Hindawi
Advances in Multimedia
Volume 2022, Article ID 1038225, 20 pages
https://doi.org/10.1155/2022/1038225

Research Article

A New Method of Pedestrian Abnormal Behavior Detection Based on Attention Guidance

Jingui Huang, Jingyi Li, and Wenya Wu

College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China

Correspondence should be addressed to Jingyi Li; 202020291631@hunnu.edu.cn

Received 5 November 2022; Revised 5 December 2022; Accepted 8 December 2022; Published 20 December 2022

Academic Editor: Yu-Chen Hu

In public places, some behavior that violates public order and endangers public safety is defined as abnormal behavior. Moreover, it is a necessary auxiliary means to maintain public order and safety by detecting abnormal behavior in a large number of surveillance videos. However, due to the small proportion of abnormal behavior in video data, the extreme imbalance of data seriously restricts the effectiveness of detection. So, weakly supervised learning has become the most suitable and effective detection method. However, existing weakly supervised methods rarely take the locality and slightness of abnormal behavior into account and ignore the details of extracted features. Based on this, an attention-directed abnormal behavior detection model is proposed. In the two common prediction and reconstruction abnormal behavior detection methods based on weak supervision, suitable attention mechanisms are introduced, respectively, and two corresponding attention-directed networks are proposed. In addition, aiming at the problem of inaccurate thresholds for abnormal behavior division, the loss function of the model is improved and a new abnormal behavior evaluation method is proposed. Experiments were carried out on three classical datasets (the USCD Ped1, USCD Ped2, and CUHK Avenue dataset) for abnormal behavior detection. The best results for the area under the curve (AUC) indicator reached 82.7%, 94.5%, and 87.3%, respectively, which are better than many existing literature results.

1. Introduction

Public safety has always received much attention. In recent years, monitoring equipment in public places has gradually increased and improved. Then, unsafe factors can be identified from massive surveillance video data [1], which can promptly find and stop suspected unsafe behavior and more effectively maintain public safety. In order to explore pedestrian behavior in monitoring, researchers first started with pedestrian trajectories and obtained complete traces of pedestrians from videos to summarize pedestrian behavior [2]. Through the analysis of traces, abnormal behaviors in monitoring can be divided into group and individual abnormal behaviors. The group abnormal behavior includes the sudden dispersion and gathering of the crowd and other abnormal behavior, which can be classified. Thus, the researchers directly applied the general deep action recognition framework [3] to the group abnormal behavior detection. They used optical flow and trajectory to classify and identify the abnormal and obvious movements of the population and obtained good results [4, 5]. Individual abnormal behavior is usually personal misconduct that violates public order, such as driving a vehicle on the sidewalk, violating the usual walking direction stipulated in public places, and so on. Since monitoring video in public places contains various kinds of individual abnormal behavior but a small number within every class and individual abnormal behavior accounts for a fairly small proportion of all behavior data, this imbalance will seriously affect the accuracy of detecting abnormal behavior according to general methods. Therefore, researchers have divided all behavior into two classes. Also, they have proposed a weakly supervised method to get the features of normal behavior data. Then, all behavior data are compared with the reconstructed or predicted behavior data by normal behavior features, and those with large differences are judged to be abnormal behavior frames [6].

At present, the weakly supervised method for abnormal behavior detection is divided into two kinds:

reconstruction and prediction according to the difference in input data [7]. The reconstruction method with a single-frame input can be applied to image and video data, and the prediction method with multiframe input can only be applied to video data. With the existing weakly supervised method for abnormal behavior detection, it is the key point to make the refactored abnormal behavior frames to be obviously different from the original data. At present, the models based on reconstruction and prediction mainly focus on this problem from two aspects. On the one hand, a variety of input data are input in the model of training normal behavior frames, and the appearance and motion features are extracted through the multibranch model to accurately refactor the normal behavior [8, 9]. On the other hand, a new module is proposed to record typical normal behavior features and increase the proportion of normal behavior features in the refactoring process [7, 10, 11]. The abovementioned studies help to realize that the refactored abnormal behavior will be far from the original data and the refactored normal behavior will be similar to the original data. So, the weak supervision method can complete the accurate detection of abnormal behavior. However, surveillance videos contain a large number of pedestrians, and pedestrian behavior usually occupies a small area. Also, the normal behavior features extracted by the existing models lack specificity, and it is easy to ignore the details of key regions, resulting in the failure of highlighting the obvious difference between abnormal behavior and normal behavior when refactoring all behaviors. To solve this problem, we introduce different attention modules in the encoder-decoder structure to capture the key areas of the input data and pay attention to the detailed features with increasing weights. At the same time, the added edge loss function focuses on the texture detail information of the data. In addition, the existing models have a single and inaccurate way to calculate abnormal scores of the refactored data. Therefore, it is difficult to select the appropriate division threshold with the abnormal scores to accurately distinguish abnormal behavior and normal behavior. To this end, we improve the abnormal evaluation method by using a combination of multiple evaluation methods to evaluate the quality of refactoring images, which increases the discrimination between refactored abnormal behavior and original abnormal behavior.

The main contributions of this paper are as follows: first, we propose an attention-directed deep model, which combines the encoder-decoder structure based on reconstruction and prediction methods, add attention modules that fit the structure of the encoder or decoder, and use edge loss to focus on the behavior region. Second, we propose a new abnormal evaluation method, which analyzes the quality of the refactored images from multiple aspects including the pixels and the whole of the images, and use the abnormal score as a standard to distinguish normal and abnormal behavior. Finally, we perform experiments on different datasets for abnormal behavior detection and verify the performance of the proposed models with the existing models through various experimental data such as ablation, quantitative, qualitative, and visualization experiments.

2. Related Work

Due to the imbalance between normal and abnormal behavior data, the method based on weak supervision is usually adopted, and the autoencoder is generally the infrastructure to extract normal behavior features. Then, all behavior data are reconstructed or predicted based on the extracted features. Finally, the normal and abnormal behavior data are distinguished according to the error between the reconstructed or predicted data and the original data. Chen and He [12] refracted video segments and video frames by using different branches to incorporate spatiotemporal information, which only contained normal behavior, and fused the reconstruction errors based on the Bayesian law. In order to increase the similarity of reconstructed normal behavior and input normal behavior, Gong et al. [7] added a memory module for recording the features of normal behavior, which were input into the decoder.

The autoencoder structure is prone to missing features due to linear compression; so, the deep learning model U-Net is used as the basic model, which is similar to the autoencoder. Li et al. [10] proposed a spatial-temporal model by combining U-Net for representing spatial information with the ConvLSTM for extracting motion information. Park et al. [13] improved the model proposed by Gong et al. [7]. Instead of the autoencoder, U-Net was selected as the basic model. The memory module was set with the appropriate amount of memory, and the feature compactness loss and separation loss were also set to effectively update the memory module. Based on the research by Park et al. [13], Lv et al. [11] set the memory module as a dynamic unit without occupying memory for reducing the cost of the model.

In order to process a large number of features in deep learning models, researchers have begun to apply attention mechanisms to various models of video anomaly detection. Inspired by the visual attention mechanism, Huan et al. [14] used the input data to obtain spatial-temporal anomaly salient maps and then combined the maps with video frames to detect abnormal behavior frames. Zhu and Newsam [15] used an attention mechanism to assign the weights to different instances in the loss function of the multiple instance method. Wei et al. [16] also used attention for multiple instance anomaly detection, and the attention mechanism was used to assign weights to the input C3D and optical flow features. Whereas, Sun and Ji [17] adopted attention-based addressing when looking for the most matching normal feature in the memory module.

Since the attention mechanism contains many variants, it can be divided into two categories according to its adaptability. One can only adapt to one model and the other can transfer to multiple models. Among them, Attention U-Net (UA) [18] belongs to the former based on U-Net, which connects and processes features with the same number of channels on both sides of the U-shaped structure to obtain feature weights. UA is currently mainly used in the field of medical image segmentation, such as liver tumor segmentation [19] and retinal vascular segmentation [20]. SE (Squeeze-and-excitation) attention [21] belongs to the latter, which can be flexibly transferred, and it selects valid features

by adjusting the channel weights of model features. SE is commonly used in the field of image classification, such as microexpression recognition [22] and ECG (electrocardiogram) classification [23].

Inspired by the abovementioned research, we use U-Net as the basic model and combine multilayer attention modules UA and SE, which are appropriate for two different networks of reconstruction and prediction methods. As the attention-directed model, two new networks strengthen the extraction of feature details for video anomaly detection.

3. Methods

3.1. Model Structure. This paper proposes an attention-directed model named Att_AE. The model structure is shown in Figure 1, which includes the encoder, decoder, memory module, loss function, and abnormal evaluation module. According to the different network structures in reconstruction and prediction methods, the parts of the encoder and decoder are different. So, the two attention mechanisms adapted to the two networks are added to the encoder or decoder module to form new structures. The multilayer attention module UA [18] is constructed in the decoder of the prediction method and the features of different layers are calculated. The multilayer SE attention module [21] is added to the encoder of the reconstruction method to calculate the feature weights between channels.

The memory module is located in the middle of the encoder and decoder, which has the same function in the reconstruction and prediction methods. The memory module is used to select and store the normal behavior features obtained by the encoder. The stored features are used to connect with the features obtained by the encoder and then input into the decoder. The memory module widens the gap between reconstructed or predicted normal behavior frames and abnormal behavior frames by increasing the proportion of normal behavior features in the input of the decoder. In addition, the training loss is also improved, and the edge constraint on the images is added by the image gradient. In addition, a new abnormal behavior evaluation module is proposed to improve the calculation method of abnormal scores among testing data, so that the model can more accurately distinguish normal behavior frame and abnormal behavior frame.

3.2. Attention-Directed Encoder-Decoder Model. The encoder and decoder modules in the U-Net reflect the attention directed by combining attention mechanisms to extract important features. In this paper, the corresponding attention mechanisms are set up according to the reconstruction and prediction model structures.

