symmetry

MDPI

*Article*

# Video Anomaly Detection Based on Attention Mechanism

**Qianqian Zhang [1,†], Hongyang Wei [1,†], Jiaying Chen [2], Xusheng Du [2] and Jiong Yu [2,*]**

[1] School of Software Engineering, Xinjiang University, Urumqi 830091, China
[2] School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China
* Correspondence: yujiong@xju.edu.cn
† These authors contributed equally to this work.

**Abstract:** Camera surveillance is widely used in residential areas, highways, schools and other public places. The monitoring and scanning of sudden abnormal events depend on humans. Human anomaly monitoring not only consumes a lot of manpower and time but also has a large error in anomaly detection. Video anomaly detection based on AE (Auto-Encoder) is currently the dominant research approach. The model has a highly symmetrical network structure in the encoding and decoding stages. The model is trained by learning standard video sequences, and the anomalous events are later determined in terms of reconstruction error and prediction error. However, in the case of limited computing power, the complex model will greatly reduce the detection efficiency, and unnecessary background information will seriously affect the detection accuracy of the model. This paper uses the AE loaded with dynamic prototype units as the basic model. We introduce an attention mechanism to improve the feature representation ability of the model. Deep separable convolution operation can effectively reduce the number of model parameters and complexity. Finally, we conducted experiments on three publicly available datasets of real scenarios (UCSD Ped1, UCSD Ped2 and CUHK Avenue). The experimental results show that compared with the baseline model, the accuracy of our model improved by 1.9%, 1.4% and 6.6%, respectively, across the three datasets. Compared with many popular models, the validity of our model in anomaly detection is verified.

**Keywords:** anomaly detection; attention mechanism; depthwise separable convolution; deep learning; symmetrical structure



**Citation:** Zhang, Q.; Wei, H.; Chen, J.; Du, X.; Yu, J. Video Anomaly Detection Based on Attention Mechanism. *Symmetry* **2023**, *15*, 528. https://doi.org/10.3390/sym15020528

Received: 17 January 2023
Revised: 10 February 2023
Accepted: 13 February 2023
Published: 16 February 2023



## 1. Introduction

Video anomaly detection [1] is unlike traditional anomaly detection [2–4], video anomaly detection refers to the detection of behavior or appearance inconsistent with expectations in a normal surveillance video or normal behavior or appearance in abnormal time or space. At present, there are many solutions, which can be roughly divided into reconstruction-based, prediction-based, classification-based and regression-based methods. The method based on reconstruction and prediction is more mainstream, so it is introduced as a key point. The reconstruction-based approach [5–9] involves constructing a model from normal video sequences and then reconstructing the test dataset. The abnormal image does not conform to the coding of the model. Researchers use the size of reconstruction error to identify the abnormal image. Depth AE is the most common detection method based on the reconstruction method. Researchers usually use an AE to learn the normal video sequence mode to reconstruct the current frame. Hasan et al. [10] proposed two detection methods based on AE, First, the feature extraction of the video sequence is carried out according to the traditional manual method, and a fully connected AE is added. Secondly, the convolutional neural network is introduced to construct a feedforward encoder to learn the local features of the video sequence. However, this method relies too much on normal data, so the sensitivity of the model to abnormal images is poor. To solve the above problems, Park et al. [11] proposed the use of a neural network model with updated memory to remember normal images. At the same time, the author introduces new features

of compactness loss and separation loss to improve the training memory of the model. Gong et al. [12] built a memory enhancement encoder, MemAE, to update the memory content at the training stage. In the test phase, it is reconstructed from the storage memory of normal data. The model has a high generalization ability. Ref. [13] couples the target motion characteristics of pre-training with the anomaly scoring function based on the cosine distance. Based on the reconstruction strategy, the author introduces additional constraints to extend the previous method. The model has achieved good results in the abnormal target locations. A temporal coherent sparse coding (TSC) is proposed in [14]. The model forces the use of similar reconstruction parameters to encode similar neighbor video sequences. Moreover, the stack recurrent neural network with a special type is used to map TSC, which shows excellent results on real datasets. Luo et al. [15] used convolutional neural networks to encode the appearance of video sequences. The convolution long and short time memory is used to detect the anomaly corresponding to the motion information, and the validity of the model is verified on a real dataset. However, when the reconstruction-based anomaly detection method encounters new detection video sequences, it needs to carry out new training to ensure that the appropriate model is obtained. Moreover, the spatial feature extraction of the model for video sequences is poor, resulting in low detection accuracy.

Prediction-based methods [16–20] of default video sequences have certain context information links. The model predicts the next frame by learning the original dataset, and the prediction error of abnormal video sequences is large. In [21], a new convolutional variational recurrent neural network (VRNN) is constructed to predict video sequences by combining the variational AE with Conv-LSTM. To solve the problem of incomplete feature extraction of a video sequence by the encoder. Wang [22] proposed a future frame prediction method based on a generation antagonism network (GAN) and attention mechanism. For the generation model, the author added the attention module in the U-Net network. For the identification model, the Markov GAN model with the attention mechanism improves the detection performance of the algorithm. Similarly, Liu [23] proposes a hybrid framework HF2-VAD that integrates stream reconstruction and frame prediction. First, the ML MemAE SC network is used to store the normal mode of optical flow reconstruction to identify abnormal frames with large flow reconstruction errors. The conditional variational automatic encoder is used to capture the correlation between the video sequence and optical flow to predict the next frame; The reconstructed optical flow of the abnormal frame further affects the quality of the predicted frame, thus identifying the abnormal frame. However, many normal events are unpredictable, and if the abnormal events occupy a small area of the picture, the prediction effect is also poor. Therefore, the false alarm rate of anomaly detection relying solely on prediction is very high.

