Article

ADL: Anomaly Detection and Localization in Crowded Scenes Using Hybrid Methods

Hequn Wu , Bai Rui , Liu Li *

Posted Date: 29 May 2023

doi: 10.20944/preprints202305.1617.v3

Keywords: Anomaly detection; YOLACT; Foreground; PWC-Net; Tracking

Article

ADL: Anomaly Detection and Localization in Crowded Scenes Using Hybrid Methods

Wu Hequn [1], Bai Rui [3,4,5] and Liu Li [2,*]

[1] Department of General Education, Zhejiang Police College, Hangzhou 310053, China; wuhequn@zjjcxy.cn
[2] School of Economics and Social Welfare, Zhejiang Shuren University, Shuren Road No.8, Hangzhou 310015, China
[3] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China; bairui@hdu.edu.cn
[4] Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province, Hangzhou 310018, China
[5] Ningxia Hui Autonomous Region Department of Agriculture and Rural Affairs Information Center, Yinchuan 750004, China
[*] Correspondence: casliuli@hotmail.com

Abstract: In recent years, video anomaly detection technology, which can intelligently analyze massive video and quickly find abnormal phenomena, has attracted extensive attention with the wide application of video surveillance technology. To address the complex and diverse problem of abnormal human behavior detection in surveillance videos, a surveillance video abnormal behavior detection and localization supervised method based on the deep network model and the traditional method is proposed. Specifically, we combined AGMM and YOLACT methods to obtain more accurate foreground information by fusing the foreground maps extracted by each technique. To further improve the accuracy, we use the PWC-Net technique to extract features of the foreground images and input them into an anomaly classification model for classification. The proposed method effectively detects and locates the abnormal behavior in the monitoring scene. In addition to the aforementioned methods, this paper also employs YOLOv5 and DeepSORT networks for object detection and tracking in the video, which allows us to track the detected objects for better understanding of the scene in the video. Experiments on the UCSD benchmark dataset and the comparison with state-of-the-art schemes prove the advantages of our method.

Keywords: anomaly detection; YOLACT; foreground; PWC-Net; tracking

1. Introduction

With the continuous population growth, human activities have become increasingly frequent, and increasing unusual emergencies have occurred, such as gathering and fleeing, fighting, and terrorist activities. Public security has become a major problem affecting social order. It is not uncommon to see news reports of modern transportation being maliciously used as a tool for committing crimes, such as deliberately driving a car onto a sidewalk with the intention of hitting pedestrians. According to FOX31 Denver, the suspect drove at high speed on the road, causing three victims to be injured, and there are also some other similar cases as shown in the Figure 1. With the development of social security, surveillance videos have been gradually adopted in ensuring people's safety in public places and for crime prevention [1]. Artificial intelligence and big data technologies can further enhance the security capabilities of public places. For example, computer vision technology can be used to automatically recognize and analyze surveillance footage, quickly detect abnormal behavior, and issue warnings. Data mining techniques can be used to analyze historical crime data, grasp criminal patterns and trends, and provide scientific basis for security decision-making. Intelligent recommendation algorithms can be used to screen people entering and exiting public places, thus achieving the detection and control of suspicious individuals. The application of these technologies can improve the safety of public places, ensuring the personal and property security of citizens.

Figure 1. According to CBS New York (1, 2, 3), a red swerves onto the sidewalk, it is a threat to public safety. According to WSAZ News Channel 3 (4, 5), a car collided with pedestrians on the sidewalk and caught fire, causing injuries. According to The Press Democral (6), a vehicle drove onto the sidewalk and hit people.

Along with the monitoring and control system is more and more huge, huge amounts of security video data more and more, and the abnormal event detection in video images and the analysis of the causes of accidents, such as business needs people to analysis and processing, the need to observe many workers for a long time without stopping the surveillance video, task workload is huge. Rely purely on people to observe the surveillance videos will lead to poor and false detection in the monitoring, which will reduce safety and the practicality of the entire security system, because people cannot maintain a high degree of concentration for a long time and require rest [2]. Therefore, to enhance the monitoring ability of video surveillance, they should be made to emulate the human brain to recognize events and provide early warning and developed to become intelligent and reduce the occurrence of all kinds of public hazards. For this goal, a technology that can automatically analyze and find abnormal conditions from surveillance videos without relying heavily on manpower should be developed, that is, video automatic anomaly detection technology.