3.2.1. Attention Encoder-Decoder Structure in Prediction Method. For the prediction method of inputting four consecutive frames to predict the next frame, a multiframe input model is proposed, and the multiframe images split from one video are input to the encoder in a multichannel way. According to the skip connection in the prediction network structure, the attention module UA that fits the model is introduced into it. The transformed encoder features are connected to the decoder features with the same number of channels for inputting into the attention module. Then, the feature weight of the next layer in the decoder is calculated. Finally, the purpose of improving the weights of local interest regions and suppressing the noninterest regions is achieved by three-layer attention modules of the decoder. The details of the prediction model are shown in Figure 2 below.

In Figure 2, the four-frame continuous frames are input into the encoder as a multichannel feature, and the spatial resolution is reduced by convolution and maxpooling to obtain advanced semantic features. Then, the output of the encoder is input into the memory module and the input feature is stitched with the normal behavior feature that is most similar stored in the memory module. After stitching, the feature is entered into the decoder. In order to avoid the loss of feature details, the skip connection is used to stitch features of the same resolution. In order to increase the attention to the key features, we combine skip connection, input the same resolution features in the encoder and decoder into the UA module, assign weights to them by UA, then realize the skip connection with the features in the encoder, and finally restore the image resolution through convolution and deconvolution to output the refactored image.

Figure 3 displays a UA structure in the prediction method. The input is a middle feature in U-Net. $\mathbf{F_g}$ and $\mathbf{F_l}$ are the outputs of different layers in the encoder and decoder. $\mathbf{F_l}$ is the output of the layer in the decoder, whose next layer has the same dimensions as $\mathbf{F_g}$. g and l indicate that they are obtained from the encoder and decoder, respectively. First, this paper uses an upsampling method to ensure that the number of channels of $\mathbf{F}_l$ and $\mathbf{F_g}$ are consistent. Then, we add the processed $\mathbf{F}_l$ and $\mathbf{F_g}$, input it into the ReLU layer for sparse processing, increase the nonlinearity through the Sigmoid layer to obtain the attention weight coefficient $\mathbf{O}$, and multiply $\mathbf{O}$ with the original input $\mathbf{F}_l$ to get the output $\mathbf{F}_1'$ in the current attention module.

$$\mathbf{F}_1' = \mathbf{O} \times \mathbf{F}_l. \tag{1}$$

This output is connected to $\mathbf{F}_l$ and is used to calculate the next attention weight coefficient with the corresponding encoder input. Finally, more efficient features can be obtained through the multilayer attention module.

3.2.2. Attention Encoder-Decoder Structure in Reconstruction Method. For the reconstruction method of inputting a single frame to refactor the input frame, Figure 4 shows the specific network structure. The input data are a single video frame and the resolution is reduced by convolution and maxpooling. To select and strengthen important features, the SE attention module is added to extract the attention between channels. The SE module is located after the convolutional block of the encoder to calculate the weight of each channel in the encoder. After adding the weighted feature to the original feature, the new feature is input to the

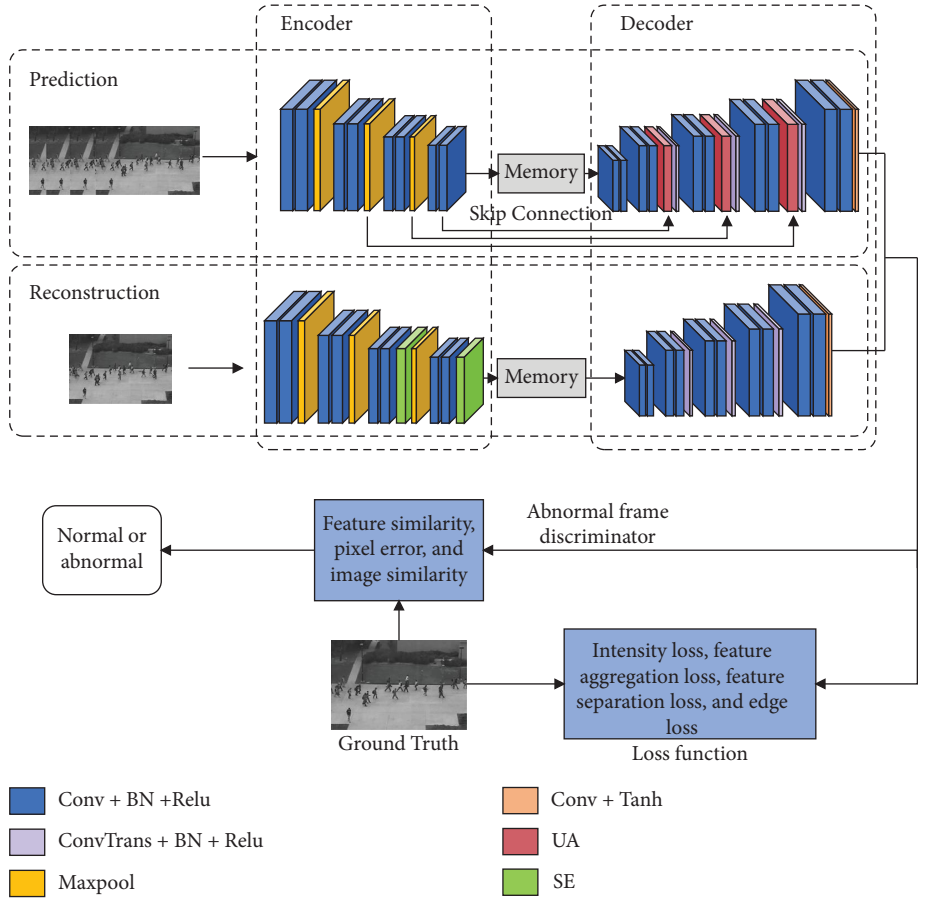

Figure 1: Att_AE model structure with two methods.

next layer of the network for continued training. Moreover, the more effective normal behavior features are selected by the multilayer attention module, so that the memory module that records the normal behavior features and the decoder can train with the more effective features.

Since the input frame is a single frame, the skip connection with the power generative ability of the deep learning model may cause the refactored image to become a replica of the input image. So, the skip connection is removed from the reconstruction model. Then, through a series of convolution and deconvolution operations, the output data are restored to the same size as the input data and the refactored image is output.

The structure of the SE attention module in the reconstruction method is shown in Figure 5. Feature **S** in the encoder is input into the SE module after convolutional blocks and its size is converted to $1 \times 1 \times Z$ from $H \times W \times Z$ by the global average pooling method. The pooling method is similar to the function of the special fully connected layer, and the calculation equation is as follows:

$$\mathbf{T} = \text{Global\_Avgpooling}(\mathbf{S}) = \frac{1}{H \times W} \sum_{m=1}^{H} \sum_{n=1}^{W} \mathbf{S}(m, n). \quad (2)$$

The output **T** is the new feature with the size of 1×1, which is obtained after averaging the pixel values of each channel. m and n represent the height and width of the pixel coordinates in the feature map, respectively. Then, **T** is input into a fully connected layer FC with the coefficient $(Z, Z/r)$, and the value of r is 16. The purpose is to reduce the number of feature channels and the amount of calculation for the model. Next, we input into the ReLU layer for nonlinear operation, then input into another fully connected layer FC′ with the coefficient $(Z/r, Z)$ to recover the original channel number, and finally input into the Sigmoid layer to normalize. The whole process can be expressed by the following equation:

$$\mathbf{U} = \sigma(E_2 \delta(E_1 \mathbf{T})), \quad (3)$$

where σ represents the Sigmoid layer, δ represents the ReLU layer, and E_1 and E_2 are the parameters in the FC and FC′. The channel attention weight **U** with the size of $1 \times 1 \times Z$ is obtained. **U** is used to multiply with the original input **S** to get the output **V** of the attention module. Finally, the obtained features according to the multilayer convolutional block and attention block in the encoder are input into the memory and decoder module to continue training.

3.3. Training Loss. The loss function has 4 parts, the intensity loss Loss_p [7], which constrains pixel similarity, the feature aggregation loss Loss_{f_g} and separation loss Loss_{f_s}[13], which ensure that the memory module can record typical normal behavior, and the edge loss Loss_g, which

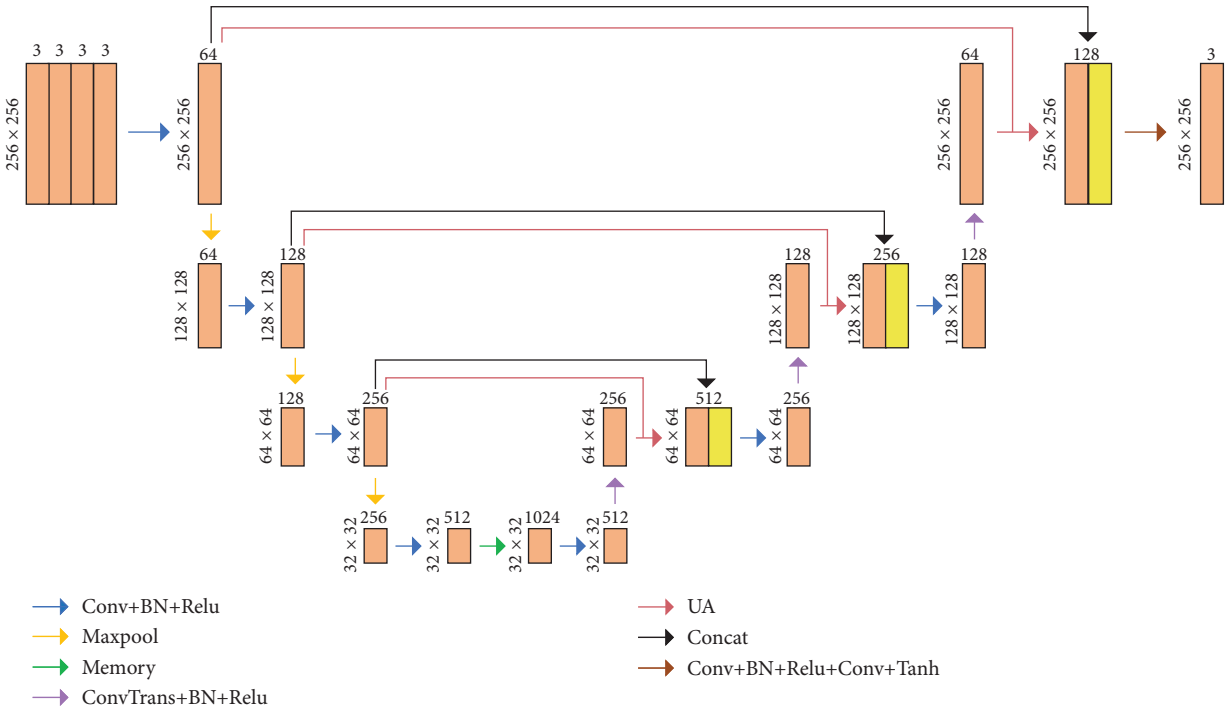

Figure 2: Prediction model structure.