Other anomaly detection methods include the classification-based method [24] and regression-based method [25]. Mehran et al. [26] proposed the use of the social force model to detect and locate crowd videos. Place a particle mesh on the image and compare it with the time-space average of the optical flow. Consider moving particles as individuals, and use the social force model to evaluate their interaction forces. Then the force is mapped to the image plane to obtain the force flow of each pixel, and the space-time volume of the force flow is used to simulate the normal behavior of the crowd to identify the abnormal frame. Ref. [27] trains a binary classifier based on unmaking to distinguish continuous video sequences; however, it eliminates too many distinguishing features at each step to distinguish abnormal events. Mahadevan [28] et al. proposed a new anomaly detection framework for congestion scenarios. This method solves the limitations of potential visual representation. The paper proposes that feature representation should jointly model the dynamics and appearance of crowd patterns, and have the ability of spatiotemporal features to ensure the integrity of feature extraction. Lu et al. [29] proposed an adaptive anomaly detection algorithm for scenes with few shots. Based on meta-learning, a few-shot learner is constructed to solve the problem of anomaly detection in multiple scenes. Cai et al. [30] proposed an appearance-motion memory consistency network (AMMC-Net) using the prior knowledge of the appearance signal and motion signal to capture their corresponding

relationship in the high-level feature space. Then, they combined this with multi-view features to increase the difference between normal and abnormal video sequences.

Among the research methods based on prediction and reconstruction, researchers are keen to use the AE (Auto-Encoder) as the detection model and then reconstruct or predict the test data by modeling the normal data. It is used to identify abnormal events based on high error. However, the AE needs to encode the normal video frame during the training process, which consumes a lot of memory, and it cannot process new scenarios in test data. Therefore, Lv [31] develops a dynamic prototype unit (DPU) as a learning model of normal features. The dynamic encoding of normal video frames is embedded into the encoder as a prototype. In addition, meta-learning is introduced to create multi-scenario learners. However, there are two main problems: in the case of limited computing power, complex models will affect the detection efficiency. At the same time, computing unimportant video background information will occupy a large amount of video memory.

To solve the above problems, this paper introduces an attention mechanism based on DPU. The circular attention module is used to collect the spatial information of the video sequence to improve the feature representation ability of the model. The depth separable convolution operation is introduced to reduce the number of parameters and enhance the model's accuracy. The contributions of this paper are summarized as follows:

- An attention module is designed and introduced into the AE. The model obtains the context information of video frames from horizontal and vertical directions to reduce the interference of video background information. To improve the feature learning ability of the neural network.
- We introduce a deep separable convolution operation into the decoder. It reduces the parameters of the model and improves detection efficiency and accuracy.
- Compared with multiple video anomaly detection models in different periods, the experimental results verify our model's powerful learning ability in complex scenes.

The structure of this paper includes the introduction, which mainly introduces the background of the paper and related research work. Next is the proposed method, focusing on how our model works. Next is the experiment, which introduces the preparation and implementation details of the experiment. Finally, we will discuss how the model works and summarize its effectiveness on the model.

## 2. Proposed Method

We propose a video anomaly detection model based on the attention mechanism, and the general framework is shown in Figure 1. Our approach consists of the following four components: (1) Encoding stage: the video sequence $(I_k - T + 1, I_k - T + 2..., I_k)$ is fed into the encoder to obtain the feature encoding of the implicit layer. (2) Prototype pool: The coding diagram in the hidden layer generates the prototype, $P_m$, through the mapping function and constructs the prototype pool. (3) Attention mechanism: Inputs the coding features obtained from the hidden layer into the attention unit. Obtains the spatial information in the feature map to form a new coding feature. (4) Decoding stage: Depth-wise (DW), a depth-separable convolution operation, is introduced to optimize the anomaly detection model.

In particular, our model has a high degree of symmetry in the encoding and decoding phases, with an overall implementation of an end-to-end connectivity architecture.

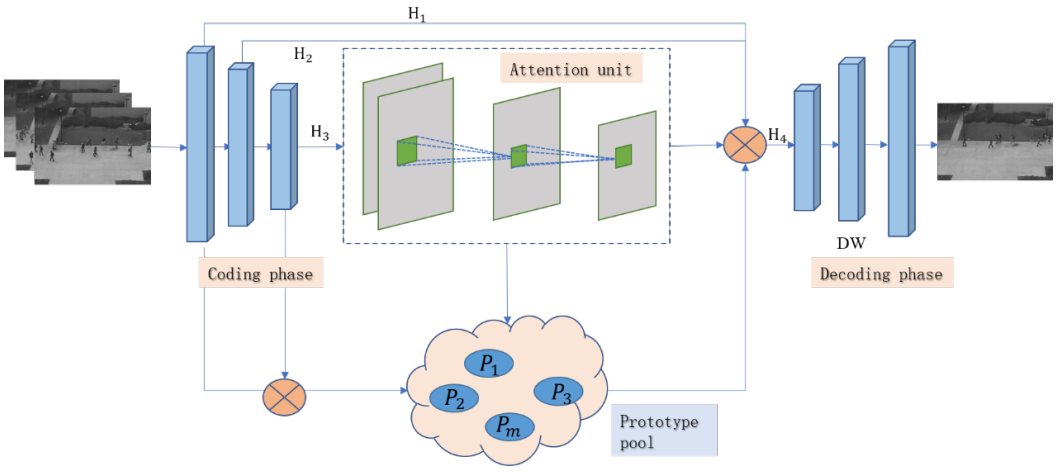


**Figure 1.** Anomaly detection model based on attention mechanism.

### 2.1. Attention Prototype Pool

We take DPU as the baseline model and add a recursive attention unit to learn normal features. The overall framework is shown in Figure 1. Assuming a set of training data frames $V = (I_k - T + 1, I_k - T + 2..., I_k)$ simplified to $V$, we feed $V$ into the AE. This outputs the result through encoding–attention model–decoding operation to predict the next frame, $Y_k = Y_k + 1$. We represent the sequence of frames at moment k as the input-output pair $(X_k, Y_k)$. The forward propagation mode of the model is: first, generate a feature prototype pool. After that, the model is encoded normally by retrieving prototype features to obtain hidden layer features. Finally, the input code is aggregated with the normal code as the output.

First, the k-th coding graph extracted from AE, $X_k = f_\theta(x_k)$, is regarded as the C-dimension vector $\{x_k^1, x_k^2 ... x_k^m\}$ with $N = w \times h$. The attention mapping function encodes and assigns weights to each pixel position of the normal image. The weight normalization calculation method is shown in Equation (1). After generating the prototype pool, use the input vector of the AE code graph to search the prototype in the prototype pool. The new coding diagram is obtained by vector summation Equation (2). Finally, the covariance matrix is introduced. The independence of the prototype in the prototype pool based on the independence of the input code is ensured, and then subsequent predictions are made through the encoder to obtain accurate judgment results.

