sensors

MDPI

Article

A Novel Unsupervised Video Anomaly Detection Framework Based on Optical Flow Reconstruction and Erased Frame Prediction

Heqing Huang [1,†], Bing Zhao [2,†], Fei Gao [1,3], Penghui Chen [1,*], Jun Wang [1] and Amir Hussain [4]

[1] School of Electronic and Information Engineering, Beihang University, Beijing 100190, China; huangheqing@buaa.edu.cn (H.H.); feigao2000@163.com (F.G.); wangj203@buaa.edu.cn (J.W.)
[2] Inspur Electronic Information Industry Co., Ltd., Beijing 100085, China; zhaobing0412@outlook.com
[3] Beihang Hangzhou Innovation Institute Yuhang, Xixi Octagon City, Yuhang District, Hangzhou 310023, China
[4] Cyber and Big Data Research Laboratory, Edinburgh Napier University, Edinburgh EH11 4BN, UK; a.hussain@napier.ac.uk
* Correspondence: chenpenghui@buaa.edu.cn
† These authors contributed equally to this work.

Abstract: Reconstruction-based and prediction-based approaches are widely used for video anomaly detection (VAD) in smart city surveillance applications. However, neither of these approaches can effectively utilize the rich contextual information that exists in videos, which makes it difficult to accurately perceive anomalous activities. In this paper, we exploit the idea of a training model based on the “Cloze Test” strategy in natural language processing (NLP) and introduce a novel unsupervised learning framework to encode both motion and appearance information at an object level. Specifically, to store the normal modes of video activity reconstructions, we first design an optical stream memory network with skip connections. Secondly, we build a space–time cube (STC) for use as the basic processing unit of the model and erase a patch in the STC to form the frame to be reconstructed. This enables a so-called ”incomplete event (IE)” to be completed. On this basis, a conditional autoencoder is utilized to capture the high correspondence between optical flow and STC. The model predicts erased patches in IEs based on the context of the front and back frames. Finally, we employ a generating adversarial network (GAN)-based training method to improve the performance of VAD. By distinguishing the predicted erased optical flow and erased video frame, the anomaly detection results are shown to be more reliable with our proposed method which can help reconstruct the original video in IE. Comparative experiments conducted on the benchmark UCSD Ped2, CUHK Avenue, and ShanghaiTech datasets demonstrate AUROC scores reaching 97.7%, 89.7%, and 75.8%, respectively.

Keywords: video anomaly detection; optical flow; incomplete event

check for updates

Citation: Huang, H.; Zhao, B.; Gao, F.; Chen, P.; Wang, J.; Hussain, A. A Novel Unsupervised Video Anomaly Detection Framework Based on Optical Flow Reconstruction and Erased Frame Prediction. *Sensors* **2023**, *23*, 4828. https://doi.org/10.3390/s23104828

Academic Editor: Alessandro Leone

Received: 26 April 2023
Revised: 14 May 2023
Accepted: 15 May 2023
Published: 17 May 2023

1. Introduction

With the development of sensors, video surveillance and sensor networks are widely used in various fields, such as traffic monitoring, environmental monitoring, industrial control, etc. [1–3]. The goal of video surveillance and sensor networks is to achieve accurate monitoring and detection of anomalous targets. When analyzing real-world datasets, a common requirement is that those instances can be detected as distinct. These instances are referred to as anomalies because they are significantly different from other normal data. The anomalies may be an operation error during collecting datasets or they may originate from unknown processes that do not conform to common sense. An example of such anomaly consists of vehicles traveling on campus roads that are usually considered abnormal. The anomaly detection task is to detect these instances that do not match the expected pattern in the dataset. To accomplish this task, artificial intelligence computer vision algorithms for automatically detecting anomalies in datasets is a viable solution [4–7].

Nowadays, video surveillance is playing a more and more crucial role in smart city monitoring applications. There is a growing interest in VAD [8–10], which aims to identify events that do not meet normal behavior by interpreting the video. This technology can predict normal/abnormal events in a given video. Using video anomaly detection technology, smart city surveillance systems can automatically detect abnormal events, such as traffic congestion, criminal activities, fires, etc. Therefore, there are important application values in some scenes such as municipal and traffic management [11].

Deep learning approaches have made great progress in many computer vision tasks in recent years(e.g., object detection [12–15] and semantic segmentation [16–19]). In related work, many anomaly detection methods based on deep learning have been performed in different ways for detecting anomalies. Many studies have used supervised learning approaches [20–22] to tackle the anomaly detection task by labeling each frame of videos to clearly define the anomaly. These studies have classified normal/abnormal samples by training a convolutional neural network. However, the inter-class imbalances and intra-class imbalances in VAD interfere with the model training process; therefore, it is challenging to obtain a representative dataset of outliers, which makes the direct application of supervised learning methods difficult.

Some studies have used unsupervised learning methods for further analysis of VAD. For learning normal activities features, the usual solution is to train a model with continuous normal samples X = [x_1, x_2, x_3....x_n]. When the model is used to analyze samples in the test set, samples that do not conform to the normal mode are called anomalous samples. At present, the mainstream video anomaly detection methods based on unsupervised learning can be divided into the following two types.

(1) Reconstruction based [23–25]. The reconstruction errors of normal samples are tiny, but the reconstruction errors of abnormal samples are large. So, we employ an auto-encoder to reconstruct the input and set a threshold to judge whether it is abnormal.
(2) Prediction based [8,26]. The model learns from the existing video frames and predicts the future frames. If there is a large difference between the prediction frame and the ground truth, the future frames are determined to be abnormal.

Although reconstruction-based/prediction-based methods have achieved good performance (more than 90% in the UCSD Ped2 dataset [27]), we consider them to have the following drawbacks.

(1) The existing studies usually process the whole image of video frames. When the objects to be detected are relatively small, they have a minor impact on the detection of abnormal events. Ideally, the precise localization of anomalous activities requires a complete representation of the subject at the object level.
(2) Although the abnormal events correspond to larger reconstruction errors, due to the problem of "over-generalizing" in deep neural network, abnormal events may also be reconstructed.
(3) Prediction-based methods can predict the future video frames from the given first few frames in the video sequence; then, the prediction error of a single frame can be regarded as a measurement for anomalies. However, these methods cannot make full use of the temporal context [28–30] and high-level semantic information of video anomaly activities.
(4) Motion is the key feature for understanding videos. Current studies generally deal with images without considering the motion information in the frames. This motion information can be provided by optical flow. Optical flow can use the change of pixels in the image sequence and the correlation with adjacent frames to find the corresponding relationship between the previous frame and the current frame.

In this paper, we approach these problems from a new perspective. Instead of training with the entire video frames, we apply an object detection algorithm to extract the objects and construct STCs [31–33]. The STCs contain not only the objects of interest in the video frames, but also the bounding boxes of the same foreground in consecutive frames so as to eliminate the possible influence of image background. In Bert [34] and MAE [35], these models recover the occluded segments/images in a logical way according to the gradually learned context. This idea is somewhat similar to the "Cloze Test", and we use a similar idea to construct "incomplete events" in VAD.

Specially, we erase a patch in STCs and let the model predict the erased frame according to the contextual information. Then, we extract the STCs of optical flow data and reconstruct the optical flow. Without loss of generality, we utilize an auto-encoder with storage mechanism in which the features of different layers are fused through skip connections. Motion information is used to predict the erased video frames in IE to prevent abnormal events from being reconstructed. Finally, we train the generator and discriminator to distinguish the normal mode and the prediction of optical flow. The major contributions of this paper are as follows:

(1) We utilize a multi-level memory auto-encoder with skip connections to reconstruct video optical flow. By combining motion information (optical flow) and appearance information (object detection), the model can provide high-level semantics as auxiliary information to analyze the motion of video frames.
(2) In order to make full use of the temporal context information of a video, we employ the idea of the "incomplete event" to predict the erased frames in the video, rather than the usual method based on future frame prediction.
(3) We exploit the GAN training method and use the conditional auto-encoder as the generator. The model applies two discriminators to classify the predicted erased frames and optical flow to further improve the performance of the VAD model.
(4) Finally, we conduct many experiments on the UCSD Ped2 [27], CUHK Avenue [36], and ShanghaiTech datasets [26]. The results show the advanced performance of our proposed method.

