, the previous methods (Liu et al. 2018; Zhou et al. 2019) impose appearance and motion constraints for prediction quality without the backwards. Their constraints only minimize the difference between a generated image and its ground truth from pixel-level in forwards direction.

Secondly, apart from employing predictive constraint in each modality prediction, we design the association constraint in terms of relevance between different modalities. The multi-modal discriminator is added to distinguish between matching versus non-matching appearance and motion predictions for association consistency. In our association consistency, we consider the consistent correlation between appearance and motion in VAD, which was ignored by previous methods (Tang et al. 2020; Yan et al. 2018). Recently, AMMC-Net(Cai et al. 2021) models the consistency between regular appearance and motion through complex memory modules. It learns two mapping functions from the appearance memory pool feature to the motion memory pool feature and vice versa. Differently, avoiding the design of an extra sophisticated network, our simple multi-modal discriminator estimates the association between the ground truth appearance and its corresponding input motion.

Finally, we add a temporal consistency constraint, ensuring the predictive network to predict more temporally consistent frames. The sequence discriminator is added to distinguish between real versus fake sequences for temporal consistency. Although the previous approaches (Cai et al. 2021; Yu et al. 2020) consider the temporal feature to regularize the prediction task through the motion information, the motion (optical flow) can only represent the short-term temporal relationship between two adjacent frames. The long-term temporal relationship of events that occurs in a video sequence was also not concerned in these approaches. Our model is able to obtain a rich yet discriminative representation of normal video events that is easily separable from abnormal events even though we do not have samples of the latter during training. Experimental evaluation shows that our method surpasses state-of-the-art on several VAD benchmarks.

We summarize our contributions below:

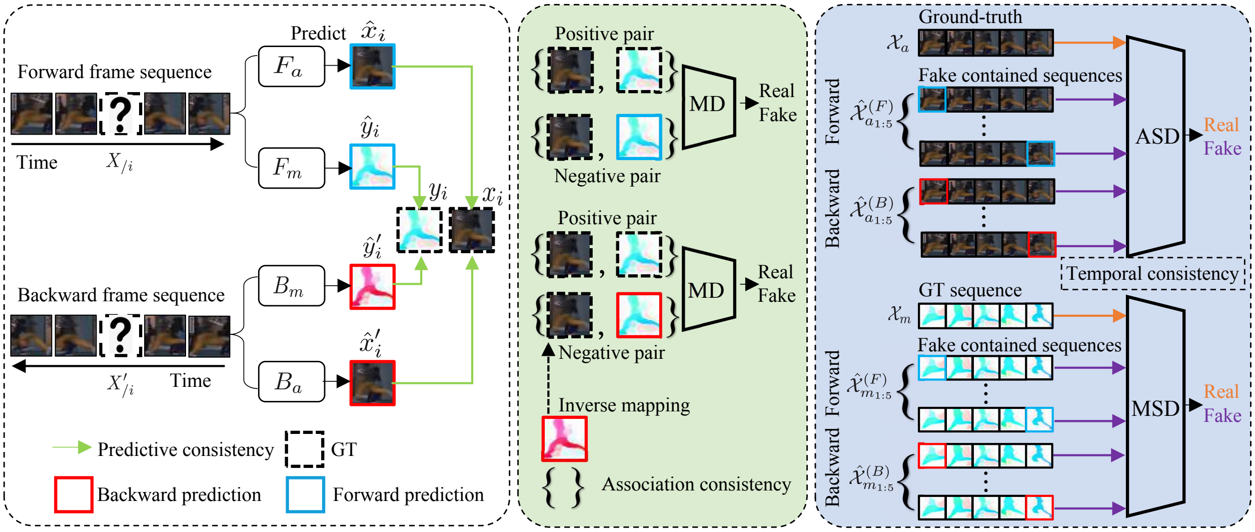
- We introduce three consistency regularizations from a pixel-wise, a cross-modality and temporal-sequence levels; these consistencies are unaccounted for in previous works.
- By assuming forwards-backwards symmetry in appearance and flow, the predictive consistency regularizes the modality prediction through a novel bi-directional predictive framework.
- The association consistency explicitly models the correction between modalities by the multi-modal discriminator. The temporal consistency captures the temporal relationship of a video sequence by the sequence-wise discriminator.
- Extensive experiments demonstrate that our method can surpass state-of-the-art methods on several VAD benchmarks. On ShanghaiTech, our method achieves the frame-level AUC of 78.1%.

Related Work

Reconstruction VAD Methods attempt to capture the distribution of normal video and reconstruct these videos with high quality in the training process. During inference, the distribution of anomaly samples should be far from the learned distribution and lead to a large reconstruction error. Some propose a Convolutional Autoencoder to reconstruct an input sequence of frames (Hasan et al. 2016; Tran and Hogg 2017). Recent works explore variants of Convolutional Autoencoders such as two-stream recurrent framework (Yan et al. 2018), a parametric density estimator (Abati et al. 2019) and a memory-augmented autoencoder (Gong et al. 2019). The reconstruction-based approaches attempt to reconstruct whole frames from scratch, but they sometimes suffer from over-fitting (Kieu et al. 2019) and can even reconstruct abnormal event well (Liu et al. 2018), which cannot distinguish between normal and abnormal data easily and successfully.