Where $P_k^m$ means that the $k$-th coding diagram is generated by the $m$-th mapping function. The prototype, $P$, is considered a set of $N$ coding vectors, where $N$ represents the number of samples. $\omega_k^{n,m}$ represents the weight value assigned by the $k$-th coding vector. $x_k^n$ represents the $k$-th vector extracted from the AE. $\theta_k^{n,m}$ represents the correlation score between the $n$-th coding vector and the $m$-th prototype in the prototype pool, and the sum obtained is taken as the output vector.

$$P_k^m = \sum_{n=1}^{N} \omega_k^{n,m} * X_k^n / \sum_{n'=1}^{N} \omega_k^{n',m} \tag{1}$$

$$X_k^n = \widehat{\sum_{m=1}^{M}} \theta_k^{n,m} p_k^m \tag{2}$$

### 2.2. Attention Unit

We introduced two repeated attention modules to collect the horizontal and vertical contextual information of video sequences to enhance the feature extraction ability of the model.

We obtain two feature maps, $Q$ and $K$, of the video sequence using two 1*1 convolution kernels. At each pixel position $S_1$ of the feature map $Q$, the feature vector $q$ is obtained. Meanwhile, the feature vector $K$ is obtained from the pixel position $S_2$ in the same row or column as $S$ in the feature map $k$. We calculate the weight value of each vector according to the affinity operation. According to the weight coefficient, the value of the feature map is weighted and summed to obtain feature map $A$ with attention. To prevent a pixel from not being closely related to its surrounding pixels. We set up two attention units to form a circular attention structure. For the feature maps, $A$ and $A^{'}$, obtained by two attention units, the weight mapping function is expressed as $A^{'} = f(A, S_1, S_2)$ .

That is, given the input video sequence X, the hidden layer characteristic graph $H$ is obtained through the AE. Then, we send the feature map $H$ into the first attention unit to generate a new feature map $H^{'}$ with attention. $H$ and $H^{'}$ have the same size. $H^{'}$ has information about the vertical and horizontal directions of each pixel in the video sequence. To ensure that the information obtained is more abundant, we put $H^{'}$ into the attention unit again and get the feature map $H^{''}$ with dense context information. To avoid extra parameter calculation, the front and back attention units share the same parameter. The loop setting of the attention unit ensures that there is an information connection between any position, $S_1$, on the feature map $H^{'}$ and any position, $S_2$, on the feature map $H^{''}$ ($S_1$ and $S_2$ are in the same direction). Therefore, we can ensure that our structure can create an informative feature map. Finally, we fuse the feature map $H''$ with contextual with the initial input video sequence $X$. The attention mechanism helps the model better focus on the video subject. We ignore the impact of the video background on the encoder to reduce the storage pressure of the model.

### *2.3. Deeply Separable Convolution*

We replace the standard convolution in the decoder with deep convolution (Depth-wise (DW)) and Pointwise convolution (PW) to reduce the model parameters. The decoder has a symmetric network structure to achieve internal end-to-end connectivity. The Depth of Separable Convolution (Depth-wise Separable Convolution, DSC) architecture, as shown in Figure 2, convolution layers after batch normalization and the Mish activation function are used. The specific convolution layer is shown in Figure 3. In deep convolution, a $3 \times 3$ convolution kernel is first used for convolution on the input channel, and the obtained feature layer $W \times H$ is the same as the target feature layer. However, the number of channels does not reach the output channel target. The model needs to use four groups of $1 \times 1$ convolution cores to reach the target channel through point-by-point convolution. In the standard convolution, assuming that the number of input channels is m and the number of output channels is n, the size of standard parameters is $m \times 3 \times 3 \times n$, while the depth convolutional parameter is $m \times 3 \times 3$, and the point-by-point convolutional parameter is $m \times 1 \times 1 \times n$. The total parameters used are $m \times 3 \times 3 + m \times 1 \times 1 \times n$. The number of parameters decreases exponentially, and the depth convolution model also contributes to the improvement in our accuracy. The Mish activation function is defined in Equation (3).