The overall structure of this paper is as follows. In Section 2, we describe the previous research on anomaly detection. Then, we design the VAD model based on hybrid tasks in Section 3. Section 4 introduces the detailed experimental setup and results of this paper. In order to analyze the effectiveness of specific components, we discuss the internal details of the model proposed in Section 5. Section 6 summarizes this paper.

2. Related Work

In this section, we describe different anomaly detection methods and summarize traditional machine learning methods (Section 2.1), reconstruction-based methods (Section 2.2), and prediction-based methods with various auxiliary information sources (Section 2.3).

2.1. Traditional Machine Learning Methods

Traditional anomaly detection models often need to design features manually to realize anomaly detection [27,37]. Early works usually extracted low-level trajectory features [38], but these approaches are not applicable in complex scenes such as occlusion. Ref. [39] proposed a hybrid model based on high-dimensional features which used the support vector machine (SVM) for normal/abnormal sample classification through unsupervised learning. The method has high scalability and computational efficiency. In addition to simple feature extraction, the help of other machine learning methods can also improve the effect of anomaly detection. Wu et al. [37] introduced chaotic dynamics into pedestrian activity events then analyzed pedestrian abnormal behavior by extracting chaotic invariant features and aggregating particle trajectories. Piciarelli et al. [40,41] analyzed the existing problems with clustering motion trajectories, which added probability information to improve the accuracy of anomaly detection. Morris et al. [42] used a hidden Markov model and maximum likelihood estimation to judge the motion route of interested nodes, which was used to forecast future activities and detect anomaly events in real-time.

2.2. Method of Reconstruction Error

Recent works contain many methods of deep learning, and reconstruction-based methods have become the mainstream. The basic assumption of the methods is that the reconstruction error of normal sample is low, whereas the reconstruction error of abnormal samples is high [43–45]. The neural network used for the reconstruction of input data is usually either an auto-encoder or GAN [46,47]. They can encode the inputs into a more compact representation and retain basic features to ensure that the reconstructed image is close to reality. Therefore, by taking advantage of auto-encoder and GAN, [24] used an auto-encoder with a convolution LSTM to model normal video frames and motion information at the same time. This method further improved the performance of the anomaly detection model. In addition, transfer learning can provide rich a priori knowledge. Salehi et al. [48] pre-trained with the ImageNet dataset to extract intermediate knowledge, then utilized the difference between the given input data and the model activation value to detect anomalies. In another work, Li et al. [49] applied a cut–paste data enhancement strategy to simulate abnormal samples and trained convolutional neural network to identify normal/abnormal samples. In the first stage, CutPaste was used to generate images for normal samples through self-supervised learning characterization. In the second stage, CNN was exploited to extract the features and calculate the anomaly score for the output features. Ref. [50] proposed a SAR image anomaly detection method based on iterative outliers and significant recursive depth, which can effectively distinguish the front background in the image and obtain an appropriate anomaly detection threshold.

2.3. Method of Frame Prediction

Many studies [8,51,52] have used video frame prediction methods to deal with the VAD problem. Chong et al. [53] introduced an unsupervised method to model video frames by convolution auto-encoder, which integrated data into a video representation and learned the temporal pattern of rules. The predicted video frame will significantly differ from the original frame when an abnormal event occurs. Feng et al. [54] proposed a convolutional transformer to predict future frames and used the double discriminator GAN training method to assist in the generation of the prediction frame. The model used prediction error to assist in the identification of abnormal video frames and achieved good results on different datasets. Ref. [55] expanded the Viola Jones algorithm to detect faces in videos and perform target tracking. By extracting the object level features and speed level features of the target, the model can apply these features to the classifier to complete VAD. In other frameworks, optical flow has also been used to assist RGB images for video anomaly detection. Ref. [56] recommended predicting intermediate frames in videos. The model is mainly composed of two parts: a middle frame predictor and an appearance detector, which considers the appearance and motion characteristics of video scene. Furthermore, based on the reconstructed optical flow, [33] designed a conditional variational auto-encoder to capture the correlation between video frames and optical flow field. The approach used this correlation to enhance the quality of frame prediction. Liu et al. [26] used optical flow features to assist in constraining the prediction of future frames in the video and forced the optical flow of predicted frames to be consistent with that of actual frames. Leyva et al. [57] designed a reasoning mechanism based on optical flow features and foreground key descriptive features to detect abnormal events. Ref. [58] proposed an optical flow feature clustering method with motion information to assist in anomaly detection which efficiently captured the directional motion information in the surveillance video.

3. The Proposed Methods

As shown in Figure 1, the pipeline proposed in this study consists of two parts: video optical flow reconstruction based on a storage mechanism and erased frame prediction based on GAN. On the one hand, we choose an object detection algorithm (Cascade R-CNN) [59] to extract local features and construct a large number of STCs for training data. This is where we explicitly focus on the object present in the video frame scene. This object detector was chosen because:

(i) it uses a recurrent head network, which can efficiently detect small objects in images;
(ii) the algorithm achieves a good balance between detection speed and accuracy.

By constructing STC for detected objects, then we perform optical flow extraction in them. The auto-encoder with the memory module is used to reconstruct optical flow.

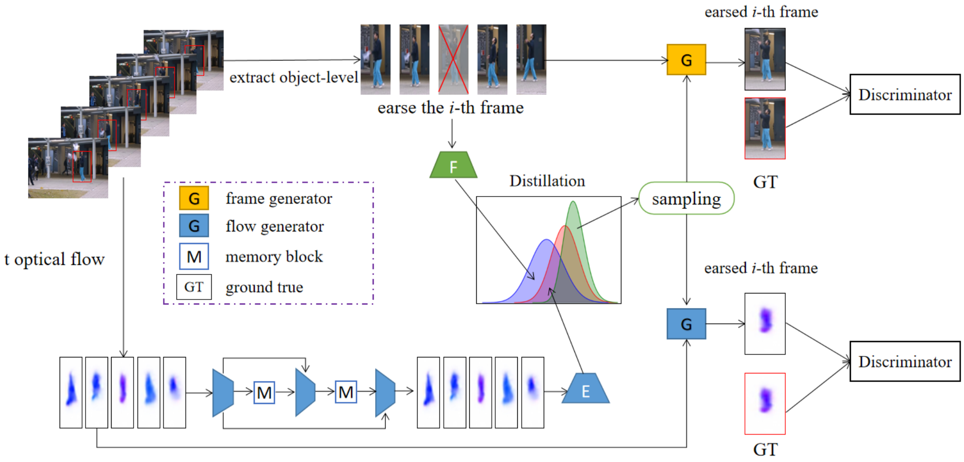


Figure 1. The anomaly detection framework proposed in this study includes the following contents. Firstly, we mask the patch in the video frame sequence. Then, the video sequence is fed to an optical flow network to extract motion information. The optical flow is reconstructed through an auto-encoder with the storage module. We sample from the distribution and input it into the variational auto-encoder to generate the prediction of the erased frame. Finally, we use a discriminator to classify it.

On the other hand, to learn more about video temporal context information, we construct IE to predict the erased patch in STC. Two distillation methods are proposed to make the reconstructed optical flow distribution as close as possible to the normal frame distribution to achieve accurate and comprehensive positioning. Next, we add two discriminators to classify the prediction of the erased frame and the reconstruction of optical flow. The whole framework is trained on normal data. During the testing, we extract five consecutive frames from the test data then generate the erased patch to complete IE and achieve VAD. Finally, we will describe the contents of each part of the framework and represent the standards for VAD.

3.1. Video Optical Flow Reconstruction with Memory Mechanism

The auto-encoder consists of two parts: the encoder and the decoder. The encoder can extract low-dimensional representations from high-dimensional data, and the decoder can use the learned representations to reconstruct the original data. For VAD models, auto-encoders usually aim to minimize the represented reconstruction error as the training objective of the network.