Prediction VAD Methods aim to predict future frames based on the context of previous frames. They (Liu et al. 2018; Lu et al. 2019; Fan, Zhu, and Yang 2019) assume that normal events are predictable while abnormal ones are unpredictable. The previous method already proposes some consistencies to regularize the prediction task. For example, (Liu et al. 2018) propose a method that predicts the future frame with higher quality for normal events by the simple intensity and gradient constraint. They regularize the predicted result by comparing the value of each pixel between the ground-truth image and the predicted image. Apart from single modality constraint, (Cai et al. 2021) attempts to model the consistency between appearance and motion information through the complex modality memory pools. It combines the multiple modality features to build a more robust feature representation of normal events. Recently, inspired by the cloze test ("fill-in-the-blank") (Taylor 1953) used in language understanding, Yu *et al.*(Yu et al. 2020) proposed a novel prediction task by predicting erased patches of incomplete video events and fully exploit temporal information in the video. However, it still simply take previous pixel-wise constraint to regularize the prediction task and ignore the correlation between optical flow and video frame.

Unlike these previous prediction approaches in VAD, our focus is on exploring and leveraging the full extent of the information contained both forwards and backwards in time within a video. Besides, we consider modeling the relationship between appearance and motion through a simple



(a) Bidirectional Predictive Networks (b) Multi-modal Discriminator (c) Sequence-wise discriminator

Figure 2: Method overview. The bi-directional predictive networks take the incomplete sequences as input to predict the corresponding motion and appearance of the erased frame, which is regularized by predictive consistency. To keep the relevance between different modalities, the multi-modal discriminator (MD) learns to classify between fake {target appearance, generated motion} and real {target appearance, target motion} tuples. Appearance and motion sequence-wise discriminators (ASD and MSD) guarantee the predictors to generate the temporally consistent appearance or motion patches by distinguishing a fake contained sequence from the ground-truth sequence.

multi-modal discriminator. It estimates the association between ground truth appearance and its corresponding input motion. Finally, the temporal relationships in the video sequence are also taken into account. It is regularized by the sequence-wise discriminator which distinguishes a fake contained sequence from the ground-truth sequence.

Methodology

The framework (see overview in Fig. 2) consists of bi-directional predictive networks (BPNs), a multi-modal discriminator (MD), and a sequence-wise discriminator (SD). Three comprehensive consistency regularizations include predictive consistency, association consistency, and temporal consistency.

Given the incomplete video sequence in forwards and backwards order, the BPNs predict the missing frame's appearance and motion in both directions. To make highly realistic appearance and motion predictions, three consistencies regularize the prediction task from pixel-wise, cross-modal, and temporal-sequence levels. Firstly, in BPNs, the predictive consistency penalizes pixel differences between the predicted frame and target frame. Secondly, association consistency ensures that each predicted motion is strongly related to the target appearance. The multi-modal discriminator (MD) determines an association between each input appearance and its corresponding motion. Finally, to predict the temporally consistent frames, the appearance and motion sequence discriminators (ASD, MSD) are used to decide whether the sequence contains a predicted sample or not. During inference, the anomaly is detected through the video event completion task in BPNs. When a video event is abnormal, the missing frame's appearance and motion prediction error should be higher than a normal video event.

Video Event Extraction

We wish to avoid the influence of the background and focus explicitly on the objects presented in the scene. To do so, we apply a pre-trained cascade R-CNN (Cai and Vasconcelos 2018) as an object detector to each frame. Since all abnormal events in the available public datasets (Lu, Shi, and Jia 2013; Luo, Liu, and Gao 2017; Tan et al. 2021) are defined by the anomaly object or behavior, we perform the basic object detection in preprocessing to avoid the inference of background and focusing on the pattern of objects. For a frame at time t, each detection bounding box is applied as an ROI to extract a track from frames $t-(n-1)$ to t. We define the resized track (to a fixed resolution of 32×32) as a video event; each event can produce n (*e.g.* $n = 5$) different incomplete sequences by removing any one of the frames (see in the supplementary). The prediction task is then to infer this missing frame. Aside from appearance defined by the RGB frames, we also estimate the corresponding n-frame optical flow track.

Bi-directional Predictive Networks

The Bi-directional Predictive Networks have a forward and a backward branch (see Fig. 2(a)). Each branch contains two U-Net architectures (Ronneberger, Fischer, and Brox 2015) to predict the missing frame of RGB and optical flow respectively from some incomplete sequences of a video event. In the forward branch, F_a and F_m denote the appearance and motion predictors; similarly, B_a and B_m, denote the appearance and motion predictors in the backward direction. For a given video event X with n frames, we denote as an incomplete sequence $X_{/i}$ where the i^{th} frame x_i is missing. The prediction $\hat{x}_i$ is found via the predictor $\hat{x}_i = F_a(X_{/i}, \theta_a)$. Similarly, if the backward sequence is denoted as X', then the corresponding missing frame x'_i can also be predicted as $\hat{x}'_i = B_a(X'_{/i}, \theta'_a)$. For motion predictions, we denote $\hat{y}_i = F_m(X_{/i}, \theta_m)$ and $\hat{y}'_i = B_m(X'_{/i}, \theta'_m)$ as the forward and backward motion predictors, respectively. All four predictors use the same U-Net architecture and differ only in their output size – the appearance predictors have three channels for RGB outputs while the motion predictors have two channels for optical flow outputs. Each predictor has its own unique set of parameters.