$$\text{Mish} = x * \tanh(\ln(1 + e^x)) \tag{3}$$

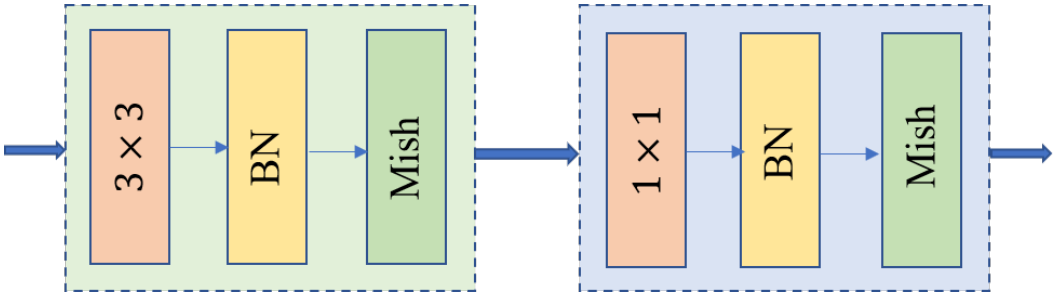


**Figure 2.** Depth-separable convolutional module architecture diagram.

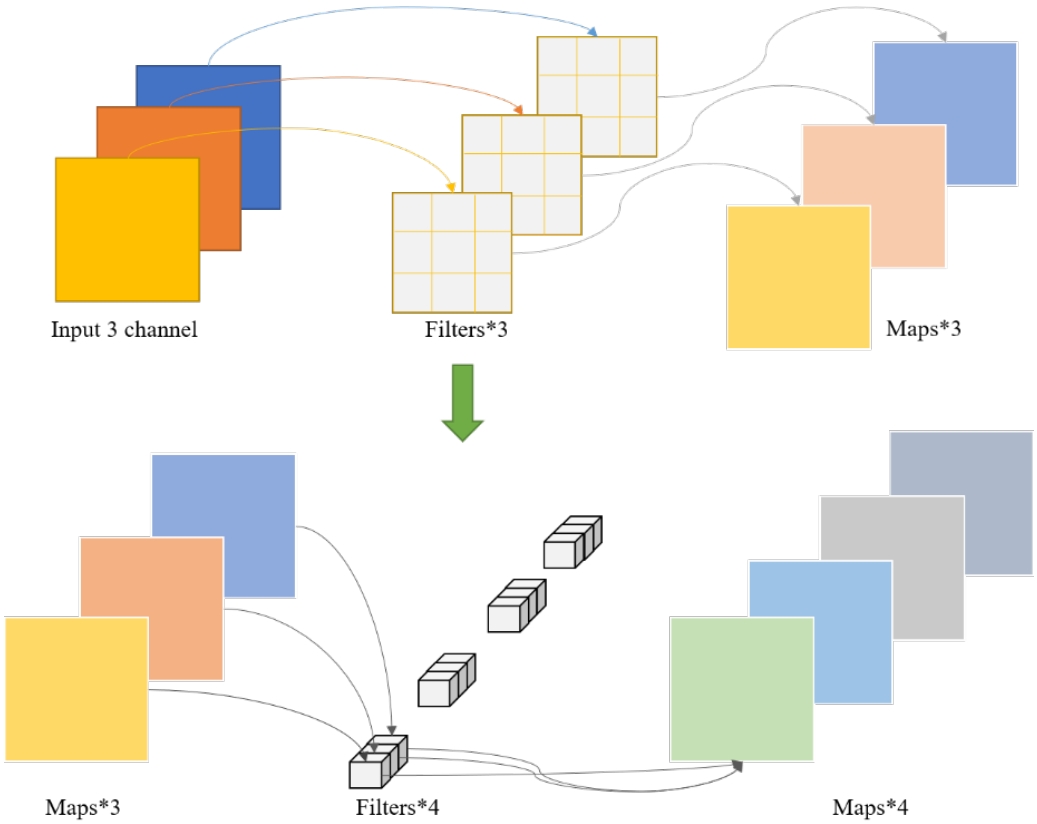


**Figure 3.** Deeply separable convolutional structure map.

### *2.4. Abnormal Score*

We use the feature reconstruction error and frame prediction error as the primary basis for anomaly scoring. The reconstruction error is first obtained by calculating the compactness error of the feature reconstruction term, and then the frame prediction error, as described in [32], can also be used as a description of the anomaly score. We use the two errors as the final anomaly score of the model by weighting the sum.

The total loss function of our model is obtained by weighting the frame prediction loss, feature reconstruction loss, and feature distance loss, as shown in Equation (4). Where the frame prediction loss is obtained by calculating the L2 distance between the actual image $y_t$ and the predicted value $y_t{}'$, marked as $L_{fra}$, as shown in Equation (5). The feature reconstruction loss is calculated from the reconstruction error obtained from the AE, marked as $L_{fea}$, as shown in Equation (6). The feature distance loss is calculated from the error between the prototype code $P_m$ and the predicted code $P_m{}'$, marked as $L_{dis}$, as shown in Equation (7).

$$L_{total} = \omega_1 L_{fra} + \omega_2 L_{fea} + \omega_3 L_{dis} \tag{4}$$

$$L_{fra} = \left\| y_t' - y_t \right\|_2 \tag{5}$$

$$L_{\text{fea}} = \left\| y_t'' - y_t \right\|_2 \tag{6}$$

$$L_{\text{dis}} = \frac{2}{M(M-1)} \sum_{m=1}^{M} \sum_{m'}^{M} (-\| P_m - P_{m'} \|_2) \tag{7}$$

## 3. Experiment

### *3.1. Experimental Implementation*

Experiments are conducted on three datasets, and the size of the input video image is 256 * 256 * 3. The circular attention mechanism is inserted behind the third layer of the AE. The parameters of the loss function are $\omega_1 = 1$, $\omega_2 = 1$, $\omega_3 = 0.001$. After 3000 training sessions for each training dataset, the loss value of the model tends to be stable, as shown in Figure 4. $Batchsize = 2$, the learning rate is $lr = 0.0001$, and the optimizer is Adam. This experiment was implemented under the deep learning framework PyTorch 1.6.0 on a 64-bit Ubuntu 18.04 LTS system with an NVIDIA GeForce RTX 2080 Ti and 11 GB of memory.