However, the "over-generalizing" of deep neural networks often gives auto-encoders the ability to reconstruct abnormal images. Therefore, we thought of adding a memory module to the auto-encoder to design a clearer data encoding method. The method incorporates memory module into the encoder and features the data by storage factors with different weights. The model places the representations of historic normal samples, then searches for the top k representations with the highest similarity to the new samples representations in memory. The historic representation multiplied by a weighted sum is used to represent the samples representation. The purpose of this is that abnormal samples can only be weighted by historic normal samples representations, which makes the reconstruction error of the anomaly samples larger and easier to distinguish.

Figure 2 shows the memory module in the auto-encoder which is composed of an encoder–decoder and memory enhancement modules. The encoder consists of convolutional blocks and downsampling layers. The decoder contains upsampling layers and convolution blocks, and it is here where we add the fusion of encoder features process. Each convolution block contains a convolution layer, a batch normalization layer, and an activation layer. We use this structure to reconstruct the optical flow motion information. Firstly, the optical flow frame is encoded to obtain the output z and the memory module is used to convert z into a query Z^*. The query obtained from the memory network is shown in Formula (1).

$$Z^* = \sum_{i=1}^{n} w_i m_i \tag{1}$$

where w_i is the attention coefficient obtained from the output of the vector encoder. Each row vector m_i represents a memory item with a dimension C equal to the dimension of encoder output feature. In this study, C is 2000. The calculation method is shown in Formulas (2) and (3). Specifically, memory is defined as a matrix containing n row vectors.

$$w_i = \frac{exp(d(z, m_i))}{\sum_{j=1}^{N} exp(d(z, m_j))} \tag{2}$$

$$d(z, m_j) = \frac{Zm_i^T}{||Z||||m_i||} \tag{3}$$

To avoid the interference of unnecessary information, we use the most representative normal pattern in the top k record matrix. The stored values are reassigned weights and the final hidden vector w_i^* is obtained for subsequent decoder operations, as shown in Formula (4).

$$w_i^* = \frac{w_i}{max_k w_i^*} \tag{4}$$

When the top k with the highest similarity in memory are found in the new samples, the reconstruction error of the new samples is calculated.

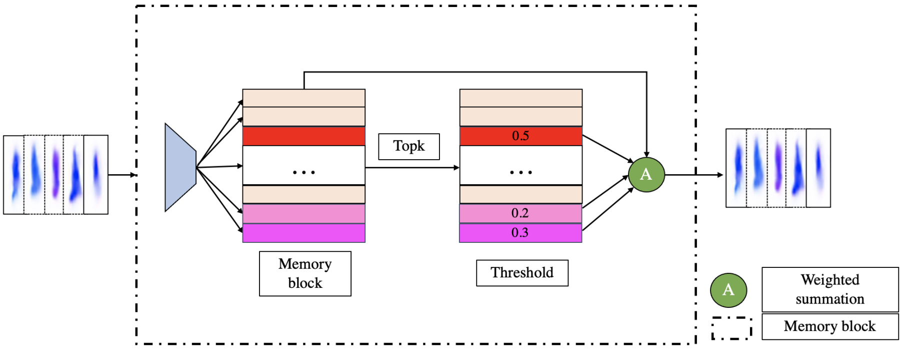


Figure 2. Structure of auto-encoder with memory modules. This memory block appears multiple times in the structure, and we fuse the blocks using skip connections to provide richer information to the model.

To provide more information for optical flow memory, we set up a skip connection. By fusing the features obtained from different storage modules, the features of the extracted normal training data are reused. In the encoder of each memory module, at each level of the encoder, we set up a convolution block and three down-sampling layers in which the convolution block consists of a convolution layer, a batch normalization layer, and an activation layer. The storage module uses the maximum value of a similar storage slot to represent the features that are input into it, and the resulting vector represents a more informative representation of the data. For anomalous data, it is more difficult to obtain a smaller reconstruction error.

We set L2 distance as the reconstruction loss: suppose x is the input data and x^* is the reconstruction result, as shown in Formula (5).

$$L_{flow_recon} = ||x - x^*||_2^2 \tag{5}$$

3.2. Erased Frame Prediction Based on Hard Distillation

Previous studies usually complete anomaly detection based on error reconstruction and future frame prediction. These methods often cannot fully use the temporal context information of video and the high-level semantic features between adjacent frames.

In this study, we extract the training data region of interest from the video frames and construct STCs as the primary input of the model rather than the video frame of the standard method. The extracted contents are all the relevant subjects with active behavior in the frame, which make the model's focus more individualized. Specifically, to extract the appearance information of the training data, we use the Cascade R-CNN [59] model pretrained on the COCO dataset to extract the region of interest in the video frame. Taking the extracted STCs as a set of sequences X = [x_1, x_2....., x_n], we use this method to mask $x_{n+1/2}$ and predict.

To better integrate appearance and motion information, we utilize the storage module to model the reconstructed of optical flow directly mapped to standard video frames. On the one hand, the reconstructed optical flow is unified into the model. On the other hand, the consistency of optical flow and video frames are encoded. The prediction accuracy is improved by trying this new mapping relationship.

Firstly, we model continuous video frames assuming that the distribution in normal mode is $p(\{x \mid x_-\{1:t\}\})$. It is worth noting that in $x_-\{1:t\}$, we erase the frames in the video sequence. At the same time, given the motion information $y_-\{1:t\}$ of the optical flow as an auxiliary condition, it is assumed that its contents have the same hidden variable z.

The conditional auto-encoder is used as the generation model directly to produce $p(\{x \mid x_-\{1:t\}, y_-\{1:t\}\})$. Based on the implicit variable Z, we use an encoder to encode the optical flow and obtain the a priori distribution of $p(\{z \mid y_-\{1:t\}\})$. The model uses another encoder F to process the information combining ordinary video frames and optical flow frames. The input of the encoder is the series of optical flow and normal video frames. After obtaining the posterior distribution of $q(z \mid x_-\{1:t\}, y_-\{1:t\})$, we sample from the posterior distribution and send it to the decoder to help predict the erased frames in IE.

We define the information distribution including the video frames and optical flow frames as the teacher network and the optical flow information distribution as the student network. In order to make the two distributions as similar as possible, we propose two distillation methods: soft distillation and hard distillation. Soft distillation can minimize the KL difference between teacher network and student network, which limits the various parts of the network as much as possible. Hard distillation distribution can deal with the global information of two distributions and take the output of teacher network as a label for student network learning. The calculation method of soft distillation is shown in Formula (6):

$$L_{SoftDis} = (1-\lambda)L_{CE}(P, y) + \lambda\tau^2 KL(P, Q) \tag{6}$$

where P is $p(\{z \mid y_-\{1:t\}\})$, Q is $q(z \mid x_-\{1:t\}, y_-\{1:t\})$. In the hard distillation distribution, we try to take the normal video frame distribution as the real label. Equation (7) is the loss function related with hard distillation.

$$L_{HardDis} = L_{CE}(P, y) + L_{CE}(P, y_t) \tag{7}$$

where L_{CE} is CE loss, y is the tag value, and y_t is shown in Formula (8). These two distillation methods can allow the two distributions to learn better paradigms by learning from each other to reduce their differences.

$$y_t = argmax_c Z_t(c) \tag{8}$$

In order to make the images have a better generation effect, we add a discriminator to judge the reconstruction effect of the erased frame. We use the erased frame prediction generating model as G. Specifically, G contains an encoder and a decoder. In the encoder, each downsampling block contains convolutional layers, batch normalization layers, and activation layers. The decoder consists of convolutional layers, batch normalization layers and dropout layers. For D, we employ the discriminator in PixelGan [60]; it first splices the input image and the features generated by the encoder and obtains the output through the combination of three downsampling and convolutional layers, batch normalization layers, and padding layers. The discriminator's loss function is shown in Formula (9).

$$L^{D}_{adv} = \sum_{i,j} \frac{1}{2} L_{MSE}(D(I)_{i,j}, 1) + \frac{1}{2} L_{MSE}(D(I)_{i,j}, 0) \tag{9}$$

During training, G and D are trained separately. The goal of training D is to judge the authenticity of the reconstructed frame. This method uses a loss function, as shown in Formulas (10) and (11). Among them, Mean Square Error (MSE) is a common method to measure the quality of the predicted image by calculating the Euclidean distance between the predicted values of all pixels in RGB color space and their real values. The goal of the training G is to generate erased frames where the weights of the discriminator are fixed. We use the MSE loss function to optimize.

$$L^{G}_{adv} = \sum_{i,j} \frac{1}{2} L_{MSE}(D(I)_{i,j}, 1) \tag{10}$$

$$L_{MSE}(Y^*, Y) = (Y^* - Y)^2 \tag{11}$$

3.3. Criteria for Abnormal Event Detection

The test standards of VAD have always been a problem discussed by researchers. In this study, our anomaly score consists of two parts:

(1) reconstruction error based on optical flow;
(2) error based on frame prediction.