Predictive Consistency: We design predictive consistency loss functions to ensure the consistency between the forward and backward predictions. For appearance prediction, we combine the pixel-wise MSE loss and a perceptual Laplacian pyramid loss (Ling and Okada 2006) to approximate the predictors $\hat{x}_i$ and $\hat{x}'_i$ to the corresponding ground-truth x_i and x'_i. We are inspired by (Bojanowski et al. 2018), which promotes the use of the Laplacian pyramid loss to capture edges and context over multiple scales to improve predictions. The predictive consistency loss for appearance can be formulated as:

$$\begin{aligned}\mathcal{L}_a = \sum_{i=1}^{n} \Big(&\|x_i - \hat{x}_i\|_2^2 + \sum_j 2^{2j} \left| \text{Lap}^j(x_i) - \text{Lap}^j(\hat{x}_i) \right|_1 \\ &+ \|x_i - \hat{x}'_i\|_2^2 + \sum_j 2^{2j} \left| \text{Lap}^j(x_i) - \text{Lap}^j(\hat{x}'_i) \right|_1 \Big),\end{aligned} \tag{1}$$

where the first and third terms are MSE losses with respect to forward and backward predictions $\hat{x}_i$ and $\hat{x}'_i$ respectively. The second and fourth terms in Eq. 1 are the Laplacian pyramid loss with respect to forward prediction $\hat{x}_i$ and backward prediction $\hat{x}'_i$, where $\text{Lap}^j(\cdot)$ is the j^{th} level of Laplacian pyramid representation. x_i is the i^{th} original frame, which is denoted as the appearance ground-truth.

Unlike the appearance prediction, the direction of backward motion prediction should be inverse to that of the forward one. We employ the ℓ_1 loss to minimize the distance between predicted motions and target ones:

$$\mathcal{L}_m = \sum_{i=1}^{n} \Big(\|y_i - \hat{y}_i\|_1 + \|y_i + \hat{y}'_i\|_1 \Big), \tag{2}$$

where y_i is the target motion from the i^{th} frame in the original sequence X. Note that the second term in Eq. 2 play a role of pushing away the directions between the target motion y_i and the predicted backward motion $\hat{y}'_i$.

Multi-modal Discriminator

The BPNs' regularization terms focus on the consistency between the forward and backward stream, but cannot associate any relevance or lack thereof between the appearance and motion itself. We therefore propose to construct an association between appearance and motion predictions by adding the multi-modal discriminator. The ground-truth appearance patch x_i and its corresponding ground-truth motion patch y_i are treated as real pairs while the ground-truth appearance patch x_i and generated motion $\hat{y}_i$ or $\hat{y}'_i$ are fake pairs. The discriminator, fed by the concatenation of an erased patch and its motion, learns to classify between fake and real pairs. The structure of multi-modal discriminator is based on the DCGAN (Radford, Metz, and Chintala 2016), where the input layer of size 32×32×5 is fed by the concatenation of a video patch and its motion. Therefore, the objective function of a multi-modal discriminator for forward direction can be formulated as:

$$\mathcal{L}_{\text{mdf}} = -\frac{1}{2} \sum_{i=1}^{n} \Big(\log \mathcal{D}(x_i, y_i) + \log\left[1 - \mathcal{D}(x_i, \hat{y}_i)\right] \Big). \tag{3}$$

Similarly, in the backwards direction, the multi-modal discriminator distinguishes the real pair from fake pair by $\mathcal{L}_{\text{mdb}}$:

$$\mathcal{L}_{\text{mdb}} = -\frac{1}{2} \sum_{i=1}^{n} \Big(\log \mathcal{D}(x_i, y_i) + \log\left[1 - \mathcal{D}(x_i, -\hat{y}'_i)\right] \Big), \tag{4}$$

where $-\hat{y}'_i$ is the inverse of the generated backward motion $\hat{y}'_i$. The multi-modal discriminator loss can be formulated as:

$$\mathcal{L}_{\text{md}} = \mathcal{L}_{\text{mdf}} + \mathcal{L}_{\text{mdb}}. \tag{5}$$

Sequence-wise Discriminator

Finally, we propose sequence-wise adversarial training to ensure temporal consistency. The structure of the sequence-wise discriminator is shown in the supplementary. It decides whether the sequence contains the predicted (fake) image or not. Suppose that N indicates a set of $2n$ fake sequences containing the generated patches both from forward and backward predictors, P is the ground-truth video event, *i.e.*, real sequence. The objective function of the sequence discriminator can be expressed as follows:

$$\mathcal{L}_{\text{sd}} = -\frac{1}{2} \sum_{i=1}^{2n} \Big(\log \mathcal{D}(P) + \log\left[1 - \mathcal{D}(N_i)\right] \Big). \tag{6}$$

Specifically, for each modal (appearance and motion), we design the sequence discriminator respectively. They share the same architecture with independent parameters, except that each sequence of the former has 3 input channels (images) while the latter has 2 (optical flow).