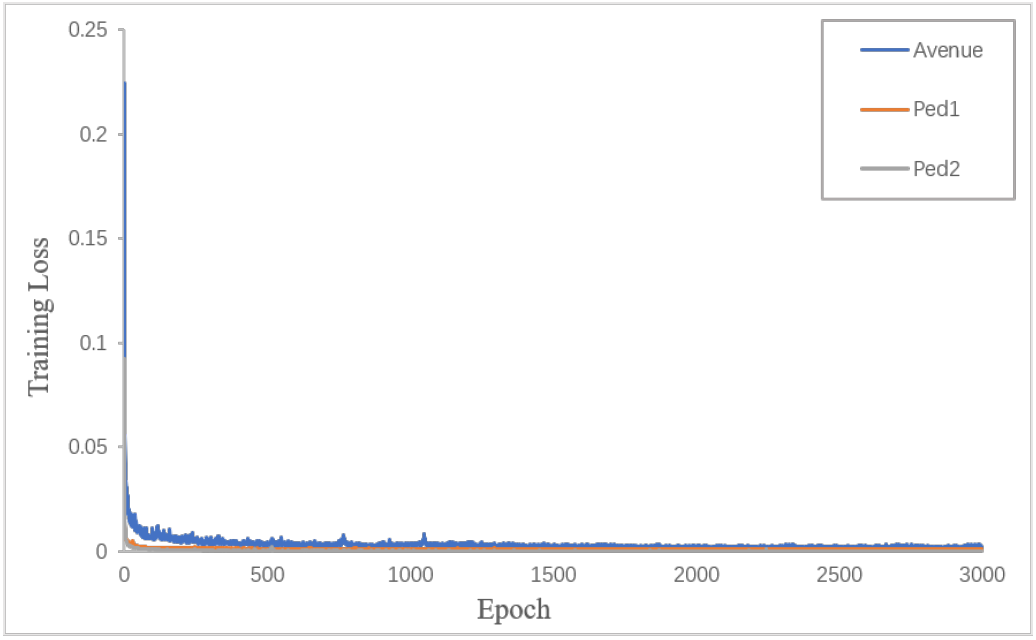


**Figure 4.** Change in training loss value with training times.

### *3.2. Datasets and Evaluation Indicators*

The model requires a large amount of training data to learn the deeper features of the video sequence. The experimental data in this paper selects three public video anomaly detection datasets, namely UCSD Ped1, Ped2 and CUHK Avenue. The UCSD Ped1 (denoted by Ped1 in the following text) includes 6800 video frames for 34 training data, 7200 video frames for 16 test data and 40 abnormal events. The UCSD Ped2 (denoted by Ped2 in the following text) includes 16 training data with a total of 1920 video frames, 21 test data with 2160 video frames and 12 abnormal events. The CUHK Avenue (denoted by Avenue in the following text) dataset consists of 16 training data, including 15,328 video frames, 21 test data of 15,324 video frames and 47 abnormal events.

The dataset attributes are shown in Table 1, and the exception event case illustration is shown in Figure 5. The Ped1 and Ped2 datasets are video scenes of school pavements, and scenes with only pedestrians in the video are considered normal data. In contrast, actions such as trucks, bicycles, wheelchairs, trampling on the lawn, and running are considered abnormal events. The Avenue dataset scene is a Central Avenue scene captured by a surveillance camera, where phenomena such as unusual movements of pedestrians, wrong direction of action and the presence of unique objects are considered extraordinary events.

**Table 1.** Dataset.

| Dataset | Number of Samples | Training Set | Testing Set | Abnormal Events |
|---|---|---|---|---|
| Ped1 | 14,000 | 6800 | 7200 | 40 |
| Ped2 | 4080 | 1920 | 2160 | 12 |
| Avenue | 30,652 | 15,328 | 15,324 | 47 |

(a) Ped1: Riding (b) Ped2: Riding (c) Avenue: Running

**Figure 5.** Dataset exception case illustration.

In video anomaly detection, we selected the Area Under the receiver operating characteristic Curve (AUC) as the final evaluation metric of the algorithm. The receiver operating characteristic curve (ROC) is a curve. Its abscissa and ordinate are false positive rate (FPR) and true positive rate (TPR). As the curves do not allow for a comparison of good and bad models, the AUC values below are used as specific evaluation indicators. According to the prediction result confusion matrix (Table 2), the calculation process is shown in Formulas (8)–(12). In general, the closer the ROC curve is to the upper left (the more significant the actual case rate), i.e., the closer the AUC score is to 1, the better the detection performance of the algorithm is demonstrated.

**Table 2.** Prediction result confusion matrix.

| Real Label | Forecast Label | |
|---|---|---|
| | Positive | Negative |
| Positive | TP | FN |
| Negative | FP | TN |

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

$$TPR = \frac{TP}{TP + FN} \tag{9}$$

$$x_{i+1} = \begin{cases} FPR & , i = 0 \\ x_i + \frac{1}{m^-} & , i > 0 \text{ and } x_i \in FP \\ x_i & , i > 0 \text{ and } x_i \in TP \end{cases} \tag{10}$$

$$y_{i+1} = \begin{cases} TPR & , i = 0 \\ y_i + \frac{1}{m^+} & , i > 0 \text{ and } y_i \in TP \\ y_i & , i > 0 \text{ and } y_i \in FP \end{cases} \tag{11}$$

$$AUC = \frac{\sum_{i=1}^{m-1}(x_{i+1} - x_i) \times (y_{i+1} - y_i)}{2} \tag{12}$$

where $x_i$ and $y_i$ are the abscissa and ordinate of the $i$-th sample point, and $i$ is a positive integer. $m$ is the total number of samples, $m^+$ is the number of positive samples and $m^-$ is the number of negative samples.

### 3.3. Parameter Sensitivity Analysis

We analyzed the influence of feature differentiation loss weight on model accuracy through experiments.

In our model, the loss function is derived from a weighted sum of three components: frame prediction loss, feature reconstruction loss and feature distance loss, with the weights of the first two losses assigned traditionally. In terms of feature distance loss, the Euclidean distance is used to calculate the feature distance. The calculation of this part has a tremendous fluctuation influence on abnormal detection results, so we set 0.0001, 0.001, 0.01, 0.10 and 1.00, respectively, for $\omega_3$ to carry out comparative experiments, and the experimental results are shown in Figure 6. The experiments show that our model has the highest accuracy when $\omega_3$= 0.001.