The abnormal fraction of optical flow is calculated as shown in Formula (12). Among them, $y^*_{1:t}$ and $y_{1:t}$ are the optical flow characteristics before and after reconstruction.

$$S_r = ||y^*_{1:t} - y_{1:t}|| \tag{12}$$

In (2), we employ the predicted generated frame I^*_i and the real label I_i to calculate the error. We do not choose the L1/L2 loss function. When the input value of the function is far from the real value, the corresponding loss value is on both sides, which may cause the gradient to explode. As the literature shows, the peak signal-to-noise ratio (PSNR) is a better image quality evaluation method [61]. If the predicted video frame is normal, the method has a high PSNR value. Compared with L1/L2 loss functions, PSNR has more stable gradients and stronger structural constraints in image generation. We calculate the value of PSNR to judge whether the generated image is abnormal. The calculation formula is shown in Equation (13).

$$PSNR(I, I^*) = 10log_{10} \frac{[max_I]^2}{\frac{1}{N}\sum_{i=0}^{N}(I_i - I^*_i)^2} \tag{13}$$

4. Experiments

4.1. Introduction to the Datasets

UCSD ped [27]: In 2010, Mahadevan et al. established the UCSD pedestrian anomaly detection dataset. The dataset was collected by installing cameras at the height of the school to capture the scene of pedestrian streets. The abnormal behaviors that were collected by the dataset include the circulation of vehicles on the sidewalk and the occurrence of abnormal pedestrian movement, for example, cyclists, skaters, pets, etc. The UCSD dataset contains two subsets: Ped1 and Ped2. Ped1 shows how people move towards and away from the camera. There are 34 video clips in the training set and 36 video clips in the test set, with 234 × 159 pixels per frame. The Ped2 dataset is a scene where pedestrians move parallel to the camera plane. The training set and the testing set contain 16 and 14 video clips, respectively, with a resolution of 360 × 240 pixels. According to the dataset statistics, about 3400 frames include abnormal events, and the other 5500 frames are normal events.

CUHK Avenue [36]: The dataset is collection of a total of 47 anomalies in a school campus. The main anomalies include the abnormal behavior of pedestrians, the wrong direction of movement, the emergence of abnormal objects, and so on. The entire dataset contains 16 training videos and 21 testing videos, with 15,328 frames in the training set and 15,324 frames in the testing set at 640 × 360 pixels.

ShanghaiTech [26]: the production of this dataset solves the problems of the existing datasets, which are relatively focused on single aspects. The ShanghaiTech dataset collected 13 different abnormal scenes and more abnormal behavior types than the previous datasets. These scenes are shot from different angles under different lighting conditions. The main anomalies are retrograde, noise, chasing, etc. The training set contains 330 videos, and the testing set contains 107 videos, each frame with the pixels of 856 × 480.

4.2. Implementation Details and Evaluation Criteria

Hardware equipment. All experiments in this study were conducted on an NVIDIA RTX 2080ti GPU (Santa Clara, CA, USA) and operating system on Ubuntu 18.04. The PyTorch version used was 1.5, and the deep learning acceleration library cudnn 10.0 was used.

Implementation details. We utilized the Cascade R-CNN based on mmdetection to extract the foreground images of all datasets, forming the foreground region sequence of t frame. These regions of interest were used to build STCs. The size of each region was normalized to 32 × 32 pixels. In this study, t = 5. We used FlowNet2-SD to extract the features of optical flow and construct the STCs of optical flow according to the same method. The final distillation strategy used by the model is hard distillation. We first trained the video optical flow reconstruction model with the memory mechanism. The top value in the memory block was set to 20 to provide salient features to the auto-encoder. The batch size was initialized to 256, the learning rate was set to 0.0001, and we adopted an Adam optimizer ($\beta1$ = 0.9, $\beta2$ = 0.999). Then, we trained the generation model of the variational auto-encoder and two discrimination models. The batch size of the generation model was 128, we trained the model for 100 epochs in total. The learning rates of the generator and discriminator were initialized to 0.0001 and 0.00001, respectively, which were adjusted by the warmup strategy.

Evaluation indicators. This paper is also applicable to the popular evaluation indexes of anomaly detection models [26,50]. There are two criteria for evaluating whether a video frame is abnormal. The first criterion is the frame level standard. In this standard, if at least one pixel of a frame is marked as abnormal, the frame is regarded as abnormal. The second criterion is the pixel level standard. If at least 40 abnormal pixels are detected in a single frame, the frame is regarded as abnormal. By combining these two standards, the area under the curve (AUC) of the receiver operating characteristic curve (ROC) is calculated to measure the final performance of the model. The advantage of AUROC is that it can effectively balance the frame-level and pixel-level anomalies and eliminate the influence of threshold as much as possible. This measure directly reflects the sorting quality of the algorithm to the test samples. The higher the value, the better the result of video anomaly detection. In order to ensure the accuracy of AUROC, the model must be able to accurately find the difference between the samples tested and the training samples.

4.3. Experimental Result

To prove the efficiency of the method proposed in this paper, we compare our model with the SOTA model, including reconstruction-based methods such as R-VAD [45], Conv-VAD [39], MEM-VAD [62], LAD [10], GMFC-VAE [63], MemAE [64], and C-VAD [65]; future frame prediction-based methods such as VEC [66], FPVAD [26], CPNet [9], ConvVRNN [67], Attention-VAD [68], D-VAD [69], and S-VAD [70]; and methods based on multiple information sources such as ASSVAD [71], MPED-RNN [72], ST-CAE [73], AnoPCN [74], and PR-AD [75].

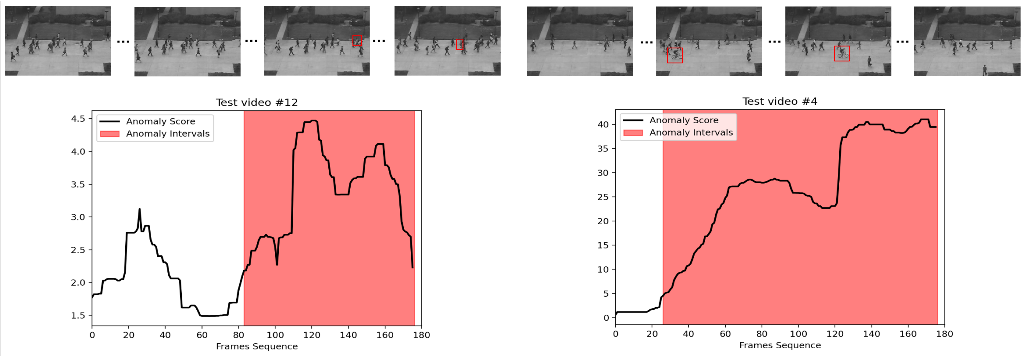
The results are shown in Table 1. It can be observed that our method achieves excellent performance on all three datasets when compared to the state-of-the-art methods, which demonstrates the advancements made by our method. Specifically, this method can reach a 97.7% AUROC score in the Ped2 dataset. In the above literature, the best VAD result for a reconstruction-based method is 95.1%, the best future frame prediction-based [66] method's result is 97.3%, and the best method that integrates multiple information sources reached a result of 97.2%. Our method is higher than the results of these methods. In the CUHK Avenue dataset, our method provided an absolute gain of 0.4%, 0.1%, and 3.3% over the best results of the above three methods, respectively. The VAD result of this paper on the ShanghaiTech dataset is 75.8%, which is higher than 74.8% in the VEC [66] literature. We believe that this is the effect of the combination of optical flow assistance and temporal context information. Our pipeline combines the advantages of the above two parts and achieves a better video anomaly detection effect.