For the appearance stream, as the fake sequence, we construct $\hat{\mathcal{X}}_{a_i}^{(F)}$ and $\hat{\mathcal{X}}_{a_i}^{(B)}$ by replacing the ground-truth sequence at the i^{th} position with the predicted $\hat{x}_i$ and $\hat{x}'_i$ from forward and backward predictors respectively as:

$$\hat{\mathcal{X}}_{a_i}^{(F)} = [x_1 : x_n \backslash \hat{x}_i], \quad \hat{\mathcal{X}}_{a_i}^{(B)} = [x_1 : x_n \backslash \hat{x}'_i]. \tag{7}$$

Therefore, we can further construct a fake appearance sequence set $\hat{\mathcal{X}}_a$ with the sequence number of $2n$ by concatenating all the predicted positions across forward and backward predictors.

The appearance sequence discriminator (ASD) attempts to distinguish between the ground-truth $\mathcal{X}_a$ and fake sequence $2n$ times. The object function Eq. 6 for appearance

Table 1: Ablation study for source information, discriminator and input modality.

Information		Discriminator		Modality		UCSD Ped2	CUHK Avenue	Traffic-Train
Forward	Backward	Sequence	Multi-modal	Motion	Appearance			
✓				✓	✓	97.3%	89.6%	64.2%
✓	✓			✓	✓	97.9%	90.0%	66.6%
✓	✓	✓		✓	✓	98.1%	90.1%	67.6%
✓	✓	✓	✓	✓		96.5%	85.0%	67.4%
✓	✓	✓	✓	✓	✓	98.3%	90.3%	68.5%

sequence discriminator can be expressed by L_{sda}:

$$\mathcal{L}_{\text{sda}} = -\frac{1}{2}\sum_{i=1}^{2n}\Big(\log \mathcal{D}(\mathcal{X}_a) + \log\Big[1-\mathcal{D}(\hat{\mathcal{X}}_{a_i})\Big]\Big). \quad (8)$$

Similarly, for the motion, we also design the object function L_{sdm} for motion sequence discriminator, which classifies an input motion sequence as real or fake. The objective function is formulated as follows:

$$\mathcal{L}_{\text{sdm}} = -\frac{1}{2}\sum_{i=1}^{2n}\Big(\log \mathcal{D}(\mathcal{X}_m) + \log\Big[1-\mathcal{D}(\hat{\mathcal{X}}_{m_i})\Big]\Big), \quad (9)$$

where $\mathcal{X}_m$ is the ground-truth motion sequence. Similar to the construction of fake appearance set $\hat{\mathcal{X}}_a$, $\hat{\mathcal{X}}_m$ is a fake motion sequence set with the sequence number of $2n$, containing the fake motion sequence $\hat{\mathcal{X}}_m^{(F)}$ and $\hat{\mathcal{X}}_m^{(B)}$ from the forward and backward predictors. Note that the fake sequence of motion from the backward predictor is constructed by:

$$\hat{\mathcal{X}}_{m_i}^{(B)} = [y_1 : y_n \backslash (-\hat{y}_i')]. \quad (10)$$

Anomaly detection

For training, the BPNs are optimized by appearance and motion prediction loss forwards and backwards. We minimize the following objective function, which consists of appearance and motion prediction losses (*i.e.*, Eqs. 1 and 2), the multi-modal and two sequence adversarial losses (*i.e.*, Eqs. 6, 8 and 9):

$$\mathcal{L} = \lambda_1\mathcal{L}_a + \lambda_2\mathcal{L}_m + \lambda_3\mathcal{L}_{\text{md}} + \lambda_4\mathcal{L}_{\text{sda}} + \lambda_5\mathcal{L}_{\text{sdm}}, \quad (11)$$

where λ_1 to λ_5 are hyperparameters for balancing five loss functions.

During inference, the video event is extracted from the current frame and four previous frames by an object detector. Each video event produces n different incomplete sequences by erasing the patch at i-th position. The trained BPNs output the predicted appearance ($\hat{x}_i$, $\hat{x}_i'$) and motion ($\hat{y}_i$, $\hat{y}_i'$) for the forwards and backwards direction. The total prediction error S_a and S_m for appearance and motion is defined as:

$$\begin{aligned} S_a &= \sum_{i=1}^{n} w_a\|x_i - \hat{x}_i\|_2^2 + w_a'\|x_i - \hat{x}_i'\|_2^2, \\ S_m &= \sum_{i=1}^{n} w_m\|y_i - \hat{y}_i\|_2^2 + w_m'\|y_i + \hat{y}_i'\|_2^2, \end{aligned} \quad (12)$$

where w_a, w_a', w_m and w_m' indicate the weights of appearance and motion prediction error in forward and backward respectively. Finally, the frame-level abnormal score S based on video event is calculated as follows:

$$S = S_a + S_m. \quad (13)$$

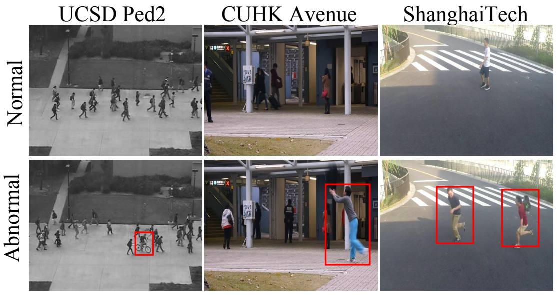


Figure 3: Some example including normal and abnormal frame in UCSD ped2, CUHK Avenue and ShanghaiTech. Red boxes denote abnormal events in abnormal frames.

Details on setting the weights w_a, w_a', w_m, w_m' and the threshold for S by cross-validation to detect abnormal events are discussed in Supplementary.

Experiments

Experimental Setting

Datasets. We experiment on five VAD benchmarks: CUHK Avenue (Lu, Shi, and Jia 2013), UCSD Ped2 (Mahadevan et al. 2010) and ShanghaiTech (Luo, Liu, and Gao 2017) in the main paper. In all the datasets, only the normal class exists in the training data. Some examples of normal and abnormal frames in the datasets are presented in Fig. 3. The optical flow of each frame as motion is computed by FlowNet2 (Ilg et al. 2017).

Evaluation metrics. For UCSD Ped2, CUHK Avenue and ShanghaiTech datasets, We adopt the most frequently-used metric (Mahadevan et al. 2010): The area under curve (AUC) of the receiver operating characteristic (ROC) curve estimated from the frame-level scores.

Implementation. Our model is implemented in PyTorch; we train with the Adam optimizer with a batch size of 128 and a learning rate of 0.0002 for the predictors and 0.00002 for the discriminators. Considering the dataset scale, model is trained by 5, 20, and 30 epochs with a batch size 128 on UCSDped2, Avenue and ShanghaiTech respectively. The anomaly scoring and balancing weights are evaluated via cross-validation (see the supplementaries). Specifically, for anomaly scoring, we set (w_a, w_a', w_m, w_m') to be $(0.5, 0.01, 1, 1.5)$ for UCSD ped2, $(1.2, 0.8, 0.8, 1.2)$, $(1, 0.01, 0.01, 3)$ for Avenue and ShanghaiTech, respectively. We set the weights for balancing five loss functions $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$ to be $(1, 2, 0.25, 1, 1)$. The number of incomplete sequence frames is set to 5, except the ablation study on video event length in next section.

Table 2: Ablation for increasing the number of predictors.

Predictors	UCSD Ped2	CUHK Avenue	Traffic-Train
One forward predictors	97.3%	89.6%	64.2%
Two forward predictors	97.3%	89.5%	64.3%
One forward predictor +One backward predictor	97.9%	90.0%	66.6%

Table 3: Ablation for Laplacian pyramid loss.

Loss Function Combination	UCSD Ped2	CUHK Avenue	Traffic-Train
MSE	98.0%	90.1%	67.1%
MSE + LAP	98.3%	90.3%	68.5%

Ablation Studies

Effectiveness of backward information. We conduct several ablation studies in Tab. 1. Comparing the first and second row, *i.e.*, with versus without the backwards stream adds the improvement of 0.6%, 0.4% and 2.4% for UCSD ped2, CUHK avenue and Traffic-Train datasets respectively. We do an additional comparison to check if the gains come simply from an ensemble effect as adding the backwards stream effectively doubles the number of predictors. Tab. 2 shows that using two forwards predictors does not provide much benefit and performance improvement, which is on par with one forwards predictor. The main reason is the lacking support from backward information.

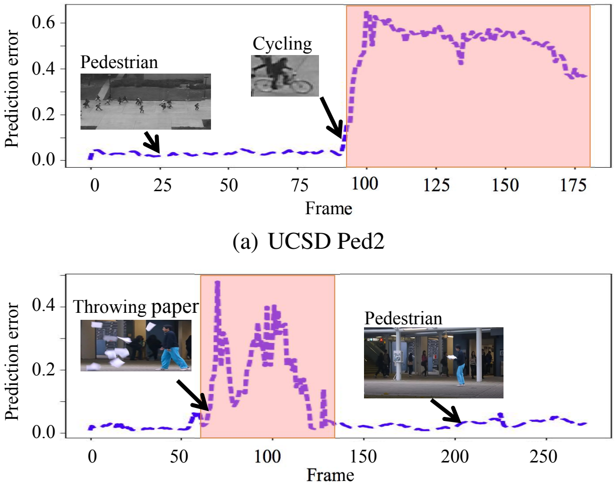
Effectiveness of consistency regularizers. Compared with the proposed model without any discriminator in the second row of Tab. 1, we observe that, in the third row, the sequence-wise discriminator (temporal consistency) brings evident improvement by 0.2%, 0.1% and 1.0% AUC gain on UCSD Ped2, CUHK Avenue, and Traffic-Train respectively. In the last row of Tab. 1, multi-modal discriminators (association consistency) achieve improvement by 0.2%, 0.2%, and 0.9%, compared to that without multi-modal discriminator (*i.e.*, the third row in Tab. 1). We adopt a strategy of typical conditional GAN in our framework. For multi-modal discriminator, it makes sure that the underlying distribution of predicted motion is highly related to its corresponding target appearance (association consistency) and closer to the distribution of ground truth motion patch simultaneously. Besides, for sequence-wise discriminator, it increases the robustness and temporal consistency of predicted frames.