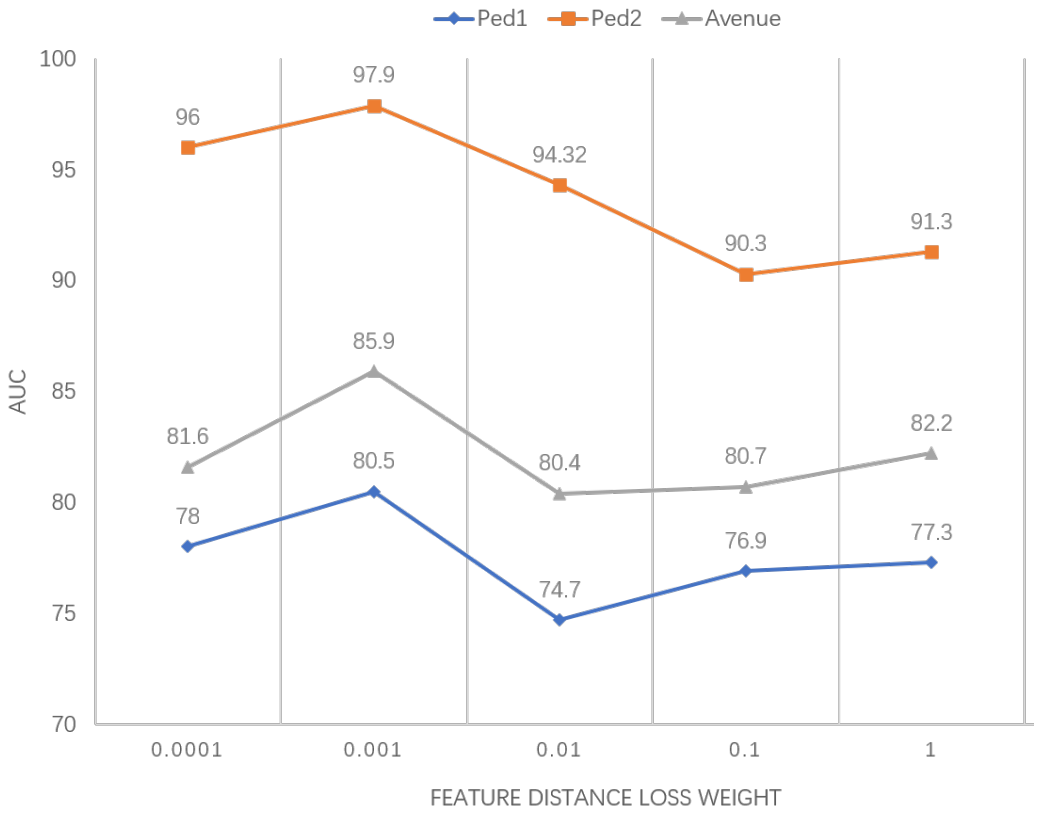


**Figure 6.** Sensitivity analysis of feature distance loss weights.

### 3.4. Time Complexity Analysis

The algorithm's primary time consumption is the attention mechanism's computational phase. In the attention unit module, the output of each encoding stage corresponds to an intermediate state $c_i$ of the video frame, which is used to record the relationship between the previous output result $s_{i-1}$ and the state of the implicit layer. To calculate the state $c_i$, we need to calculate the n weights obtained by attentional encoding each time. $n$ denotes the number of images that go to the encoding state, and the decoder has a total of n states, so the time complexity is O(n*n).

### 3.5. Results and Analyses

We have compared popular algorithms in different periods. The specific detection accuracy is shown in Table 3. The data in the table is obtained from the original paper. Since our model performs best on dataset Ped2; we have carried out experimental verification on it using abnormal scores. As shown in Figure 7, the part with a high abnormal score represents the abnormal picture. We marked the abnormal part with a red detection box.


**Figure 7.** Results of abnormal scores of some frames in Ped2.

The attention mechanism helps our model pay more attention to the main part of the video frame. After two attention unit modules, the main body of the video frame is gradually clear. Our model can better learn video features to achieve better detection results. Figure 8 shows the effect of the attention unit on datasets Ped1 and Ped2. The first column is the real ground map, and the second and third columns are the feature map after the attention unit. It can be seen that the model pays more attention to the main body of the picture after circulating the attention unit. The model can better learn the features of the subject so as to achieve higher detection accuracy.

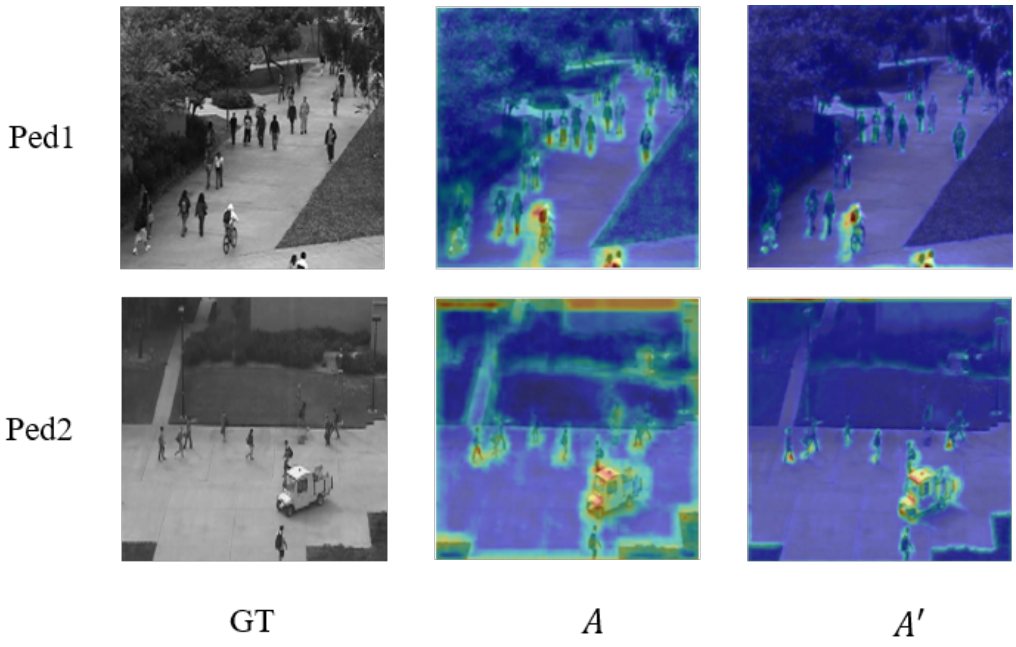


**Figure 8.** Example of the effect of attention mechanism in the detection process.

Some interesting information can be found in Table 3.

**Table 3.** Comparison of the experimental results with those of advanced algorithms.

| Method | Ped1 | Ped2 | Avenue |
|---|---|---|---|
| SF [26] | 67.5 | 55.6 | - |
| MDT [28] | 81.8 | 82.9 | - |
| Unmasking [27] | 68.4 | 82.2 | 80.6 |
| TSC [14] | - | 89.1 | 80.6 |
| Mem-AE [12] | - | 94.1 | 83.3 |
| Conv-AE [10] | 75.0 | 85.0 | 70.2 |
| Conv-LSTM [15] | 75.5 | 86.1 | 77.0 |
| r-GAN [29] | 83.7 | 95.9 | 85.3 |
| AMMC-Net [30] | - | 96.6 | 85.6 |
| DPU [31] | 78.9 | 96.5 | 79.3 |
| Ours | 80.5 | 97.9 | 85.9 |

(1) Our models are relative to SF, MDT, Unmasking, conv-AE, conv-LSTM, r-GAN and the baseline model DPU. On the Ped1 dataset, the AUC values of the samples were elevated by 13%, −1.3%, 12.1%, 5.5%, 5%, −3.2% and 1.6%, respectively. On the Ped2 dataset, the values were 42.3%, 15%, 15.7%, 8.8%, 3.8%, 12.9%, 11.8%, 2%, 1.3%, 1.4% compared to SF, MDT, Unmasking, TSC, Mem-AE, conv-AE, conv-LSTM, r-GAN, AMMC-Net and the baseline model, respectively. On the Avenue dataset, the improvements over Unmasking, TSC, Mem-AE, conv-AE, conv-LSTM, r-GAN, AMMC-Net and the baseline model were 5.3%, 5.3%, 2.6%, 15.7%, 8.9%, 0.6%, 0.3% and 6.6%, respectively.