Table 1. Video abnormal detection results (in %) in terms of AUROC score on the UCSD Ped2, CUHK Avenue, and ShanghaiTech datasets. Our method is compared with various start-of-the-art approaches, which are sorted by type of method.

Category	Method	Ped2	CUHK Avenue	ShanghaiTech
Restructure	R-VAD [45]	92.2	81.7	68.0
	Conv-VAD [39]	90.0	70.2	60.9
	MEM-VAD [62]	90.2	82.8	69.8
	LAD [10]	95.1	89.3	/
	GMFC-VAE [63]	92.2	83.4	/
	MemAE [64]	94.1	83.3	69.8
	C-VAD [65]	87.5	84.4	
Prediction	VEC [66]	97.3	89.6	74.8
	FPVAD [26]	95.4	85.1	72.8
	CPNet [9]	96.1	85.1	/
	ConvVRNN [67]	96.1	85.8	/
	Attention-VAD [68]	96.0	86.0	/
	D-VAD [69]	95.6	84.9	73.7
	S-VAD [70]	/	89.6	74.7
Hybrid Frame	ASSVAD [71]	96.7	86.4	71.6
	MPED-RNN [72]	/	/	73.4
	ST-CAE [73]	91.2	80.9	/
	AnoPCN [74]	96.8	86.2	73.6
	PR-AD [75]	96.3	85.1	73.0
	Ours	**97.7**	**89.7**	**75.8**

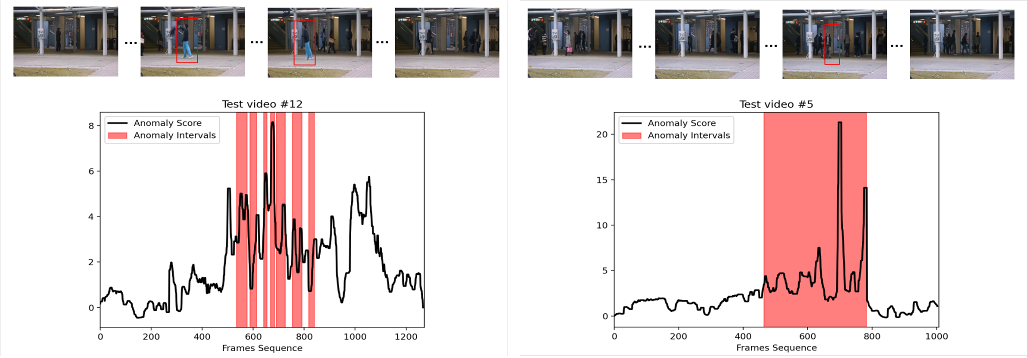
Figure 3 shows the effect of VAD on the UCSD Ped2 dataset. We visualized a two-frame video sequence. In the drawn graph, the abscissa is the video frame sequence, whereas the ordinate is the abnormal response value. The red areas represent the real label of the anomaly, and the black curves represent the response of the model proposed in this study. In Figure 3a, "retrograde" is identified as an abnormal event on the pedestrian street. When the video is played to approximately frame 83, a pedestrian is moving retrograde to the camera. This behavior is different from most behaviors in the normal frame, and the exception value increases accordingly. In Figure 3b, at about frame 25, "cyclist" appears in the scene. The model can effectively detect this as an anomaly.

We show the VAD effect on the CUHK Avenue dataset in Figure 4. The abnormal behaviors in Figure 4a,b are "the person throwing the schoolbag" and "the person pushing the bicycle".



(**a**) (**b**)

Figure 3. We show two abnormal events on the UCSD Ped2 dataset. The red border in the diagram represents abnormal behavior. The red area is a part of the real label of the anomaly, and the black curve represents the score of the anomaly. The higher the value, the more likely the exception will occur. (**a**) On a normal street from left to right, a pedestrian turned retrograde; (**b**) At frame 25 of the video, a man on a bicycle appeared in the street where pedestrians were walking.



(**a**) (**b**)

Figure 4. Two anomaly examples from the CUHK Avenue dataset. (**a**) A man in blue trousers scattered pieces of paper on the road. (**b**) In frames 500 to 780 of the video, a man walks out with a bicycle.

4.4. Ablation Experiment

To analyze the effects of different components of the model proposed in this study, we conducted ablation experiments on the UCSD Ped2 and CUHK Avenue datasets and evaluate it with the AUROC evaluation index. The results are shown in Table 2. Our model is divided into three components: optical flow reconstruction based on the memory module, intermediate frame prediction, and multi-discriminator optimization. The baseline we used is a simple prediction of future frames model. The ablation experiments were carried out by accumulating these three parts, respectively. In this experiment, the distillation method for optical flow distribution and overall distribution is hard distillation.

It can be seen from the table that the prediction method with only future frames is 94.1%. A 95.23% AUROC score was obtained by adding optical flow motion information with the memory module. Adding optical flow information can make the model more sensitive to abnormal activities. After using video context information, the prediction result of the erased frame is 96.07%; this shows that rich temporal information is crucial for improving model accuracy. The final result is 97.7% after adding multiple discriminators. We believe that this multi-discriminator pipeline is beneficial to balance the training of the network and improve the quality of the generated images. On the CUHK Avenue dataset, our method was 3.4% higher than the original baseline. All the methods are superior to the previous methods that can only predict future frames, which reflects the superiority of the method proposed in this study.

Table 2. We analyzed different components on the UCSD Ped2 and CUHK Avenue dataset. By superimposing the baseline, the contribution of each component to the model is evaluated.

Dataset	Baseline	Optical Flow	Erased Frame Prediction	Multi-Discriminator	AUROC (%)
Ped2	✓				94.1
	✓	✓			95.2
	✓	✓	✓		96.1
	✓	✓	✓	✓	97.7
CUHK Avenue	✓				86.3
	✓	✓			87.2
	✓	✓	✓		87.6
	✓	✓	✓	✓	89.7

We show the abnormal score images of different components on the Ped2 dataset in Figure 5. In Figure 5d, although the method of using only future frame prediction can achieve results, the score value of abnormal behavior is low in about 100–160 frames. In Figure 5c, by erasing frames, we utilize temporal context information to assist prediction. Then we use GAN's training method, which can obtain more response. Figure 5b shows the prediction of the optical flow reconstructed by the memory module. In Figure 5a, we combine all the information to test. Around frames 100–120, our model can detect obvious abnormalities, and the detection effect is smoother than others.

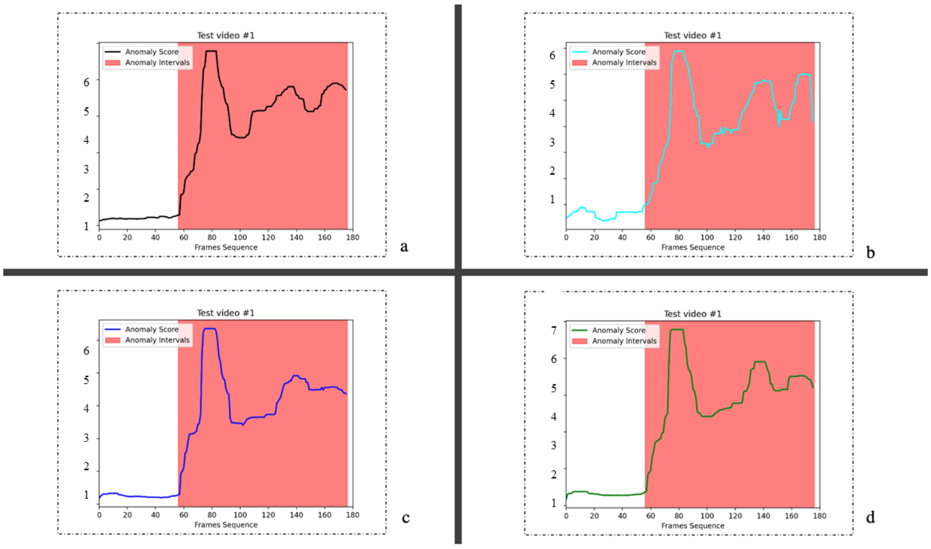


Figure 5. The different structures in the ablation experiment are analyzed on the Ped2 dataset, in which (**a**) is our overall model, (**b**) is the optical flow reconstruction model based on the memory module, (**c**) is the prediction of the erased frame with two discriminators, and (**d**) is the prediction of the original future frame.