Video anomaly behavior detection and localization refers to automatically detecting and locating abnormal behavior by utilizing the difference between the representation of normal and abnormal behavior features. With the increasing demand for automated anomaly detection in various applications, many practical anomaly detection frameworks have been proposed. The GMM method has attracted great attention for background modeling techniques for detection and location in crowded scenes [7,8]. The GMM generates a model based on the Gaussian probability density function by calculating the intensity, mean, and variance parameters [9]. Sabokrou et al. [10] used the GMM model to detect and locate abnormal behaviors in crowded scenes. Leyva et al. [11] combined the GMM model , Markov chains, and Bag-of-Words for video anomaly detection. Lu et al. [12] merged the Markov random field and the GMM to detect abnormal behavior through the calculation of the confidence measure. Marsden et al. [13] blended support vector machine (SVM) and the GMM to classify abnormal behaviors. Optical flow is an important feature that describes the motion patterns of visual features (such as points, objects, and shapes) through the continuous observation of the environment [32]. Yuan et al. [33] proposed an abnormal event detection method based on statistical hypothesis testing, which identifies abnormal events as events with high scores. In [34], a new feature descriptor, that is, hybrid optical flow histogram, is proposed, using sparse reconstruction cost to detect abnormal behavior because sparse representation has high recognition rate and stability. Li et al. [35] proposed a novel motion feature descriptor, that is, the histogram of maximal optical flow projection, to detect abnormal events in crowded scenes, using SVM to classify abnormal frames. As can be seen from the above elaboration, surveillance video anomaly detection efficiently detects abnormal events from many videos, and it usually consists of three parts: foreground extraction and motion target detection, feature extraction, classification and anomaly behavior detection, each of which is described in more detail below.

Traditional motion target detection algorithms mainly include background subtraction method, frame difference, edge detection, and optical flow methods. With the development of deep learning in the field of object detection, object detection networks have been widely used in foreground extraction and anomaly behavior detection. YOLACT (You Only Look At CoefficienTs) is an object detection technique based on instance segmentation, which can output both the position and mask information of objects. This method can be used for video anomaly detection because it can accurately segment and locate abnormal objects. Methods based on YOLOACT have achieved good results in video anomaly detection and have high practical value. Xu et al. [3] proposed a real-time video anomaly detection method based on Region Proposal Network and YOLACT, which can efficiently detect abnormal behaviors in a scene. Object detection and instance segmentation are performed using YOLACT, and multi-level convolutional networks are used for feature extraction to achieve fast and accurate anomaly detection.

Feature extraction is crucial for anomaly behavior detection. The higher the discriminability of features between normal and abnormal behaviors, the higher the detection accuracy. Traditional hand-crafted features represent behavior using manually defined low-level visual features (e.g. LUV, Cascade Step Search, OP and HOG), such as using Histograms of Oriented Gradients (HOG) to represent human body shape and contour information in static images, using optical flow to describe the changes in pixel grayscale values between adjacent frames to represent motion information, and using trajectories to describe the trajectory of moving targets. Features extracted based on deep learning can automatically learn the distribution rules of data from massive datasets, extract more robust high-level semantic features, and are less sensitive to crowded scenes. Gradually, they have replaced traditional feature extraction techniques.

Anomaly detection techniques require training a classifier to detect behavior after extracting features. For a given specific scene with video data samples, the motion and appearance features of video frames or images within video windows are first extracted, and a model is built to learn the distribution of normal samples. During testing, the extracted features of the test sample are input into the model, and the model judges the sample as normal or abnormal anomaly score. Commonly used classifiers include SVM, Naive Bayes classifiers and decision tree.

Deep learning have shown remarkable potential in learning appearance representations from images. Methods based on deep learning are widely used in the field of computer vision, such as image detection and classification [14,15] and behavior detection [16–23]. Hu et al. [24] used the deep learning network framework to calculate the number of individuals in a population extracted from a picture. Shao et al. [25] presented the slicing CNN (S-CNN) for effectively extracting appearance and dynamic information in crowd video scenarios. Karen Simonyan and Andrew Zisserman [26] proposed a method based on two independent recognition streams; the spatial stream extracts action recognition frames from still videos, while the temporal stream is trained to recognize actions from motion in the form of dense optical flow and then fuses them. Feichtenhofer et al. [27] proposed a new spatio-temporal architecture for two-stream networks to spatially and temporally learn the correspondence between highly abstract deep features. Yi et al. [28] proposed Behavior-CNN) to model pedestrian behaviors in crowded scenes. Luo et al. [29] used the YOLOv3 target detection network to detect pedestrians holding sticks, guns, knives and facial shielding. The local resolution enhancement network (LDA-Net) proposed by Gong et al. [30] took the foreground human body extracted by YOLO network as the input of 3DCNN, so as to extract the spatio-temporal characteristics of behaviors and classify normal and abnormal behaviors. Zou [31] input the feature vector extracted from YOLO network into LSTM network. This structure of CNN and LSTM makes full use of the spatio-temporal fusion features of video, thus improving the recognition accuracy.