(2) The attention mechanism does obtain the complete spatial information of the video sequence, helping us to get better video frame features and thus achieve higher accuracy. Especially on the Avenue dataset, the advantages of our model are more prominent. We hypothesized that the Avenue dataset is significant, and the attention mechanism is more advantageous for feature extraction in large datasets. Compared to other datasets and models, our model has the best lift on this dataset.

(3) We introduced separable convolution operation in the decoding stage. As can be seen from Tables 4 and 5, we improved the detection accuracy while reducing model parameters, which verified our model's effectiveness in video anomaly detection.

**Table 4.** Comparison of ablation results.

| Module | Ped1 | Ped2 | Avenue |
|---|---|---|---|
| base | 78.9 | 96.5 | 79.3 |
| +DSC | 80.5 | 96.6 | 82.6 |
| +Attention | 80.1 | 97.3 | 82.4 |
| +DSC+Attention | 80.8 | 97.9 | 85.9 |

To verify the influence of our depth-separable convolution operation on model parameters, we calculated model parameters and Floating point operations per second (Flops), and the specific results are shown in Table 5. The number of parameters in our model is reduced by more than four times. That is, the complexity of the model can be reduced while the accuracy can be further improved.

**Table 5.** Comparison of model complexities.

| Module | Params | Flops |
|---|---|---|
| Base | 13.19 M | 49.11 GMac |
| Ours | 3.27 M | 35.31 GMac |

Ablation Study

We added attention mechanisms and depth-separable convolution operations to the baseline model DPU. To verify the effectiveness of each part of the model, we conducted ablation studies on three datasets, and the experimental results are shown in Table 4 and Figure 9. Four model combinations, baseline model, baseline model + attention, baseline model + depth-separable convolution and baseline model + attention + depth-separable convolution, are compared and investigated. The operation with depth-separable convolution improves the accuracy by 1.6%, 0.1% and 3.3% on the three datasets. The models with the attention mechanism improved accuracy by 1.2%, 0.8% and 3.1% on the three datasets, respectively. In contrast, the combined model improved accuracy by 1.9%, 1.4% and 6.6% on the three datasets, respectively. Both the deep separable convolution operation and the attention mechanism provide performance improvements to the original model, while the combined model offers the best performance improvement to the original model.

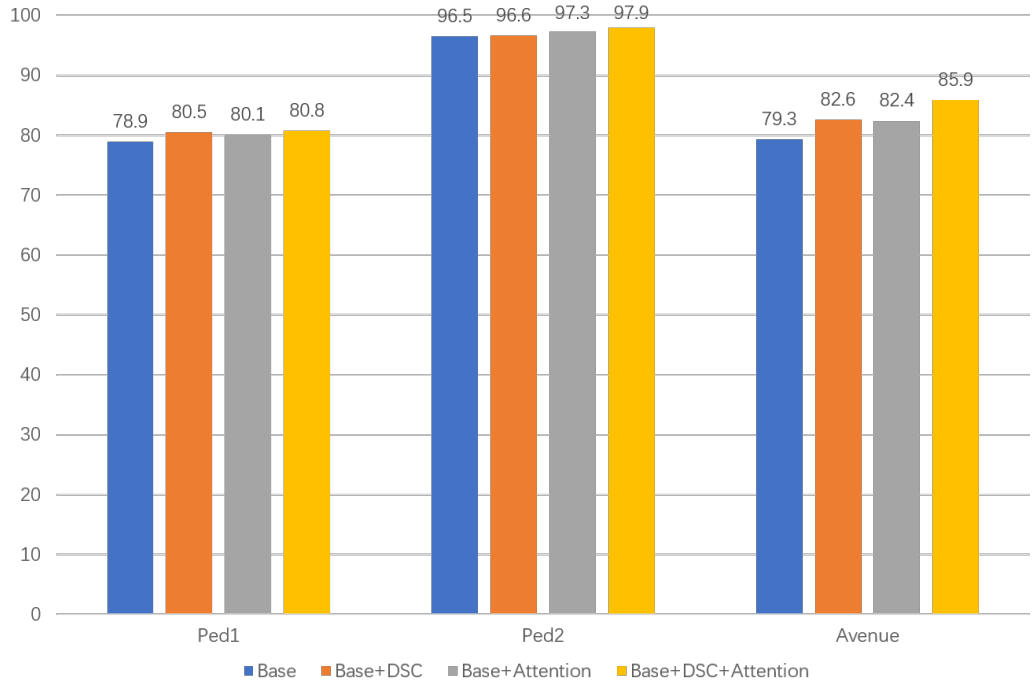


**Figure 9.** Histogram of ablation experiment results.

## 4. Discussion and Summary

In this paper, we first propose two main problems in video anomaly detection. Then we introduce the attention mechanism into the DPU model to improve the model detection accuracy. Then we introduce separable convolution operation to reduce the complexity of the model. Finally, we perform experimental research on three public datasets to prove the effectiveness of our algorithm in video detection. In addition, we also conducted ablation experiments to verify the effectiveness of the methods used in the model. Finally, we discuss the influence of feature distance loss weight on the detection ability of the model.

The excellent performance of the model in video anomaly detection mainly comes from the following two aspects:

(1) Our model eliminates the influence of the video background through dual attention units, thus facilitating the model to extract the main features. This improves the detection accuracy of the model.

(2) The separable convolution operation reduces the model parameters by nearly four times and accelerates the detection speed of the model. This improves the efficiency of model detection.