Similarly, we also performed ablation experiments on the CUHK Avenue dataset, and the abnormal score image of each part is shown in Figure 6. When we combined optical flow information and temporal context information, the model was more sensitive to anomaly detection. When an abnormal event had just occurred, the response of the model was also stronger than others.

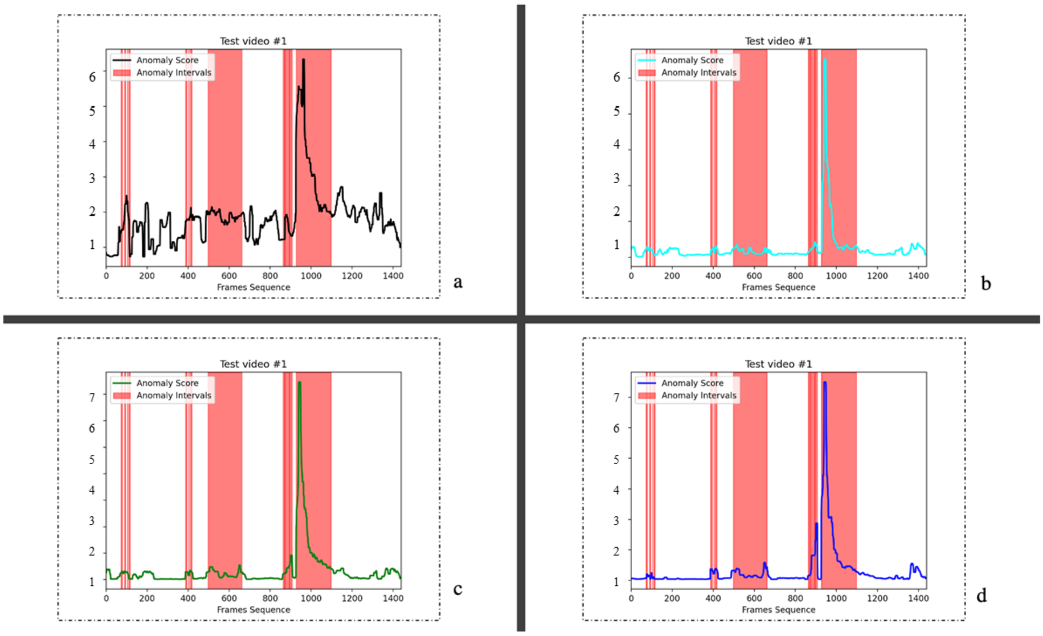


Figure 6. Test effects of different components of the model in the CUHK Avenue dataset, all legends are the same as shown in Figure 5.

In Figure 7, we visualize a heatmap of anomaly detection results on two datasets. The heatmap contains rich color changes and full information, which helps us add color representations of detected objects to the image. In the VAD task, we use the heatmap to indicate which areas are abnormally high-risk areas. In Figure 7a, we show that the anomalies are “cyclists” and “skateboarders”, and their trajectories can be clearly distinguished. In Figure 7b, the abnormal event we visualized is a “runner”. When an abnormal event occurs, its activity will be more intuitively represented.

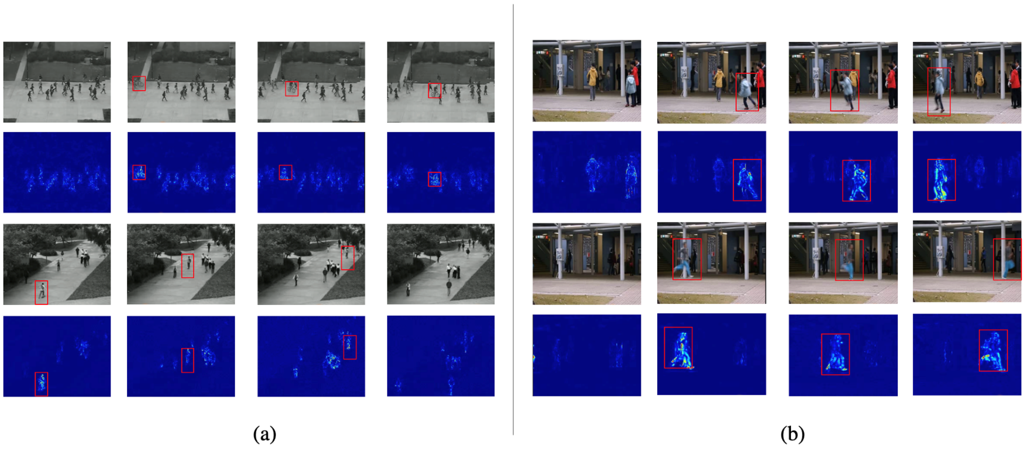


Figure 7. Visual example of frame prediction comparison In different datasets, we show normal and abnormal data. The first and third rows are the raw video frames in the dataset, and the second and fourth rows are the heatmaps of the frames. From left to right, we show (**a**) the Ped2 dataset, where the anomalies are “bicyclist” and “skateboarder” and (**b**) the CUHK Avenue dataset, where the anomalies are all “running”. Brighter colors in the heatmap indicate larger prediction errors.

5. Discussion

5.1. Top k in Memory Mechanism

We attempted to experiment with the memory mechanism of optical flow reconstruction, and the experimental results on the two datasets are shown in Table 3. In Ped2 dataset, through the sigmoid and weighted sum of the weights in all dimensions, the AUROC score was 95.6%. This method generates a lot of computational effort and affects the speed of model training. Next, considering the sparse addressing method, we set different top k values for experimental analysis. When K was 5, 10, 20, and 30, the final AUROC values were 95.4%, 95.9%, 97.7%, and 96.9%, respectively. When K was 20, our model's result was the best. In the CUHK Avenue dataset, an AUROC score of 89.3 % was obtained using the weighted average method. When K = 20, the AUROC value was still the highest, at 89.7%. The sparse addressing method enables the model to learn the representation of data with less but more relevant memory values. The latent space obtained by the encoder is used to retrieve the most relevant weight of the memory module to minimize the reconstruction error.

Table 3. In the module of memory reconstruction optical flow, we analyze the influence of different top k on the model results.

Dataset	**Top k**	**AUROC (%)**
Ped2	K = 5	95.4
	K = 10	95.9
	K = 20	97.7
	K = 30	96.9
	Weight mean	95.6
CUHK Avenue	K = 5	88.1
	K = 10	88.6
	K = 20	89.7
	K = 30	89.1
	Weight mean	89.3

5.2. Different Numbers of Temporal Context Information

In order to better complete incomplete events, we analyzed the number of contexts in the erased frames. The experiment was carried out on two datasets, and the results are shown in Table 4. Firstly, the video frames were preprocessed and the t_{th} frames were extracted as video sequences. In Ped2, when t = 7 and the number of front and back contexts was three frames, the model obtained a resulting AUROC score of 96.6%. When t = 5 and the number of front and back contexts was 2, the model obtained an AUROC score of 97.7%. When t = 3, the AUROC score was only 94.2%. In the CUHK Avenue dataset, the resulting AUROC score was only 81.0% when t = 3. When t = 5, the best result was 89.7%. The analysis results show that more context information may provide redundant information. Too little context cannot provide enough temporal information to effectively complete VAD tasks.

Table 4. We compared the effects of different numbers of before and after frames on intermediate frame prediction.

Dataset	**t = ?**	**AUROC (%)**
Ped2	t = 3	94.2
	t = 5	97.7
	t = 7	96.6
CUHK Avenue	t = 3	81.0
	t = 5	89.7
	t = 7	88.9

5.3. Different Distillation Methods

Knowledge distillation can compress the information learned by a teacher network into the student network. We transfer this learning inductive bias method to the optical flow model and take the normal video frame model as the teacher network. At the same time, we carry out experiments on the two distillation methods proposed in this study, and the results are shown in Table 5. In the Ped2 dataset, the final result obtained by the hard distillation method was 0.2% higher than that achieved by the soft distillation method. In the CUHK Avenue dataset, the performance is more obvious, and the model's effect was improved by 1.6%. Using the hard distillation method can make the student network more dependent on the hard labels provided by the teacher network. We think that this method can make the model learn the global information more effectively. The experiments showed that this hard distillation method can more effectively restore the original video of IE with the help of optical flow motion information.