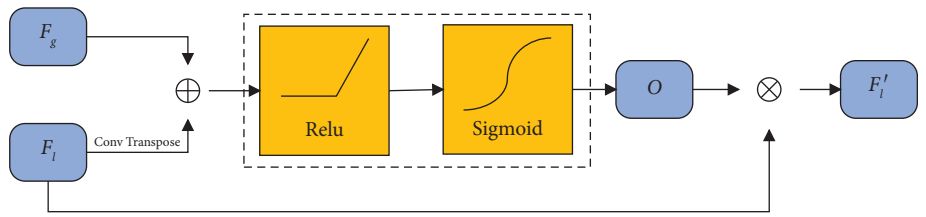

Figure 3: Attention module in prediction method.

characterizes the edges of the image. The loss function is defined as follows:

$$\text{Loss} = \text{Loss}_p + \gamma \text{Loss}_{f\_g} + \theta \text{Loss}_{f\_s} + \beta \text{Loss}_g, \tag{4}$$

where γ, θ, and β are weight factors corresponding to the different parts of the loss function ($0 < \gamma, \theta, \beta < 1$).

The intensity loss reduces the difference between real and refactored data by penalizing the distance between them calculated by pixel similarity. Specifically, the intensity loss Loss_p is calculated by using the mean squared difference between two images with the following equation:

$$\text{Loss}_p(x_i, x_i') = \frac{1}{I} \sum_{i=1}^{I} (x_i' - x_i)^2, \tag{5}$$

where x_i and x_i' represent the pixel value of the original image and the prediction or reconstruction image, respectively. I represents the total number of pixels in the image.

The feature aggregation loss is set for normal behavior features stored in memory modules. The feature aggregation loss Loss_{f_g} ensures that the input features are similar to the recorded features in the memory module, and the calculation equation is as follows:

$$\text{Loss}_{f\_g} = \sum_{k}^{K} \left\| \mathbf{f_k} - \mathbf{mem_c} \right\|_2, \tag{6}$$

where $\mathbf{f_k}$ represents the features extracted by the model. K is the total number of extracted features. $\mathbf{mem}_c$ is the closest normal behavior feature to $\mathbf{f_k}$ of the memory module. By constraining the difference between the input features and the features in the memory module by equation (6), it is convenient to ensure the similarity between the stored features in the memory module and the input normal features. Thus, it is conducive to the refactoring quality of normal behavior frames after inputting the two into the decoder.

Based on the fact that the input feature is matched with the most similar feature in the memory module, the number of normal behavior features in the memory module is also required to control for reducing the memory. Therefore, the features stored in the memory module are required to be representative and different. The feature separation loss

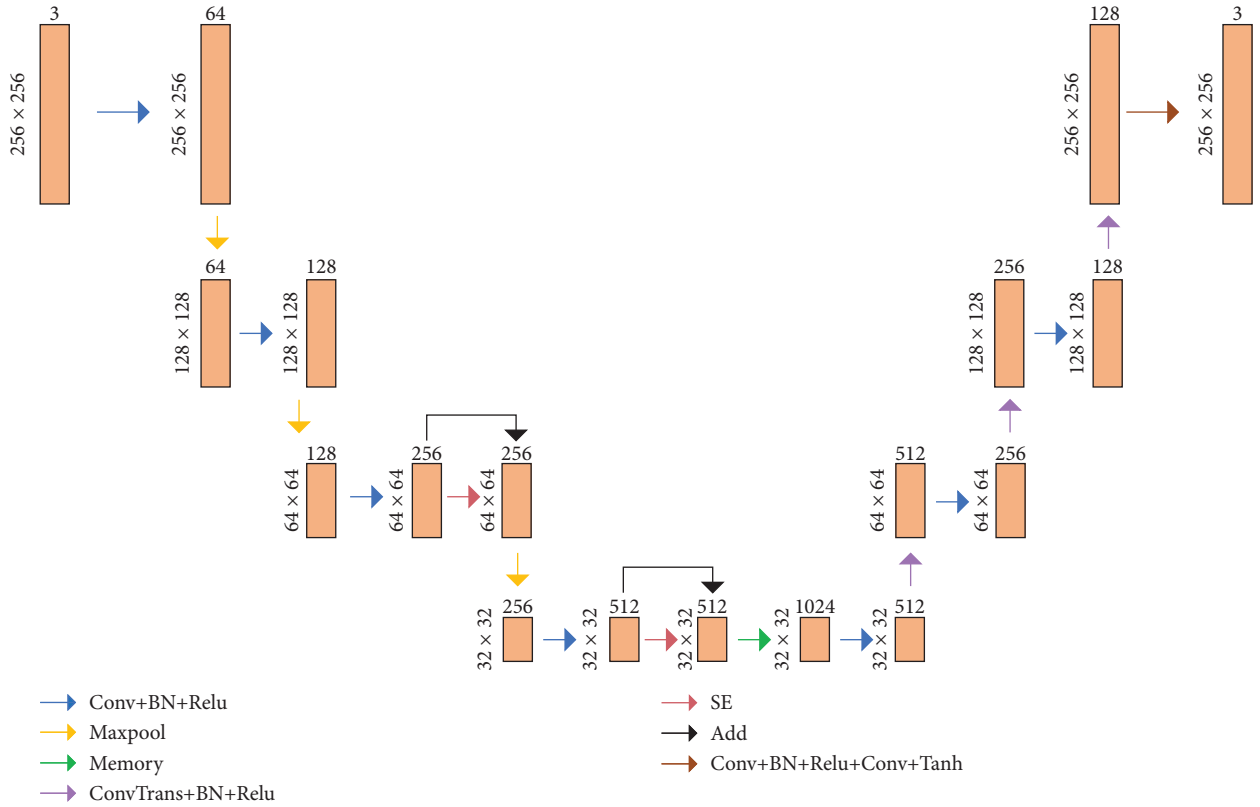

Figure 4: Reconstruction model structure.

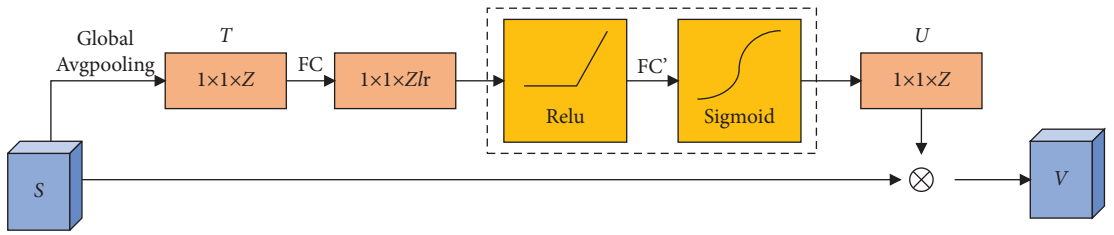

Figure 5: Attention module in reconstruction method.

Loss_{f_s} is used to guarantee the diversity of features in the memory module, and the equation is as follows:

$$\text{Loss}_{f\_s} = \sum_{k}^{K} \left(\left\| \mathbf{f_k} - \mathbf{mem_c} \right\|_2 - \left\| \mathbf{f_k} - \mathbf{mem_s} \right\|_2 + \alpha \right), \tag{7}$$

where $\mathbf{mem_s}$ is the second closest normal behavior feature to $\mathbf{f_k}$ of the memory module. α is set to make the loss greater than zero. The feature separation loss limits the distance between the input feature and the second similar feature in the memory module, which increases the difference between $\mathbf{mem_c}$ and $\mathbf{mem_s}$. Thus, the feature separation loss increases the difference among all features in the memory module.

For more clearly delineating the outline edges of the contents in the input frame, the model sets the edge loss Loss_g to consider the details of texture structure well. This loss mainly constrains the horizontal and vertical gradients of the image calculated by the Sobel operator [24].

$$\mathbf{G_x} = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \mathbf{G_y} = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}, \tag{8}$$

where $\mathbf{G_x}$ and $\mathbf{G_y}$ are the convolutional kernel that computes the image gradient horizontally and vertically, respectively. $\mathbf{G}$ and $\mathbf{G}'$ are the gradients of the original image and the reconstruction or prediction image, respectively. $\mathbf{X_g}$ and $\mathbf{Y_g}$ represent the horizontal and vertical gradients obtained by the convolution between the original image and the Sobel operator, respectively. $\mathbf{X_g'}$ and $\mathbf{Y_g'}$ represent the horizontal and vertical gradients of the reconstruction or prediction

image and the Sobel operator, respectively. The image gradient is calculated as follows:

$$\begin{aligned}\mathbf{G} &= \mathrm{sqrt}\left(\mathbf{X}_{\mathbf{g}}^2 + \mathbf{Y}_{\mathbf{g}}^2\right),\\ \mathbf{G}' &= \mathrm{sqrt}\left(\mathbf{X}'^2_{\mathbf{g}} + \mathbf{Y}'^2_{\mathbf{g}}\right).\end{aligned} \tag{9}$$

We define the gradient difference as the edge loss and use the L_1 distance to calculate the edge loss Loss_g. In equation (10), the calculated edge loss improves the detail of the refactored image by constraining the gradient difference between the real image and the output refactored image.

$$\mathrm{Loss}_g = \left\|\mathbf{G}' - \mathbf{G}\right\|. \tag{10}$$

3.4. Abnormal Behavior Evaluation. After training with the normal behavior data, the abnormal behavior evaluation is used to provide abnormal scores for testing frames. This paper proposes a new method to judge abnormal behavior, which is shown in equation (11). It includes feature similarity D, pixel error PSNR, and image similarity SSIM. Then, we fuse three values according to a certain proportion to calculate the final abnormal score. If the value is high, the frame will be much more possible to include abnormal behavior.

$$\mathrm{Score}_{\mathrm{anomaly}} = \lambda D\left(\mathbf{f}, \mathbf{mem}\right) + \eta\left(1 - \mathrm{PSNR}\left(\mathbf{x}, \mathbf{x}'\right)\right) + \varphi\left(1 - \mathrm{SSIM}\left(\mathbf{x}, \mathbf{x}'\right)\right), \tag{11}$$

where $0 < \lambda, \eta, \varphi < 1$ and $\lambda = 1 - \eta - \varphi$, $\mathbf{x}, \mathbf{x}'$ are vectors of the input image and output image.