Effectiveness of modality. In the last two rows of Tab. 1, It is clear that the two modalities have different impacts on the performance of the proposed method in these benchmark datasets. In UCSD Ped2, the video quality of a stationary camera mounted at an elevation is too low. In the Traffic-Train, people are blocked by obstacles in the train. The appearance information of the object in both datasets can not be employed to detect the abnormal event effectively. Differently, the videos captured in CUHK avenue are much more clear, which supports the network to learn the normal pattern from spatial–temporal information in the video. Therefore, with motion information, the appearance information could further achieve great improvement to the detection performance, 5.3%, in CUHK avenue dataset.

Effect of predictive consistency loss. The effectiveness of the Laplacian pyramid loss is further verified in Tab. 3. The Laplacian loss is a better measure of image reconstruction according to perceived visual quality than L1 or MSE error. This has been verified in previous works (Bojanowski et al. 2018; Hou and Liu 2019).

Table 4: Ablation for the length of the complete video event.

Length of video event	UCSD Ped2	CUHK Avenue	Traffic-Train
3	93.5%	88.1%	65.7%
5	98.3%	90.3%	68.5%
7	97.1%	89.3%	66.7%
9	96.6%	89.1%	64.5%



(a) UCSD Ped2

(b) CUHK Avenue

Figure 4: Abnormal score (prediction error) curves evolved along time on two example abnormal events in UCSD Ped2 and CUHK Avenue. Red areas contain ground-truth abnormal frames.

Effect of Video event length n**.** We experiment the number of sequence frames n with 3, 5, 7 and 9; in each case, only one frame is erased and the proposed model learns the normal patterns by predicting the erased frame from these incomplete sequences. Tab. 4 shows that with short events (3 frames), we cannot fully exploit the temporal context information and obtain the worst performance. Longer video events of 7 and 9 frames adds computational expense but also learns some patches which are not strongly related to the erased patch. Therefore, we fix the video event at 5 frames, which achieves the best performance in UCSD Ped2, CUHK Avenue, and Traffic-Train.

Comparison with the State-of-the-art

CUHK Avenue, UCSD Ped2 and ShanghaiTech Our method belonging to frame prediction-based methods achieves the state-of-the-art performance on ShanghaiTech, CUHK Avenue, and UCSD Ped2 in Tab. 5. Our method uses additional backward information and several consistencies to regularize the appearance and motion prediction from the incomplete sequence, which improves the performance by 1.0%, 0.7%, and 3.3% AUC, compared with VEC (*i.e.* the

Table 5: Comparison of frame-level performance (AUC) of anomaly detection; The methods are ordered chronologically. Our method achieves the best performance marked in boldfaced and the second ranking performance marked in blue.

Types	Methods	Citation & Year	UCSD Ped2	CUHK Avenue	ShanghaiTech
Reconstruction Based Methods	Memory-guided	(Park, Noh, and Ham 2020)	97.0%	88.5%	70.5%
	Clustering Driven	(Chang et al. 2020)	96.5%	86.0%	73.3%
	SIGnet	(Fang et al. 2020)	96.2%	86.8%	-
	DGN	(Saypadith and Onoye 2021)	93.6%	86.8%	73.0%
	Memory Consistency	(Cai et al. 2021)	96.6%	86.6%	73.7%
Frame Prediction Based Methods	Attention-Prediction	(Zhou et al. 2019)	96.0%	86.0%	-
	VEC	(Yu et al. 2020)	97.3%	89.6%	74.8%
	MONAD	(Doshi and Yilmaz 2021)	97.2%	86.4%	70.9%
	STCEN	(Hao et al. 2021)	96.9%	86.6%	73.8%
	VPC	(Liu et al. 2021)	93.6%	85.4%	-
Hybrid Methods	Skeleton-Trajectories	(Morais et al. 2019)	-	-	73.4%
	AnoPCN	(Ye et al. 2019)	96.8%	86.2%	73.6%
	Prediction&Reconstruction	(Tang et al. 2020)	96.3%	85.1%	73.0%
	sRNN	(Luo et al. 2021)	92.2%	83.5%	69.6%
	Ours		**98.3%**	**90.3%**	**78.1%**