Table 5. We carried out experiments on different distillation methods.

Dataset	Soft Distillation	Hard Distillation	AROUC (%)
Ped2	✓		97.5
		✓	97.7
CUHK Avenue	✓		88.1
		✓	89.7

5.4. Differing Numbers of Model Parameters

The number of different model parameters can directly affect the performance and efficiency of the model. On the Ped2 dataset, we compared the number of parameters of the model proposed in this paper with three other methods, and the results are shown in Table 6. The number of parameters of the proposed method reached 287 M, which is higher than in MemAE and HF2VAD. We believe that this is brought on by the computation of optical flow and the GAN training method, both of which create a large number of parameters to be computed. In our future work, we will reduce the number of model parameters as much as possible while maintaining accuracy and reducing the storage and computational costs of the model.

Table 6. Comparison of different method parameter quantities.

Model	Parameters (M)
MemAE [64]	22.8
HF2VAD [33]	252
FPVAD [26]	339
Ours	287

6. Conclusions

In this work, we combined the methods of reconstruction-based/frame prediction-based to obtain a high-precision VAD model. Our design idea is novel, predicting future frames by utilizing a variational auto-encoder as a generator while taking previous video frames and optical flow as inputs, and it proposes an efficient reconstruction method, i.e., an auto-encoder with memory structures. Instead of simply combining reconstruction and prediction, our architecture constrains the reconstructed optical flow and video frames by designing different distillation methods. At the same time, we make full use of the temporal information in the video to make predictions by erasing a patch in consecutive video frames. To generate high-quality predictions, we also add two discriminators to ensure that the generated predictions are consistent with the ground truths. This way, we can guarantee that events that are normal in terms of appearance and motion are identified as such, whereas events predicted to differ significantly from the ground truth are classified as anomalous. To verify the validity of this method, we performed experiments on three different datasets. The AUROC score achieved by our model on the UCSD Ped2 datasets was 97.9%, and those of the CUHK Avenue and ShanghaiTech datasets reached 89.7% and 75.8%, respectively. We performed ablation experiments with the first two datasets, achieving performance increases of 3.6% and 3.4%, respectively, when compared with the baseline. Our experiments show that this method can effectively manage VAD tasks. The results of this paper can play an important role in practical applications such as smart city surveillance and security fields and provide new ideas and methods to achieve more accurate and efficient target anomaly detection.

Author Contributions: Methodology, H.H. and B.Z.; Resources, F.G. and J.W.; Data curation, P.C.; Funding acquisition, A.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Engineering and Physical Sciences Research Council, grant number EP/M026981/1, EP/T021063/1, EP/T024917/.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jakob, P.; Madan, M.; Schmid-Schirling, T.; Valada, A. Multi-perspective anomaly detection. *Sensors* **2021**, *21*, 5311. [CrossRef] [PubMed]
2. Sun, L.; Wang, Z.; Zhang, Y.; Wang, G. A Feature-Trajectory-Smoothed High-Speed Model for Video Anomaly Detection. *Sensors* **2023**, *23*, 1612. [CrossRef]
3. De Paula, D.D.; Salvadeo, D.H.; de Araujo, D.M. CamNuvem: A robbery dataset for video anomaly detection. *Sensors* **2022**, *22*, 10016. [CrossRef]
4. Sidibe, D.; Sankar, S.; Lemaitre, G.; Rastgoo, M.; Massich, J.; Cheung, C.Y.; Tan, G.S.; Milea, D.; Lamoureux, E.; Wong, T.Y.; et al. An anomaly detection approach for the identification of DME patients using spectral domain optical coherence tomography images. *Comput. Methods Programs Biomed.* **2017**, *139*, 109–117. [CrossRef]
5. Chen, J.; Li, J.; Chen, W.; Wang, Y.; Jiang, T. Anomaly detection for wind turbines based on the reconstruction of condition parameters using stacked denoising autoencoders. *Renew. Energy* **2020**, *147*, 1469–1480. [CrossRef]
6. Abdallah, A.; Maarof, M.A.; Zainal, A. Fraud detection system: A survey. *J. Netw. Comput. Appl.* **2016**, *68*, 90–113. [CrossRef]
7. Lv, H.; Chen, C.; Cui, Z.; Xu, C.; Li, Y.; Yang, J. Learning Normal Dynamics in Videos with Meta Prototype Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 15425–15434.
8. Georgescu, M.I.; Barbalau, A.; Ionescu, R.T.; Khan, F.S.; Popescu, M.; Shah, M. Anomaly detection in video via self-supervised and multi-task learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 12742–12752.
9. Jin, Y.; Hong, J.; Han, D.; Ko, H. CPNet: Cross-Parallel Network for Efficient Anomaly Detection. In Proceedings of the 2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Washington, DC, USA, 16–19 November 2021; pp. 1–8.
10. Wan, B.; Jiang, W.; Fang, Y.; Luo, Z.; Ding, G. Anomaly detection in video sequences: A benchmark and computational model. *arXiv* **2021**, arXiv:2106.08570.
11. Zhang, M.; Li, T.; Shi, H.; Li, Y.; Hui, P. A decomposition approach for urban anomaly detection across spatiotemporal data. In Proceedings of the IJCAI International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019.
12. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
13. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578.
14. Yue, Z.; Gao, F.; Xiong, Q.; Wang, J.; Huang, T.; Yang, E.; Zhou, H. A novel semi-supervised convolutional neural network method for synthetic aperture radar image recognition. *Cogn. Comput.* **2021**, *13*, 795–806. [CrossRef]
15. Ma, F.; Sun, X.; Zhang, F.; Zhou, Y.; Li, H.C. What catch your attention in SAR images: Saliency detection based on Soft-Superpixel lacunarity cue. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 1–17. [CrossRef]
16. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]

17. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
18. Zhao, Y.X.; Zhang, Y.M.; Song, M.; Liu, C.L. Adaptable Global Network for Whole-Brain Segmentation with Symmetry Consistency Loss. *Cogn. Comput.* **2022**, *14*, 2246–2259. [CrossRef]
19. Wang, Y.N.; Tian, X.; Zhong, G. FFNet: Feature fusion network for few-shot semantic segmentation. *Cogn. Comput.* **2022**, *14*, 875–886. [CrossRef]
20. Chalapathy, R.; Borzeshi, E.Z.; Piccardi, M. An investigation of recurrent neural architectures for drug name recognition. *arXiv* **2016**, arXiv:1609.07585.
21. Wulsin, D.; Blanco, J.; Mani, R.; Litt, B. Semi-supervised anomaly detection for EEG waveforms using deep belief nets. In Proceedings of the 2010 Ninth International Conference on Machine Learning and Applications, Washington, DC, USA, 12–14 December 2010; pp. 436–441.
22. Song, H.; Jiang, Z.; Men, A.; Yang, B. A hybrid semi-supervised anomaly detection model for high-dimensional data. *Comput. Intell. Neurosci.* **2017**, *2017*, 8501683. [CrossRef] [PubMed]
23. Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A.K.; Davis, L.S. Learning temporal regularity in video sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 733–742.
24. Luo, W.; Liu, W.; Gao, S. Remembering history with convolutional lstm for anomaly detection. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 439–444.
25. Nguyen, T.N.; Meunier, J. Anomaly detection in video sequence with appearance-motion correspondence. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1273–1283.
26. Liu, W.; Luo, W.; Lian, D.; Gao, S. Future frame prediction for anomaly detection–a new baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6536–6545.
27. Mahadevan, V.; Li, W.; Bhalodia, V.; Vasconcelos, N. Anomaly detection in crowded scenes. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1975–1981.
28. Chang, X.; Ren, P.; Xu, P.; Li, Z.; Chen, X.; Hauptmann, A. A comprehensive survey of scene graphs: Generation and application. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *45*, 1–26. [CrossRef]
29. Zhang, L.; Chang, X.; Liu, J.; Luo, M.; Li, Z.; Yao, L.; Hauptmann, A. TN-ZSTAD: Transferable Network for Zero-Shot Temporal Activity Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 3848–3861. [CrossRef]
30. Li, M.; Huang, P.Y.; Chang, X.; Hu, J.; Yang, Y.; Hauptmann, A. Video Pivoting Unsupervised Multi-Modal Machine Translation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 3918–3932. [CrossRef]
31. Saligrama, V.; Chen, Z. Video anomaly detection based on local statistical aggregates. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2112–2119.
32. Zhang, Y.; Lu, H.; Zhang, L.; Ruan, X.; Sakai, S. Video anomaly detection based on locality sensitive hashing filters. *Pattern Recognit.* **2016**, *59*, 302–311. [CrossRef]
33. Liu, Z.; Nie, Y.; Long, C.; Zhang, Q.; Li, G. A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 13588–13597.
34. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
35. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. *arXiv* **2021**, arXiv:2111.06377.
36. Lu, C.; Shi, J.; Jia, J. Abnormal event detection at 150 fps in matlab. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2720–2727.
37. Wu, S.; Moore, B.E.; Shah, M. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2054–2060.
38. Tung, F.; Zelek, J.S.; Clausi, D.A. Goal-based trajectory analysis for unusual behaviour detection in intelligent surveillance. *Image Vis. Comput.* **2011**, *29*, 230–240. [CrossRef]
39. Erfani, S.M.; Rajasegarar, S.; Karunasekera, S.; Leckie, C. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognit.* **2016**, *58*, 121–134. [CrossRef]
40. Piciarelli, C.; Foresti, G.L. On-line trajectory clustering for anomalous events detection. *Pattern Recognit. Lett.* **2006**, *27*, 1835–1842. [CrossRef]
41. Piciarelli, C.; Micheloni, C.; Foresti, G.L. Trajectory-based anomalous event detection. *IEEE Trans. Circuits Syst. Video Technol.* **2008**, *18*, 1544–1554. [CrossRef]
42. Morris, B.T.; Trivedi, M.M. Trajectory Learning for Activity Understanding: Unsupervised, Multilevel, and Long-Term Adaptive Approach. *IEEE Trans. Softw. Eng.* **2011**, *33*, 2287–2301. [CrossRef] [PubMed]