The difference is calculated between the normal behavior features in the memory module and the features obtained in testing. Moreover, it is an effective way of determining whether it is an abnormal behavior frame. If the distance is large, the probability of the abnormal behavior frame will be large. The model uses L_2 distance to calculate the distance, and the used equation is as follows:

$$D\left(\mathbf{f}, \mathbf{mem}\right) = \frac{1}{K}\sum_{k}^{K}\left\|\mathbf{f}_{\mathbf{k}} - \mathbf{mem}_{\mathbf{c}}\right\|_2. \tag{12}$$

PSNR is one of the most widely used image evaluation indicators [8] and is usually used to measure the gap between the distorted image and the original image. Moreover, this paper uses it to judge the difference in pixels between the output image and the input image. The equation PSNR is as follows:

$$\mathrm{PSNR}\left(\mathbf{x}, \mathbf{x}'\right) = 10\ \log_{10}\left(\frac{\max\left(\mathbf{x}'\right)}{1/I\sum_{i=1}^{I}\left(\mathbf{x}' - \mathbf{x}\right)^2}\right), \tag{13}$$

where $\max\left(\mathbf{x}'\right)$ represents the maximum pixel value of the input image. The high value of PSNR indicates that the output image is similar to the input image, and the input image is judged as a normal behavior frame. Otherwise, the input image is judged as an abnormal behavior frame.

PSNR is only concerned with pixel error, so another image similarity evaluation criterion SSIM [25] is added to calculate the similarity between whole images from three aspects, brightness, contrast, and structural similarity. Moreover, the lower the value, the greater the probability of an abnormal behavior frame. The calculation equation is as follows:

$$\mathrm{SSIM}\left(\mathbf{x}, \mathbf{x}'\right) = l\left(\mathbf{x}, \mathbf{x}'\right)\cdot c\left(\mathbf{x}, \mathbf{x}'\right)\cdot s\left(\mathbf{x}, \mathbf{x}'\right). \tag{14}$$

$\mu_{\mathbf{x}}$ and $\mu_{\mathbf{x}'}$ represent the average gray level of the input image and the output image, respectively. The intensity similarity $l\left(\mathbf{x}, \mathbf{x}'\right)$ is calculated as follows:

$$l\left(\mathbf{x}, \mathbf{x}'\right) = \frac{2\mu_{\mathbf{x}}\mu_{\mathbf{x}'} + C_1}{\mu_{\mathbf{x}}^2 + \mu_{\mathbf{x}'}^2 + C_1}. \tag{15}$$

C_1 is set to prevent the denominator from being zero, and $l\left(\mathbf{x}, \mathbf{x}'\right)$ is always in the range of $(0, 1]$. $\sigma_{\mathbf{x}}$ and $\sigma_{\mathbf{x}'}$ represent the standard deviation of the input image and output image, respectively. The contrast similarity $c\left(\mathbf{x}, \mathbf{x}'\right)$ is calculated as follows:

$$c\left(\mathbf{x}, \mathbf{x}'\right) = \frac{2\sigma_{\mathbf{x}}\sigma_{\mathbf{x}'} + C_2}{\sigma_{\mathbf{x}}^2 + \sigma_{\mathbf{x}'}^2 + C_2}. \tag{16}$$

The function of C_2 is the same as C_1, and the range of $c\left(\mathbf{x}, \mathbf{x}'\right)$ is the same as $l\left(\mathbf{x}, \mathbf{x}'\right)$. Finally, the structural similarity $s\left(\mathbf{x}, \mathbf{x}'\right)$ of images is calculated, and the calculation equation is as follows:

$$s\left(\mathbf{x}, \mathbf{x}'\right) = \frac{\sigma_{\mathbf{x}\mathbf{x}'} + C_3}{\sigma_{\mathbf{x}}\sigma_{\mathbf{x}'} + C_3}. \tag{17}$$

C_3 is also set to prevent the denominator from being zero. $\sigma_{\mathbf{x}\mathbf{x}'}$ represents the covariance of the input image and output image.

The similarity between the input image and output image can be calculated according to equation (14). Since the mean and variance of the entire image usually have a large variation, the sliding window method is used to calculate the multiregion SSIM of the image. Finally, the average value of the multiregion SSIM is taken as the final result.

4. Experiments

For verifying the performance of the model, the proposed model is trained on three classic abnormal behavior detection datasets. The experiments include parameter selection, ablation experiments, comparison experiments, and visualization of abnormal behavior.

4.1. Dataset and Experimental Environment. In this paper, three public datasets are selected: USCD Ped1 [26], USCD Ped2 [26], and CUHK Avenue [27].

The USCD Ped1 dataset and the USCD Ped2 dataset [26] are pedestrian monitoring videos at the University of California, San Diego. The Ped1 dataset contains 34 training set videos and 36 testing set videos with a total of 14000 frames. The Ped2 dataset contains 16 training set videos and 12 testing set videos with a total of 4560 frames. The labels of the two datasets are frame-level labels, and the training set only contains the normal behavior frames and the testing set

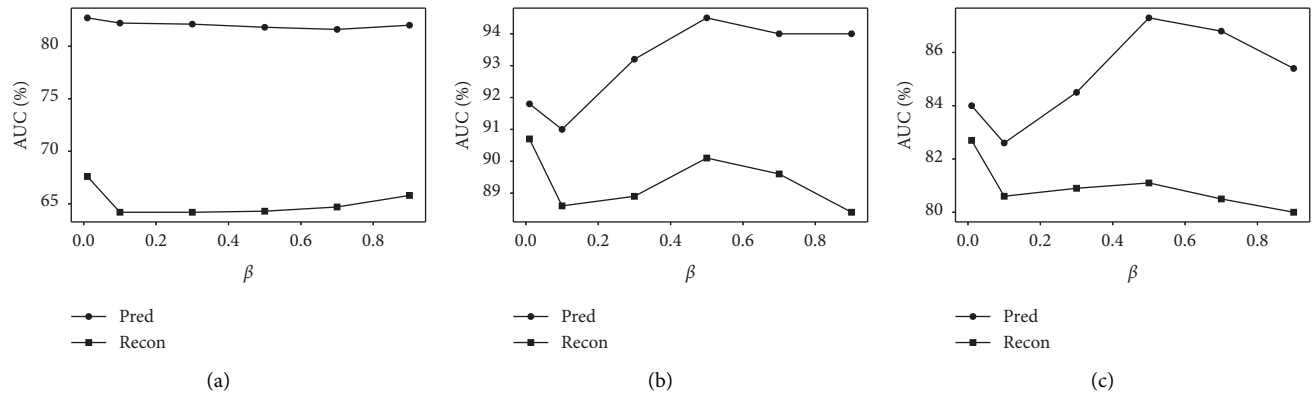

Figure 6: Prediction and reconstruction model results with different β values ($\beta = 0.01$, 0.1, 0.3, 0.5, 0.7, and 0.9). (a) Ped1 dataset. (b) Ped2 dataset. (c) Avenue dataset.

contains the normal and abnormal behavior frames. The videos record crowd behavior on campus sidewalks, and the labeled abnormal behavior mainly includes cycling, skateboarding, and driving.

The CUHK Avenue dataset [27] contains 16 training set videos and 21 testing set videos with a total of 30652 frames, and the labels are also frame-level labels. The division of training and testing sets is similar to the Ped1 and Ped2 datasets. The Avenue dataset records the crowd behavior at the subway entrance, and the abnormal behavior mainly includes fast running, abnormal actions in places, and wrong walking direction.

The experimental environment framework is Python 3.6, the compilation tool is PyCharm, the graphics card is TITAN XP, and the video memory is 12 GB. The number of training epochs is 60, the initial learning rates used by the prediction and reconstruction methods are $2e-4$ and $2e-5$, and the optimizer is Adam. The two models use the same loss function, as defined in 3.3 Training Loss. Based on the abovementioned parameters, the training time of the Ped1, Ped2, and Avenue datasets is about 5, 2, and 10 hours, respectively. The time performance of the Ped1, Ped2, and Avenue datasets is 58.5 FPS (Frames Per Second), 58.5 FPS, and 50 FPS, respectively.

4.2. Model Parameters. The model parameters mainly include the weights of the loss function and the weights of the abnormal behavior evaluation module. Moreover, the parameters are selected according to the experimental results by combining different weights. The evaluation indicator of the model is AUC [7], which is used for binary classification models. The larger the AUC, the better the result of the model.

4.2.1. Weights of Training Loss. The loss function of the model is defined in 3.3 Training Loss, which consists of 4 parts, including intensity loss, feature aggregation loss, feature separation loss, and edge loss.

According to equation (4) in Section 3.3, only the weight parameters of the three losses are required to determine, and they are defined as γ, θ, and β. γ and θ are still set to 0.01 in the reconstruction method and 0.1 in the prediction method, respectively [13]. Taking β as the only variable, the weight parameters that correspond to the optimal result are chosen by training different models with different values. Figure 6 shows the experimental results of prediction and reconstruction models with different β on three datasets. According to Figure 6, the construction model has the best performance with β of 0.01 on three datasets. The prediction model has the best performance with β of 0.01 on the Ped1 dataset, and with β of 0.5 on the Ped2, Avenue dataset.

Table 1: Abnormal behavior evaluation for different weights of prediction model on Ped2 dataset (different combinations of η and φ).