From the above analysis, we can understand that compared with the baseline model; our proposed model has excellent detection accuracy and speed. However, our model performs poorly on Ped1 and Avenue datasets. This means that the model is not very universal, which is also the common fault of almost all video anomaly detection models. In the future, we will consider combining more advanced deep learning methods to study more general detection models. It is convenient to solve more complex video data in the real world.

**Author Contributions:** Conceptualization, Q.Z. and H.W.; methodology, Q.Z. and H.W.; software, H.W.; validation, X.D., J.C. and J.Y.; resources, Q.Z.; writing—review and editing, Q.Z. and H.W.; supervision, X.D. and J.Y.; project administration, Q.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by The National Natural Science Foundation of China under Grant 61862060, Grant 61462079, Grant 61562086 and Grant 61562078. This work was supported in part by the National Natural Science Foundation of China Project under Grant 62262064, Grant 62266043, and Grant 61966035; in part by the Key R&D projects in Xinjiang Uygur Autonomous Region under Grant XJEDU2016S106; and in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region of China under Grant 2022D01C56.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data are contained within the article.

**Acknowledgments:** The authors would like to thank the editors and referees for their precious remarks and comments.

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

## References

1. Lv, H.; Zhou, C.; Xu, C.; Cui, Z.; Yang, J. Localizing anomalies from weakly-labeled videos. *arXiv* **2020**, arXiv:2008.08944.
2. Lin, I.-C.; Chang, C.-C.; Peng, C.-H. An Anomaly-Based IDS Framework Using Centroid-Based Classification. *Symmetry* **2022**, *14*, 105. [CrossRef]
3. Zhang, Y.; Lei, Y. Data Anomaly Detection of Bridge Structures Using Convolutional Neural Network Based on Structural Vibration Signals. *Symmetry* **2021**, *13*, 1186. [CrossRef]
4. Alsulami, A.A.; Abu Al-Haija, Q.; Alqahtani, A.; Alsini, R. Symmetrical Simulation Scheme for Anomaly Detection in Autonomous Vehicles Based on LSTM Model. *Symmetry* **2022**, *14*, 1450. [CrossRef]
5. Sabokrou, M.; Fathy, M.; Hoseini, M. Video anomaly detection and localization based on the sparsity and reconstruction error of auto-encoder. *Electron. Lett.* **2016**, *52*, 1122–1124. [CrossRef]
6. Sabokrou, M.; Khalooei, M.; Fathy, M.; Adeli, E. Adversarially Learned One-Class Classifier for Novelty Detection. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3379–3388.
7. 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 Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 12737–12747.
8. Deepak, K.; Chandrakala, S.; Mohan, C.K. Residual spatiotemporal autoencoder for unsupervised video anomaly detection. *Signal Image Video Process.* **2021**, *15*, 215–222. [CrossRef]
9. Zhang, Y.; Nie, X.; He, R.; Chen, M.; Yin, Y. Normality Learning in Multispace for Video Anomaly Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 3694–3706. [CrossRef]
10. Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A.K.; Davis, L.S. Learning Temporal Regularity in Video Sequences. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 733–742.
11. Park, H.; Noh, J.; Ham, B. Learning Memory-Guided Normality for Anomaly Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 14360–14369. [CrossRef]
12. 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 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 1705–1714.
13. Bergaoui, K.; Naji, Y.; Setkov, A.; Loesch, A.; Gouiffes, M.; Audigier, R. Object-Centric and Memory-Guided Normality Reconstruction for Video Anomaly Detection. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 2691–2695. [CrossRef]

14. Luo, W.; Liu, W.; Gao, S. A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 341–349.
15. 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, 10–14 July 2017; pp. 439–444.
16. Chang, Y.; Tu, Z.; Xie, W.; Luo, B.; Zhang, S.; Sui, H.; Yuan, J. Video anomaly detection with spatio-temporal dissociation. *Pattern Recognit.* **2022**, *122*, 108213. [CrossRef]
17. Jin, P.; Mou, L.; Xia, G.-S.; Zhu, X.X. Anomaly Detection in Aerial Videos Via Future Frame Prediction Networks. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 8237–8240. [CrossRef]
18. Sun, Y.; Cui, T.; An, G.; Ruan, Q. A Video Abnormal Detection Framework based on Appearance-Motion Fuse Memory. In Proceedings of the 2022 16th IEEE International Conference on Signal Processing (ICSP) 1, Beijing, China, 21–24 October 2022; pp. 535–539.
19. Ingle, P.Y.; Kim, Y.-G. Real-Time Abnormal Object Detection for Video Surveillance in Smart Cities. *Sensors* **2022**, *22*, 3862. [CrossRef]
20. Bian, Y.; Tang, X. Abnormal Detection in Big Data Video with an Improved Autoencoder. *Comput. Intell. Neurosci.* **2021**, *2021*, 9861533. [CrossRef] [PubMed]
21. 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.
22. Wang, C.; Yao, Y.; Yao, H. Video anomaly detection method based on future frame prediction and attention mechanism. In Proceedings of the 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 27–30 January 2021; pp. 405–407. [CrossRef]
23. 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 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 13568–13577. [CrossRef]
24. Narasimhan, M.; SowmyaKamath, S. Dynamic video anomaly detection and localization using sparse denoising autoencoders. *Multimed. Tools Appl.* **2017**, *77*, 13173–13195. [CrossRef]
25. Sultani, W.; Chen, C.; Shah, M. Real-World Anomaly Detection in Surveillance Videos. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6479–6488.
26. Mehran, R.; Oyama, A.; Shah, M. Abnormal crowd behavior detection using social force model. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 935–942.
27. Ionescu, R.T.; Smeureanu, S.; Alexe, B.; Popescu, M. Unmasking the Abnormal Events in Video. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2914–2922.
28. 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.
29. Lu, Y.; Yu, F.; Reddy, M.K.K.; Wang, Y. Few-shot Scene-adaptive Anomaly Detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
30. Cai, R.; Zhang, H.; Liu, W.; Gao, S.; Hao, Z. Appearance-Motion Memory Consistency Network for Video Anomaly Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021.
31. 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 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), virtual, 19–25 June 2021; pp. 15420–15429. [CrossRef]
32. Liu, W.; Luo, W.; Lian, D.; Gao, S. Future frame prediction for anomaly detection–A new baseline. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6536–6545.