43. Xu, D.; Ricci, E.; Yan, Y.; Song, J.; Sebe, N. Learning deep representations of appearance and motion for anomalous event detection. *arXiv* **2015**, arXiv:1510.01553.
44. Hinami, R.; Mei, T.; Satoh, S. Joint detection and recounting of abnormal events by learning deep generic knowledge. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3619–3627.
45. Luo, W.; Liu, W.; Gao, S. A revisit of sparse coding based anomaly detection in stacked rnn framework. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 341–349.
46. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks: An overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [CrossRef]
47. Yan, C.; Chang, X.; Li, Z.; Guan, W.; Ge, Z.; Zhu, L.; Zheng, Q. Zeronas: Differentiable generative adversarial networks search for zero-shot learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 9733–9740. [CrossRef] [PubMed]
48. Salehi, M.; Sadjadi, N.; Baselizadeh, S.; Rohban, M.H.; Rabiee, H.R. Multiresolution knowledge distillation for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14902–14912.
49. Li, C.L.; Sohn, K.; Yoon, J.; Pfister, T. CutPaste: Self-Supervised Learning for Anomaly Detection and Localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9664–9674.
50. Cui, Z.; Qin, Y.; Zhong, Y.; Cao, Z.; Yang, H. Target Detection in High-Resolution SAR Image via Iterating Outliers and Recursing Saliency Depth. *Remote Sens.* **2021**, *13*, 4315. [CrossRef]
51. Pang, G.; Yan, C.; Shen, C.; Hengel, A.v.d.; Bai, X. Self-trained deep ordinal regression for end-to-end video anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 12173–12182.
52. Saypadith, S.; Onoye, T. An Approach to Detect Anomaly in Video Using Deep Generative Network. *IEEE Access* **2021**, *9*, 150903–150910. [CrossRef]
53. Chong, Y.S.; Tay, Y.H. Abnormal event detection in videos using spatiotemporal autoencoder. In Proceedings of the International Symposium on Neural Networks, Sapporo, Japan, 21–26 June 2017; pp. 189–196.
54. Feng, X.; Song, D.; Chen, Y.; Chen, Z.; Ni, J.; Chen, H. Convolutional Transformer based Dual Discriminator Generative Adversarial Networks for Video Anomaly Detection. In Proceedings of the 29th ACM International Conference on Multimedia, Online, 20–24 October 2021; pp. 5546–5554.
55. Veluchamy, S.; Karlmarx, L.; Mahesh, K.M. Detection and Localization of Abnormalities in Surveillance Video Using Timerider-Based Neural Network. *Comput. J.* **2021**, *64*, 1886–1906. [CrossRef]
56. Lee, S.; Kim, H.G.; Ro, Y.M. BMAN: Bidirectional Multi-Scale Aggregation Networks for Abnormal Event Detection. *IEEE Trans. Image Process.* **2019**, *29*, 2395–2408.
57. Leyva, R.; Sanchez, V.; Li, C.T. Video anomaly detection with compact feature sets for online performance. *IEEE Trans. Image Process.* **2017**, *26*, 3463–3478. [CrossRef] [PubMed]
58. Athanesious, J.; Srinivasan, V.; Vijayakumar, V.; Christobel, S.; Sethuraman, S.C. Detecting abnormal events in traffic video surveillance using superorientation optical flow feature. *IET Image Process.* **2020**, *14*, 1881–1891. [CrossRef]
59. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
60. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
61. Mathieu, M.; Couprie, C.; LeCun, Y. Deep multi-scale video prediction beyond mean square error. *arXiv* **2015**, arXiv:1511.05440.
62. Park, H.; Noh, J.; Ham, B. Learning memory-guided normality for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14372–14381.
63. Fan, Y.; Wen, G.; Li, D.; Qiu, S.; Levine, M.D.; Xiao, F. Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder. *Comput. Vis. Image Underst.* **2020**, *195*, 102920. [CrossRef]
64. Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M.R.; Venkatesh, S.; Hengel, A.v.d. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1705–1714.
65. Liu, Y.; Li, C.L.; Póczos, B. Classifier Two Sample Test for Video Anomaly Detections. In Proceedings of the British Machine Vision Conference, Newcastle, UK, 3–6 September 2018; p. 71.
66. Yu, G.; Wang, S.; Cai, Z.; Zhu, E.; Xu, C.; Yin, J.; Kloft, M. Cloze test helps: Effective video anomaly detection via learning to complete video events. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 583–591.
67. Lu, Y.; Kumar, K.M.; Shahabeddin Nabavi, S.; Wang, Y. Future frame prediction using convolutional vrnn for anomaly detection. In Proceedings of the 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Taipei, Taiwan, 18–21 September 2019; pp. 1–8.
68. Zhou, J.T.; Zhang, L.; Fang, Z.; Du, J.; Peng, X.; Xiao, Y. Attention-driven loss for anomaly detection in video surveillance. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 4639–4647. [CrossRef]
69. Dong, F.; Zhang, Y.; Nie, X. Dual discriminator generative adversarial network for video anomaly detection. *IEEE Access* **2020**, *8*, 88170–88176. [CrossRef]

70. Sun, C.; Jia, Y.; Hu, Y.; Wu, Y. Scene-aware context reasoning for unsupervised abnormal event detection in videos. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 184–192.
71. Doshi, K.; Yilmaz, Y. Any-shot sequential anomaly detection in surveillance videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 934–935.
72. Morais, R.; Le, V.; Tran, T.; Saha, B.; Mansour, M.; Venkatesh, S. Learning regularity in skeleton trajectories for anomaly detection in videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 11996–12004.
73. Zhao, Y.; Deng, B.; Shen, C.; Liu, Y.; Lu, H.; Hua, X.S. Spatio-temporal autoencoder for video anomaly detection. In Proceedings of the 25th ACM international conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1933–1941.
74. Ye, M.; Peng, X.; Gan, W.; Wu, W.; Qiao, Y. Anopcn: Video anomaly detection via deep predictive coding network. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1805–1813.
75. Tang, Y.; Zhao, L.; Zhang, S.; Gong, C.; Li, G.; Yang, J. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognit. Lett.* **2020**, *129*, 123–130. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.