η	φ			
	0.01	0.05	0.1	0.3
0.5	94.34	94.24	94.04	92.62
0.55	94.45	94.33	94.0	92.50
0.6	**94.49**	94.31	94.05	92.26
0.65	94.45	94.25	93.94	91.90
0.7	94.36	94.14	93.79	91.46

4.2.2. Weights of Abnormal Behavior Evaluation. The abnormal behavior evaluation module consists of three parts, feature similarity, pixel error, and image similarity. And the corresponding weight parameters are set as λ, η, φ in Section 3.4. Since the sum of feature similarity and pixel error weights is set to 1 [13], this paper sets $\lambda = 1 - \eta - \varphi$. η is set to 0.7 and 0.6 in reconstruction and prediction methods [13]. Therefore, this paper sets η around the value of literature [13] and then adjusts the value of the newly added weight φ. Table 1 shows the results (AUC%) with different weights of the prediction model on Ped2 dataset.

According to the experimental results in Table 1, when η is 0.6 and φ is 0.01, the performance of the prediction model on the Ped2 dataset is the best. Based on the abovementioned method, when η is 0.6 of the prediction models on the Ped2 dataset and Avenue dataset, 0.75 of other models on other datasets, and φ is 0.01 of all models, the trained models achieve their optimal performance.

Table 2: Ablation experiments of attention-directed abnormal behavior detection model.

Method	Attention module	Edge loss	Abnormal behavior evaluation	Ped1 dataset	Ped2 dataset	Avenue dataset
Reconstruction	✗	✗	✗	66.3	89.3	80.5
	✓	✗	✗	67.1	89.7	82.1
	✗	✓	✗	67.0	90.5	81.8
	✗	✗	✓	66.7	89.4	80.5
	✓	✓	✗	67.4	90.6	82.6
	✓	✗	✓	67.3	89.8	82.1
	✗	✓	✓	67.3	90.6	81.9
	✓	✓	✓	**67.6**	**90.7**	**82.7**
Prediction	✗	✗	✗	79.8	92.0	83.4
	✓	✗	✗	81.4	92.8	85.6
	✗	✓	✗	80.9	93.6	87.0
	✗	✗	✓	81.0	92.3	84.1
	✓	✓	✗	81.1	**94.5**	87.2
	✓	✗	✓	82.1	93.0	85.6
	✗	✓	✓	81.9	93.5	87.0
	✓	✓	✓	**82.7**	**94.5**	**87.3**

4.3. Ablation Experiments. To analyze each added part that affects the accuracy of the proposed model, ablation experiments are performed on three datasets, and the results are shown in Table 2. The data of the first row in the two method areas are the baseline of three datasets and the other rows include the model results of different adding parts. Through data comparison, it can be seen that the single part and different combinations of parts have a promotion effect on the performance of abnormal behavior detection.

Regarding the single added part, in the reconstruction model, the attention module increases to the most effective values of 0.8%, 0.4%, and 1.6% on the Ped1, Ped2, and Avenue datasets. In the prediction model, the edge loss increases to the most effective values of 1.1%, 1.6%, and 3.6% on three datasets. Moreover, other adding parts also improve the performance. Thus, for the single added part, the attention module and the edge loss bring the most improvements in reconstruction and prediction models, respectively.

About the combination model parts, in the reconstruction model, the combination of the attention module and the edge loss brings more increases with 1.1%, 1.3%, and 2.1% on the Ped1, Ped2, and Avenue datasets; the final model brings the most improvement with 1.3%, 1.4%, and 2.2% on three datasets. In the prediction model, the combination of the attention and the abnormal behavior evaluation brings more increases with 2.3% on the Ped1 dataset; the combination of the attention module and the edge loss brings more increases with 2.5% and 3.8% on Ped2 and Avenue datasets; the final model brings the most improvement with 2.9%, 2.5%, and 3.9% on three datasets. With the results of the combinations, the attention module, edge loss, and abnormal behavior evaluation are positively correlated with the model result. Moreover, all combinations of the three parts are beneficial to improve performance.

4.4. Quantitative Experiments. In this section, the experimental results of the model in this paper are compared with the existing abnormal behavior detection models with the same indicator on three datasets. In the comparison models, unmasking [28] was the model to learn the effective classifier through sliding windows, the level set method [29] was the model that used horizontal set detection to extract image descriptors, SRNN [32] and sRNN-AE [35] used RNN as the basic model, GAN_pred [33] used GAN combined with U-Net, PST [36] used pose components for abnormal behavior detection, and others were variant models based on autoencoder and U-Net. Table 3 shows the results of the proposed model and the comparison models on three datasets. In Table 4, we use another evaluation index, Equal Error Rate (EER) [30], to compare the results of this model with others on the same dataset. EER is used to measure the error rate of the model, and a smaller EER value means a lower error rate of the model.

The model in this paper includes two methods for detecting abnormal behavior. Tables 3 and 4 are separately divided into three areas to compare the proposed model with other models separately. The first area includes some abnormal behavior detection models that are not based on reconstruction or prediction methods; the second area includes the models using the reconstruction method; the third area includes the models using the prediction method. Att_AE_recon and Att_AE_pred are the proposed networks with reconstruction and prediction methods in this paper. In Tables 3 and 4, the results of Att_AE_recon and Att_AE_pred are significantly better than those of the comparison models in the corresponding areas. Moreover, the result of Att_AE_pred is better than those of all models.

4.5. Qualitative Experiments

4.5.1. Abnormal Behavior Scores. The abnormal behavior dataset selected in this paper is labeled with frame-level labels, and the specific labels are 0 and 1. 0 represents the normal behavior frame and 1 represents the abnormal behavior frame. In this paper, the model scores the testing videos containing normal and abnormal behavior frames through the abnormal behavior evaluation defined in

Table 3: Comparative experiments of the attention-directed abnormal behavior detection model (AUC%).

Method	Model	Ped1	Ped2	Avenue
—	Unmasking [28]	68.4	82.2	80.6
	Level set method [29]	79.3	89.8	—
	Two-stream R-ConvVAE [30]	75.0	91.0	79.6
	Two stream with OFF [24]	—	—	84.3
Reconstruction	CAE [31]	58.5	84.7	77.2
	AE_mem_recon [13]	66.3	89.3	80.5
	Att_AE_recon	**67.6**	**90.7**	**82.7**
Prediction	SRNN [32]	—	92.2	81.7
	GAN_pred [33]	82.4	93.0	84.6
	AE_mem_pred [13]	79.8	92.0	83.4
	ST-3DCAE [34]	80.7	85.3	81.0
	sRNN-AE [35]	—	92.2	83.5
	Two stream 3DAE [9]	79.6	90.3	82.0
	Memory_att [17]	—	—	85.7
	PST [36]	—	—	86.7
	Att_AE_pred	**82.7**	**94.5**	**87.3**

Table 4: Comparative experiments of the attention-directed abnormal behavior detection model (EER%).

Method	Model	Ped1	Ped2	Avenue
—	Unmasking [28]	31.2	—	—
	Level set method [29]	25.8	14.0	—
	Two-stream R-ConvVAE [30]	32.4	15.5	27.5
	Two stream with OFF [24]	—	—	22.9
Reconstruction	CAE [31]	43.1	24.5	27.0
	AE_mem_recon [13]	37.6	18.7	27.4
	Att_AE_recon	**37.0**	**17.2**	**24.6**
Prediction	SRNN [32]	—	—	24.7
	GAN_pred [33]	24.0	15.6	22.5
	AE_mem_pred [13]	26.9	15.0	23.3
	ST-3DCAE [34]	25.1	21.8	24.9
	sRNN-AE [35]	-	14.9	23.2
	Two stream 3DAE [9]	27.2	15.5	24.8
	Memory_att [17]	—	—	21.6
	PST [36]	—	—	20.8
	Att_AE_pred	**23.9**	**13.7**	**19.5**

Section 3.4 and sets the appropriate threshold to divide normal and abnormal behavior frames. In Figures 7–9, it is shown that the partial anomaly scores and corresponding division thresholds on the test set given by the prediction model of this paper and the literature [13] in the Ped1, Ped2, and Avenue datasets.

In Figures 7–9, score1 is used to represent the model scores in literature [13], score2 is used to represent the model scores in this paper, the label is used to represent the label of the testing videos, and the threshold is the division value of abnormal and normal behavior. The abnormal score is the credential that divides the normal and abnormal behavior frames. If the abnormal scores are close to the corresponding labels, it will be easy to distinguish normal and abnormal behavior. In Figures 7–9, it can be seen that the abnormal scores of the proposed model are closer to the dataset labels. It can also be seen from the thresholds in Figures 7–9that the model in this paper has a higher accuracy rate of dividing normal and abnormal behavior. So, the model in this paper can more accurately divide the normal and abnormal behavior frames than the model in the literature [13].

4.5.2. Abnormal Behavior Detection Effect. For abnormal behavior detection, we refactor the test set data that mix abnormal behavior and normal behavior according to the features of normal behavior and obtain the refactoring error by the output image and the input image. If it is a normal behavior frame, its refactoring error should be small; if it is an abnormal behavior frame, the refactoring error should be large. In order to evaluate the model performance, we visualize the refactoring error obtained by the prediction model of literature [13] and this paper. The good performance of the proposed model in abnormal behavior detection can be seen through the comparison of figures. Figures 10 and 11 show some refactoring error pictures obtained from three datasets.