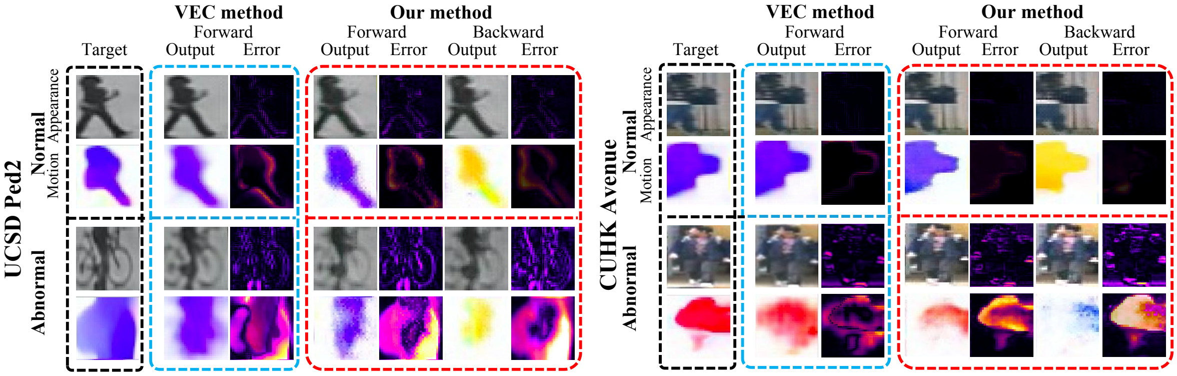


Figure 5: Visualization of erased patches and their optical flow (Target), completed patches (Output) and prediction error (Error) from our method and VEC. Brighter color in Error indicates larger error shown in the rightest bar.

previous best SOTA method) on these benchmark datasets respectively. We plot the prediction errors over time for two sequences from UCSD Ped2 and CUHK Avenue in Fig. 4. The prediction error increases dramatically when abnormal events (cycling and throwing paper) occur and decreases again when the events are complete. Such an observation indicates that our method is effective in detecting the occurrence of anomalies. However, it still has some problems that need to be solved, especially on ShanghaiTech, which is a larger-scale dataset and consists of 13 scenes, several types of anomalies. Previous models have the risk of overfitting to normal training patterns and being sensitive to hard normal patterns. Their method may results in irregular responses to normal data during inference. In our method, to avoid overfitting issues, three consistency constraints to comprehensively regularize the prediction task.

Appearance and Motion Predictions We visually compare the predictions between VEC and our method in Fig. 5. The first two rows of Fig. 5, our model makes high-quality appearance and motion predictions for the normal event of a pedestrian; the errors are also lower compared to VEC. For the abnormal events of bikers in UCSD and people walking in the wrong direction for CUHK, our method has higher prediction errors compared to VEC. Therefore, our method distinguishes between normal and abnormal events more easily and correctly. In addition, it is worth noting that for the appearance prediction, the error for the forward and backward direction are quite similar. For the motion, however, the gap between target and predicted flow is larger for the backward direction than the forwards. In other words, compared with forwards information, the backwards flow plays a more essential role in discriminating between normal and abnormal samples. This is a key point which VAD methods to date have overlooked.

Conclusion

In this paper, we propose a novel bi-directional predictive framework based on video event completion for video anomaly detection. To learn more discriminative representation, we introduce three consistencies to regularize the output prediction from pixel-wise, cross-modal, and temporal-sequence levels. Extensive experiments on five benchmark datasets show superior performance gains over state-of-the-art methods.

Acknowledgments

We thank for the support from National Natural Science Foundation of China (61972157, 62176092, 61876161, 61902129, 62102151, 72192821), Zhejiang Lab (No. 2020NB0AB01), Natural Science Foundation of Shanghai (20ZR1417700), National Key Research and Development Program of China (2019YFC1521104, 2020AAA0108301), Shanghai Municipal Commission of Economy and Information (XX-RGZN01-19-6348), Shanghai Science and Technology Commission No.21511100700, Shanghai Jiao Tong University, Shanghai Sailing Program (21YF1411200), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102). Yuan Xie is supported by CAAI-Huawei MindSpore Open Fund.

References

Abati, D.; Porrello, A.; Calderara, S.; and Cucchiara, R. 2019. Latent space autoregression for novelty detection. In *CVPR*, 481–490. 2

Bojanowski, P.; Joulin, A.; Lopez-Pas, D.; and Szlam, A. 2018. Optimizing the Latent Space of Generative Networks. In *ICML*, 600–609. 4, 6

Cai, R.; Zhang, H.; Liu, W.; Gao, S.; and Hao, Z. 2021. Appearance-Motion Memory Consistency Network for Video Anomaly Detection. In *AAAI*, volume 35, 938–946. 2, 7

Cai, Z.; and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 6154–6162. 3

Chang, Y.; Tu, Z.; Xie, W.; and Yuan, J. 2020. Clustering Driven Deep Autoencoder for Video Anomaly Detection. In *ECCV*, 329–345. 7

Doshi, K.; and Yilmaz, Y. 2021. Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. *Pattern Recognition*, 114: 107865. 7

Fan, H.; Zhu, L.; and Yang, Y. 2019. Cubic LSTMs for video prediction. In *AAAI*, 8263–8270. 2

Fang, Z.; Liang, J.; Zhou, J. T.; Xiao, Y.; and Yang, F. 2020. Anomaly Detection With Bidirectional Consistency in Videos. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14. 7

Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M. R.; Venkatesh, S.; and Hengel, A. v. d. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *ICCV*, 1705–1714. 1, 2

Hao, Y.; Li, J.; Wang, N.; Wang, X.; and Gao, X. 2021. Spatiotemporal consistency-enhanced network for video anomaly detection. *Pattern Recognition*, 121: 108232. 7

Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A. K.; and Davis, L. S. 2016. Learning temporal regularity in video sequences. In *CVPR*, 733–742. 2

Hou, Q.; and Liu, F. 2019. Context-aware image matting for simultaneous foreground and alpha estimation. In *ICCV*, 4130–4139. 6

Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; and Brox, T. 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2462–2470. 5

Kieu, T.; Yang, B.; Guo, C.; and Jensen, C. S. 2019. Outlier Detection for Time Series with Recurrent Autoencoder Ensembles. In *IJCAI*, 2725–2732. 2

Ling, H.; and Okada, K. 2006. Diffusion distance for histogram comparison. In *CVPR*, 246–253. 4

Liu, B.; Chen, Y.; Liu, S.; and Kim, H.-S. 2021. Deep Learning in Latent Space for Video Prediction and Compression. In *CVPR*, 701–710. 7

Liu, W.; Luo, W.; Lian, D.; and Gao, S. 2018. Future frame prediction for anomaly detection–a new baseline. In *CVPR*, 6536–6545. 1, 2

Lu, C.; Shi, J.; and Jia, J. 2013. Abnormal event detection at 150 fps in matlab. In *ICCV*, 2720–2727. 3, 5

Lu, Y.; Kumar, K. M.; shahabeddin Nabavi, S.; and Wang, Y. 2019. Future frame prediction using convolutional VRNN for anomaly detection. In *AVSS*, 1–8. 1, 2

Luo, W.; Liu, W.; and Gao, S. 2017. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *ICCV*, 341–349. 3, 5

Luo, W.; Liu, W.; Lian, D.; Tang, J.; Duan, L.; Peng, X.; and Gao, S. 2021. Video Anomaly Detection with Sparse Coding Inspired Deep Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3): 1070–1084. 7

Mahadevan, V.; Li, W.; Bhalodia, V.; and Vasconcelos, N. 2010. Anomaly detection in crowded scenes. In *CVPR*, 1975–1981. 5

Morais, R.; Le, V.; Tran, T.; Saha, B.; Mansour, M.; and Venkatesh, S. 2019. Learning regularity in skeleton trajectories for anomaly detection in videos. In *CVPR*, 11996–12004. 7

Nguyen, T.-N.; and Meunier, J. 2019. Anomaly detection in video sequence with appearance-motion correspondence. In *CVPR*, 1273–1283. 1

Park, H.; Noh, J.; and Ham, B. 2020. Learning memory-guided normality for anomaly detection. In *CVPR*, 14372–14381. 7

Perera, P.; and Patel, V. M. 2019. Learning deep features for one-class classification. *TIP*, 28(11): 5450–5463. 1

Radford, A.; Metz, L.; and Chintala, S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*. 4

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. 3

Saypadith, S.; and Onoye, T. 2021. Video Anomaly Detection Based on Deep Generative Network. In *ISCAS*, 1–5. 7

Tan, X.; Xu, K.; Cao, Y.; Zhang, Y.; Ma, L.; and Lau, R. W. H. 2021. Night-time Scene Parsing with a Large Real Dataset. *IEEE Transactions on Image Processing*, 30: 9085–9098. 3

Tang, Y.; Zhao, L.; Zhang, S.; Gong, C.; Li, G.; and Yang, J. 2020. Integrating prediction and reconstruction for anomaly detection. *PR*, 129: 123–130. 2, 7

Taylor, W. L. 1953. Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30(4): 415–433. 2

Tran, H. T.; and Hogg, D. 2017. Anomaly detection using a convolutional winner-take-all autoencoder. In *BMCV*. 2

Yan, S.; Smith, J. S.; Lu, W.; and Zhang, B. 2018. Abnormal event detection from videos using a two-stream recurrent variational autoencoder. *TCDS*, 12(1): 30–42. 2

Ye, M.; Peng, X.; Gan, W.; Wu, W.; and Qiao, Y. 2019. Anopcn: Video anomaly detection via deep predictive coding network. In *ACMMM*, 1805–1813. 7

Yu, G.; Wang, S.; Cai, Z.; Zhu, E.; Xu, C.; Yin, J.; and Kloft, M. 2020. Cloze Test Helps: Effective Video Anomaly Detection via Learning to Complete Video Events. In *ACMMM*, 583–591. 1, 2, 7

Zhou, J. T.; Du, J.; Zhu, H.; Peng, X.; Liu, Y.; and Goh, R. S. M. 2019. AnomalyNet: An Anomaly Detection Network for Video Surveillance. *TIFS*, 14(10): 2537–2550. 1

Zhou, J. T.; Zhang, L.; Fang, Z.; Du, J.; Peng, X.; and Xiao, Y. 2019. Attention-driven loss for anomaly detection in video surveillance. *TCSVT*, 30(12): 4639–4647. 1, 2, 7