In Figures 10 and 11, the first column is the original video frame, the second column is the refactoring error

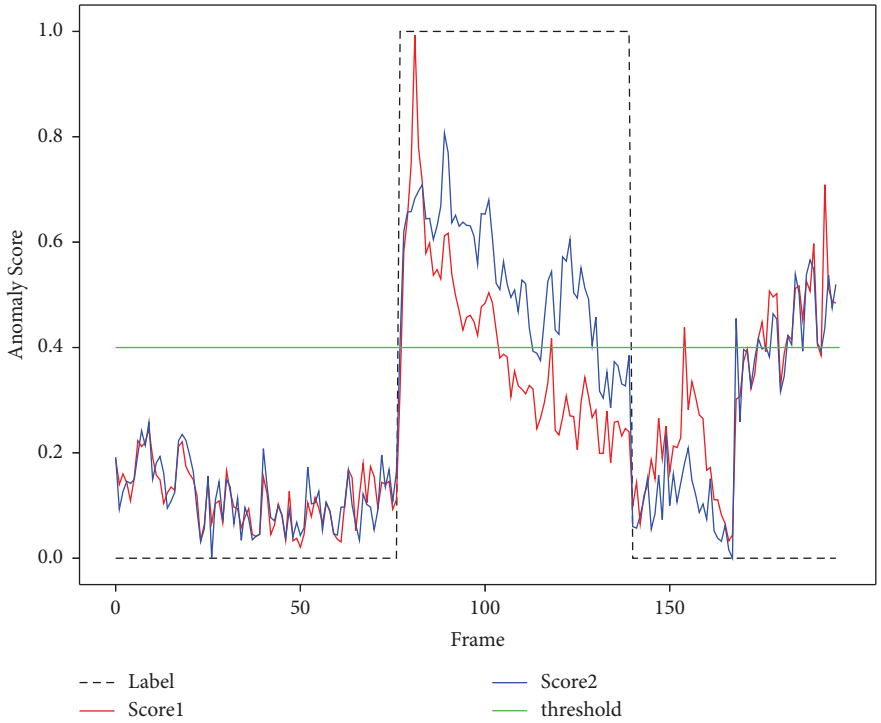

Figure 7: Anomaly scores on the Ped1 testing set.

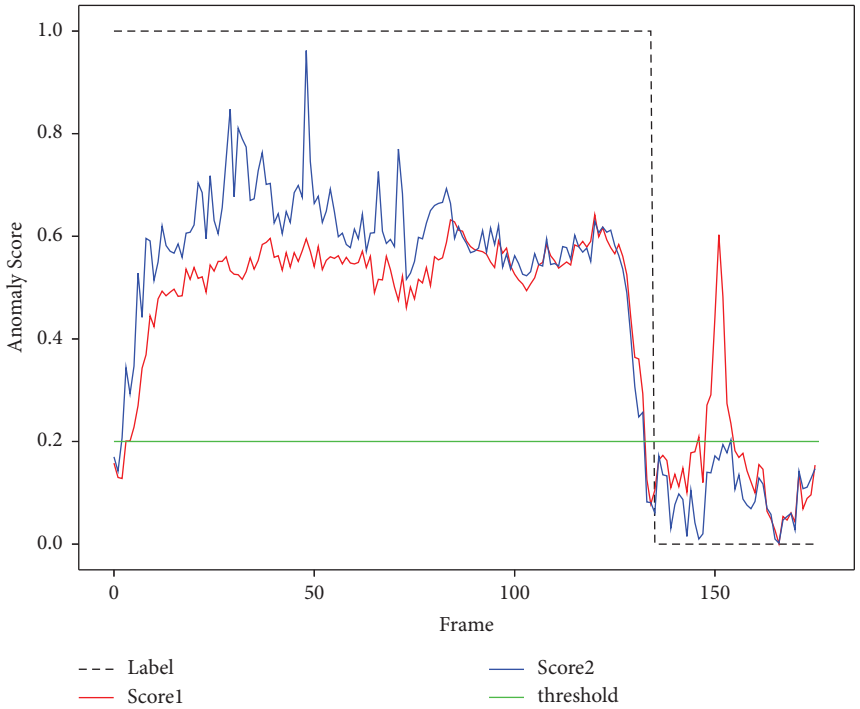

Figure 8: Anomaly scores on the Ped2 testing set.

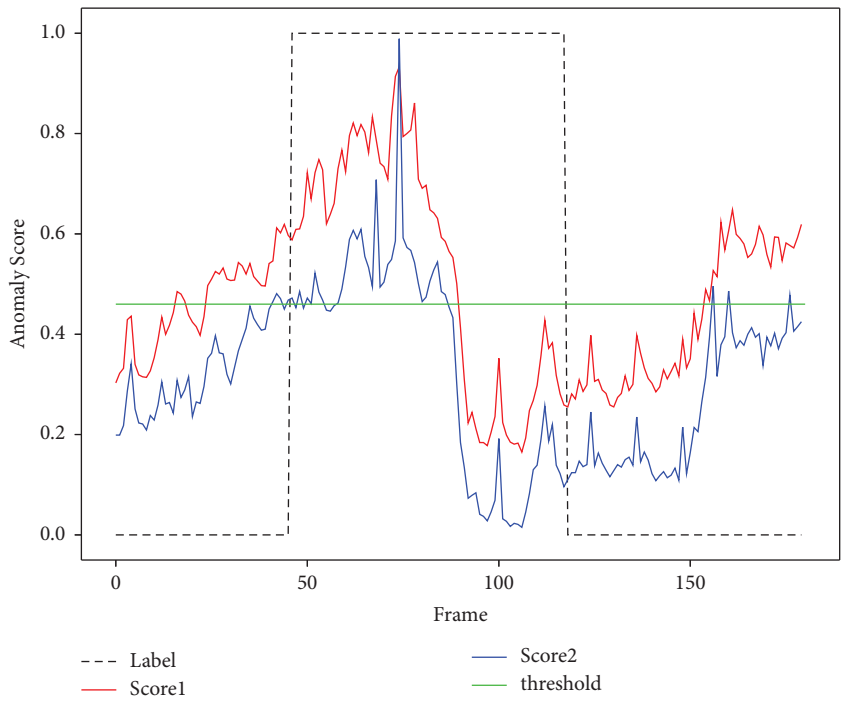

Figure 9: Anomaly scores on the Avenue testing set.

obtained by the literature [13], and the third column is the refactoring error obtained by this paper. The images in the figures are selected from the Ped1, Ped2, and Avenue datasets by row major. Figure 10 shows the refactoring error images of the normal behavior frames in the literature [13] and this paper, and it can be seen that the refactoring error of the normal behavior frames in this paper is less than that of the literature [13]. Figure 11 shows the refactoring error images of abnormal behavior in the literature [13] and this paper, where abnormal behavior is marked with a green rectangle box. Through the comparison of the second and third columns in Figure 11, it can be seen that the refactoring error obtained in this paper for the abnormal behavior frame is larger than that of the literature [13], especially the marked abnormal behavior area. Therefore, based on the visual comparison, the abnormal behavior detection performance of the proposed model is better.

4.6. Comparison of Attention Mechanisms. For exploring the effectiveness of the attention mechanisms in this paper, the CBAM module for calculating spatial and channel attention [37], and the ECA module for calculating channel attention [38] are selected for comparison. The same number of different attention modules are added in the same way to train the Ped2 dataset with the prediction and reconstruction networks. Moreover, the AUC (%) of the different reconstruction and prediction networks is shown in Figures 12 and 13.

SE module in the reconstruction method and the UA module in the prediction method can achieve better results, which are used in this paper. In addition, the average time performance of these networks is around 55FPS, which proves that the attention mechanisms used in this paper are the more effective scheme.

4.7. Visualization of Abnormal Behavior. The model in this paper fully learns the normal behavior features in the training process, so that the normal behavior frames can be effectively reconstructed or predicted in the testing set. There is a large reconstruction or prediction error of abnormal behavior frames in the testing set. The reconstruction or prediction error is the difference between the input image and the output image. Also, the specific areas of abnormal behavior can be observed by the visualization of the error. As shown in Figures 14 and 15, the prediction error of the Ped2 dataset and the Avenue dataset is visualized by the prediction network.

In Figure 14, the first column includes the abnormal behavior frames selected in the original testing set. The second column includes the difference between input images and output images, which can more clearly identify the specific areas of abnormal behavior. The third column includes the mark of the abnormal behavior area on the original frames. By marking the prediction error obtained by the proposed model in red, the scene and specific meaning of the abnormal behavior can be explored. In Figure 14, it can

Figure 10: Refactoring error of normal behavior frames.

be seen that the abnormal behavior selected in the Ped2 dataset includes cycling, driving cars, and skateboarding.

There are also three columns of images in Figure 15, and each column is with the same meaning as in Figure 14. Through the red annotation of the third column, it can be seen that the abnormal behavior selected in the Avenue dataset includes walking in the wrong direction, running fast, and making abnormal actions (throwing the backpack in place). Therefore, while correctly distinguishing between normal and abnormal behavior frames, the proposed model can get the specific area of abnormal behavior, which is effective for more accurate analysis and induction of abnormal behavior combined with the scenes.

4.8. Runtime. With a TITAN XP, the average time of the two baseline models [13] is 54.8 FPS and the average time of the models in this paper is 55.7 FPS. Also, the average time of the models in this paper is faster than other state-of-art models. For example, the average time of unmasking [28], SRNN [32], GAN_pred [33], and sRNN-AE [35] are 20 FPS, 50 FPS, 53.4 FPS, and 10 FPS, respectively.

5. Discussion

The proposed model is trained and tested on three abnormal behavior detection datasets, and the good performance of the proposed model is verified by comparison with other

Figure 11: Refactoring error of abnormal behavior frames.

model results. However, in the process of analyzing and induction, we also find that there are still some problems in the proposed model, which lead to errors in distinguishing between abnormal behavior and normal behavior. For further work, we will discuss some problems arising from the proposed model and analyze the possible causes and then propose the solutions as the direction of future work.

Figure 16 shows some detecting errors, and the data are selected from different datasets. In the abnormal scoring chart including the error frame, the location of the detection error frame is marked with a green circle. Also, the original data and corresponding refactoring error are provided for analyzing the cause of the detection error. It can be seen that the detection error frames are located near the demarcation between abnormal and normal behavior. The abnormal behavior frame in the first row is judged to be a normal behavior frame. From the abnormal behavior area marked by the green box, it can be seen that the abnormal behavior in Figure 16 is a cyclist in the crowd, which may be judged to be a normal behavior frame because it is located at the edge of the image and is obscured by the pedestrian in front. The second row includes a normal behavior frame. The abnormal behavior data in front of the normal behavior frame are that the pedestrian throws the backpack in place and walks quickly. We think the reason why this frame is judged to be an abnormal behavior frame may be without considering changes in motion speed and the similarity between this frame and its front abnormal data.

Based on the abovementioned analysis, we believe that the next step should be concerned with how to extract comprehensive features and solve judgment errors caused by

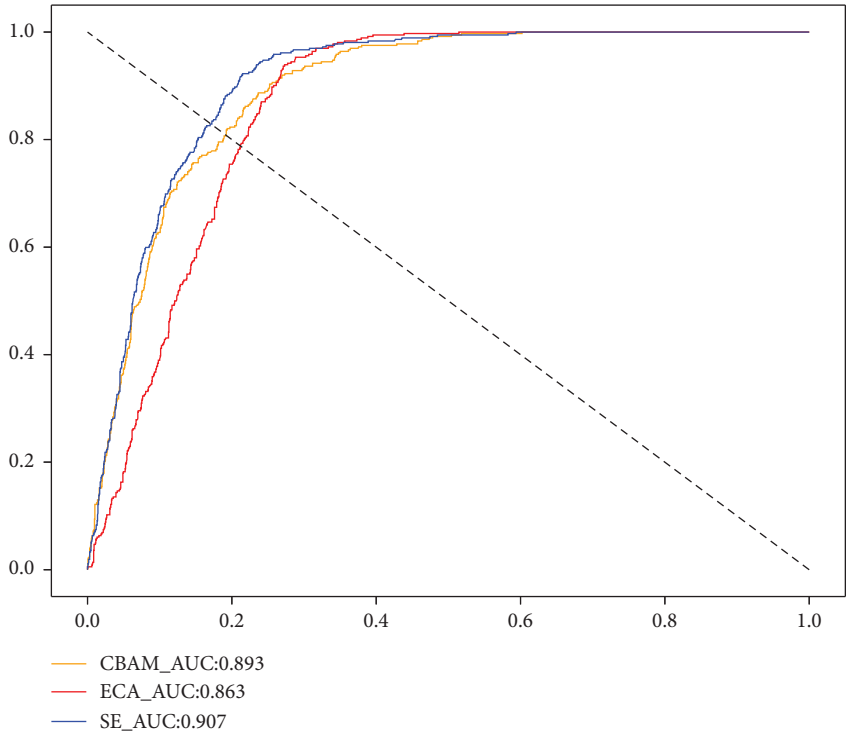

Figure 12: Reconstruction method results in different attention mechanisms.

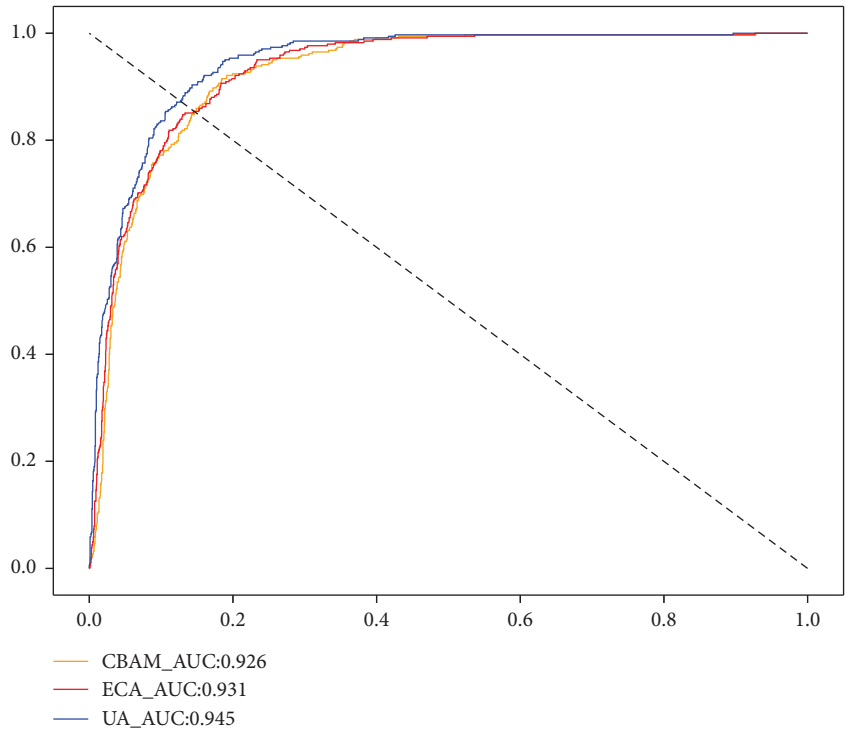

Figure 13: Prediction method results in different attention mechanisms.

occlusion and movement. Due to the difference in the motion state between abnormal behavior and normal behavior frames, we consider extracting motion features based on interframe variation to capture the difference for improving the judgment result of frames around the dividing line.

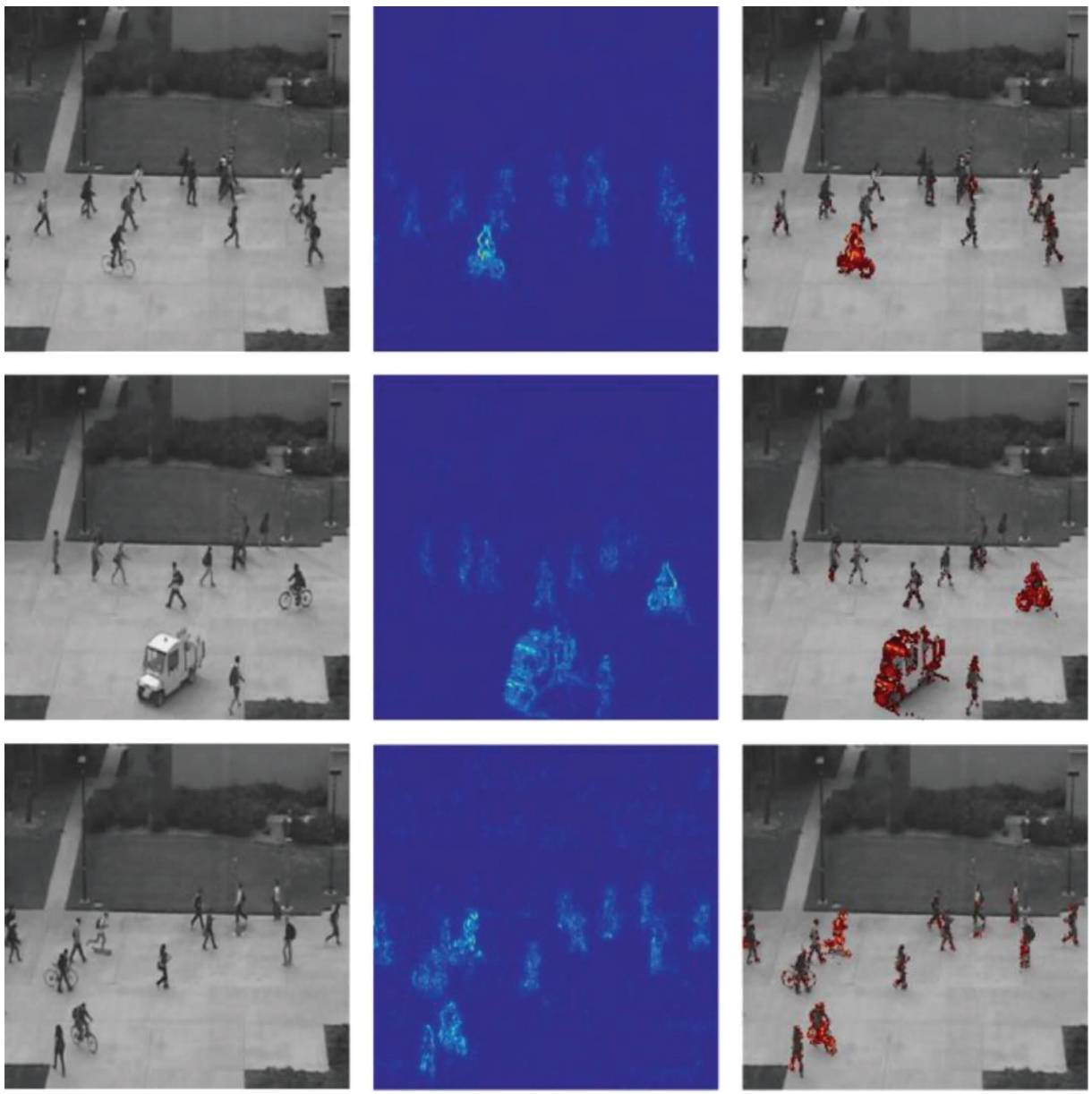

Figure 14: Visualization results of the Ped2 dataset.

Figure 15: Visualization results of the Avenue dataset.

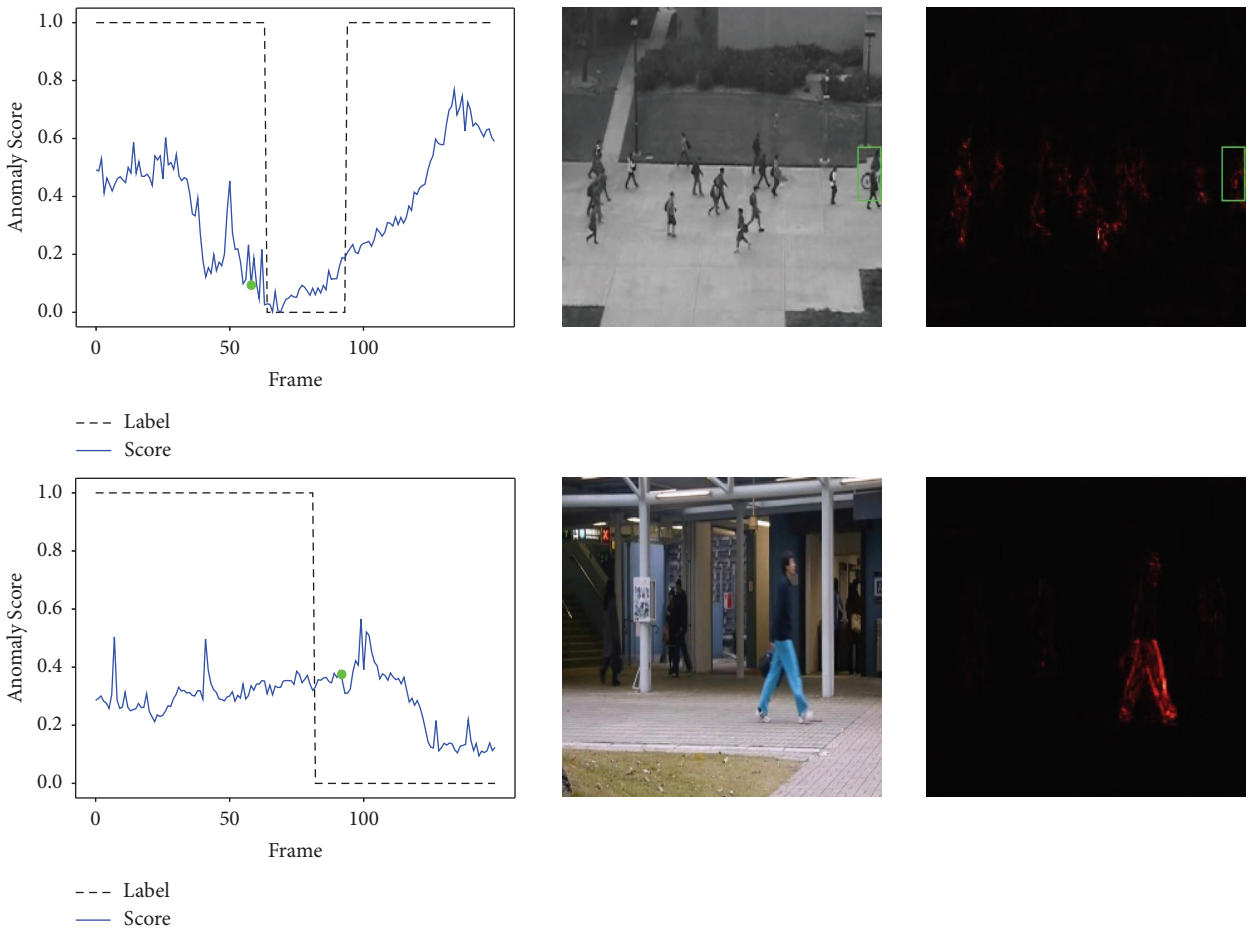

Figure 16: Abnormal behavior detection error.

6. Conclusions

Based on the weak supervision, this paper proposes an attention-directed abnormal behavior detection model for the situation that cannot detect local abnormal behavior effectively. The multilayer attention modules are added to obtain key features adapted to different structures of prediction and reconstruction methods. On this basis, this paper also modifies the loss function and proposes a new abnormal behavior evaluation module to increase the gap between normal and abnormal behavior frames after reconstruction or prediction, which is beneficial to the effective detection of abnormal behavior. Experiments on the Ped1, Ped2, and Avenue datasets have verified the advancement of the proposed model.

The proposed model in this paper has improved the performance of abnormal behavior detection to a certain extent, but the model still has some problems. For example, at the partial boundaries between the normal and abnormal behavior frames in videos, the difference between the abnormal behavior data and the normal behavior features extracted by the model is still small, which may lead to misjudgment. Therefore, further work is required to study how to extract the full and discriminating features of normal behavior for more effective abnormal behavior detection.

Data Availability

The three abnormal behavior datasets are public datasets. The download URL of USCD Ped1 and Ped2 is http://www.svcl.ucsd.edu/projects/anomaly/dataset.html. The download URL of CUHK Avenue is https://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62077014) and Research and Development Projects in Key Areas of Hunan Province, China (No. 2019SK2161).

References

[1] S. H. Cho and H. B. Kang, "Abnormal behavior detection using hybrid agents in crowded scenes," *Pattern Recognition Letters*, vol. 44, pp. 64–70, 2014.

[2] S. D. Khan, S. Bandini, S. Basalamah, and G. Vizzari, "Analyzing crowd behavior in naturalistic conditions: identifying

sources and sinks and characterizing main flows," *Neurocomputing*, vol. 177, pp. 543–563, 2016.

[3] M. Ullah, M. M. Yamin, A. Mohammed, and L. Chen, "Attention-based LSTM network for action recognition in sports," *Electronic Imaging: Intelligent Robotics and Industrial Applications using Computer Vision*, vol. 10, pp. 302–311, 2021.

[4] M. U. Farooq, M. N. M. Saad, and S. D. Khan, "Motion-shape-based deep learning approach for divergence behavior detection in high-density crowd," *The Visual Computer*, vol. 38, no. 5, pp. 1553–1577, 2022.

[5] A. J. Alzahrani and S. D. Khan, "Characterization of different crowd behaviors using novel deep learning framework," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 29, no. 1, pp. 169–185, 2021.

[6] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, Souel Korrea, June 2018.

[7] D. Gong, L. Liu, and V. Le, "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*, pp. 1705–1714, Seoul, South Korea, 2019.

[8] Z. Li, Z. Wang, H. Chen, L. Li, and X. He, "Video abnormal behavior detection based on dual prediction model of appearance and motion features," *Journal of Computer Applications*, vol. 41, no. 10, pp. 2997–3003, 2021.

[9] Z. Wang, X. Zhou, H. Yan, and J. Wang, "Abnormal behavior detection model based on two stream structure," *Computer Applications and Software*, vol. 39, no. 2, pp. 188–193, 2022.

[10] Y. Li, Y. Cai, J. Liu, S. Lang, and X. Zhang, "Spatio-temporal unity networking for video anomaly detection," *IEEE Access*, vol. 7, pp. 172425–172432, 2019.

[11] H. Lv, C. Chen, Z. Cui, L.Y. Xu, and J. Yang, "Learning normal dynamics in videos with meta prototype network," in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15420–15429, Nashville, TN, USA, November 2021.

[12] Y. Chen and D. He, "Spatial-temporal stream anomaly detection based on bayesian fusion," *Journal of Electronics and Information Technology*, vol. 41, no. 5, pp. 1137–1144, 2019.

[13] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14360–14369, Souel Korea, June 2020.

[14] W. Huan, H. Guo, and X. Wu, "Saliency attention based abnormal event detection in video," in *Proceedings of the 2014 IEEE International Conference on Robotics and Biomimetics*, pp. 1039–1043, Beijing China, October 2014.

[15] Y. Zhu and S. Newsam, "Motion-aware feature for improved video anomaly detection," in *Proceedings of the British Machine Vision Conference*, pp. 270–281, NEW York China, September 2019.

[16] S. Wei, G. Ji, Z. Xu, and X. Xiao, "Attention mechanism based multiple instance learning video anomaly detection," *Journal of Chinese Computer Systems*, vol. 1, pp. 1–9, 2022.

[17] J. Sun and J. Ji, "Memory-augmented deep autoencoder model for video anomaly detection," *Infrared and Laser Engineering*, vol. 51, no. 6, pp. 368–374, 2022.

[18] J. Schlemper, O. Oktay, M. Schaap et al., "Attention gated networks: learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197–207, 2019.

[19] C. Li, Y. Tan, W. Chen, X. Luo, and Y. Gao, "Attention Unet++: a nested attention-aware U-Net for liver CT image segmentation," in *Proceedings of the 2020 IEEE International Conference on Image Processing*, pp. 345–349, Abu Dhabi, United Arab Emirates, September 2020.

[20] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, and C. Fa, "SA-UNet: spatial attention U-Net for retinal vessel segmentation," in *Proceedings of the 2020 25th International Conference on Pattern Recognition*, pp. 1236–1242, Milan, Italy, May 2021.

[21] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[22] L. Yao, X. Xiao, R. Cao, and F. Xiao, "Three stream 3D CNN with SE block for micro-expression recognition," in *Proceedings of the 2020 International Conference on Computer Engineering and Application*, pp. 439–443, Beijing China, July 2020.

[23] J. Chen, T. Chen, B. Xiao, and S. Chen, "SE-ECGNet: multi-scale SE-Net for multi-lead ECG data," in *Proceedings of the 2020 Computing in Cardiology*, pp. 1–4, Berlin Garmany, May 2020.

[24] J. Cui, Z. Zhang, Q. Fang, and F. Zhang, "Video abnormal behavior detection based on optical flow constraints of adjacent frames," *Journal of Suzhou University of Science and Technology (Natural Science Edition)*, vol. 38, no. 3, pp. 76–84, 2021.

[25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[26] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 18–32, 2014.

[27] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proceedings of the 2013 IEEE International Conference on Computer Vision*, pp. 2720–2727, Sydney, NSW, Australia, March 2013.

[28] R. T. Ionescu, S. Smeureanu, B. Alexe, and R. Zhang, "Unmasking the abnormal events in video," in *Proceedings of the 2017 IEEE International Conference on Computer Vision*, pp. 2914–2922, Seoul Korea, April 2017.

[29] L. Kangwei, W. Jianhua, and H. Zhongzhi, "Abnormal event detection and localization using level set based on hybrid features," *Signal, Image and Video Processing*, vol. 12, no. 2, pp. 255–261, 2018.

[30] S. Yan, J. S. Smith, W. Lu, and B. Zhang, "Abnormal event detection from videos using a two-stream recurrent variational autoencoder," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 1, pp. 30–42, 2020.

[31] M. Ribeiro, A. E. Lazzaretti, and H. S. Lopes, "A study of deep convolutional auto-encoders for anomaly detection in videos," *Pattern Recognition Letters*, vol. 105, pp. 13–22, 2018.

[32] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 341–349, Berlin Germany, May 2017.

[33] W. Liu, W. Luo, D. Lian, and C. Chen, "Future frame prediction for anomaly detection - a new baseline," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6536–6545, New York, October 2018.

[34] J. Lian, X. Hu, and Y. Huang, "Video anomalous behavior detection based on 3D convolutional auto-encoder,"

Intelligent Computer and Applications, vol. 11, no. 6, pp. 70–75, 2021.

[35] W. Luo, W. Liu, D. Lian et al., "Video anomaly detection with sparse coding inspired deep neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1070–1084, 2021.

[36] W. Pang, Q. He, and Y. Li, "Predicting skeleton trajectories using a Skeleton-Transformer for video anomaly detection," *Multimedia Systems*, vol. 28, no. 4, pp. 1481–1494, 2022.

[37] S. Woo, J. Park, J. Lee, and S. Liu, "CBAM: convolutional block attention module," in *Proceedings of the 2018 European Conference on Computer Vision*, pp. 3–19, NY China, July 2018.

[38] Q. Wang, B. Wu, P. Zhu, and L. Xiang, "ECA-net: efficient channel attention for deep convolutional neural networks," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11531–11539, Beijing China, June 2020.