Given the success of deep neural network (DNN) in feature representation, the features extracted by a DNN represent the appearance and motion pattern in different scenes more specifically than the traditional anomaly detection approaches. The proposed hybrid method, which combines deep learning feature extraction with traditional foreground extraction and abnormal behavior detection, is effective in detecting and locating abnormal behaviors in crowded scenes. In this paper, we use the adaptive Gaussian mixture mode (AGMM) and to YOLOACT extract the foreground and the foreground masks as a preprocessing procedure of feature extraction. Then PWC-Net is applied to extract the motion information from the foreground map. Furthermore, we input the motion information to the classification network to output the anomaly score. Finally, this paper proposes a real-time object detection and tracking method using YOLOv5 and DeepSORT for surveillance videos. Specifically, after object detection is performed using YOLOv5, the results will contain the position and class information of each detected object. Then, these object information are fed into DeepSORT, which assigns a unique ID to each target and performs real-time tracking of the targets. In the subsequent video frames, DeepSORT can accurately predict the position of the targets by calculating their motion information and appearance features, achieving continuous tracking of the targets. The main contributions of this paper are as follows:

* This paper fine-tunes the YOLACT network and introduces a mask generation module to meet the requirements of the proposed method.
* This paper combines traditional methods with deep learning networks to extract foreground mask images from video frames, thereby enhancing the richness and accuracy of the foreground masks.
* In this paper, PWC-Net is used to extract foreground object features and these features are used to train the anomaly detection classifier, resulting in improved accuracy of anomaly detection classification.
* The method proposed in this paper consists of two stages. The first stage involves detecting and locating anomalies in video frames, while the second stage focuses on tracking the objects in the video frames to facilitate better understanding and analysis.
* This paper adopts a hybrid methods to detect and locate anomalous video frames. Specifically, we use deep features to construct the feature space instead of handcrafted features, and then use traditional machine learning methods to detect anomalies. By leveraging the strengths of these two approaches, we improve the performance of the method.

The remainder of this paper is organized as follows. Section 2 reviews the related works on anomaly detection and localization. Section 3 provides a detailed description of the proposed method. First, the framework of the proposed method is introduced. Then, the components of surveillance video anomaly detection are discussed. Section 4 presents the experimental results and comparisons. Finally, the conclusion is presented in Section 5.

2. Related works

2.1. Anomaly Detection Analysis

Video anomaly detection can be divided into local anomalies and global anomalies. Local anomalies usually refer to the activity of an individual that deviates significantly from its neighboring individuals in a moderately or densely crowded environment. Global anomalies refer to the overall abnormality in a specific scene, and the activities of individual locals may be normal. Apart from that, there are two more methods for anomaly detection. One is the abnormal appearance or motion attributes in videos, and the other is the normal appearance or motion attributes in abnormal time or space [36]. In terms of anomaly types, appearance anomaly refers to spatial anomaly, including local anomaly at the pixel level and global anomaly at the frame level. Motion anomaly refers to time anomaly, that is, the context anomaly related to time sequence. The task of video anomaly detection is to detect the temporal and spatial anomalies in videos [38]. Given the diversity of abnormal samples, the video anomaly detection method models the distribution of normal and abnormal samples and the trained model is used to distinguish the different properties of abnormal samples and normal samples to detect abnormal samples in the test [37]. Given that the background of surveillance videos in a specific scene is often fixed, the video frequency of surveillance is a typical single-scene video. The research on video anomaly detection based on a single scene is the focus of this paper.

2.2. The Learning Paradigms for Video Anomaly Detection

Video anomaly detection has four main learning paradigms, namely, supervised, unsupervised, weakly supervised, and self-supervised. This paper mainly introduced supervised learning. Supervised learning refers to the process of mapping all data samples to labels of different categories through model training, given data samples and their corresponding labels. Video anomaly detection uses normal samples, abnormal samples, and their corresponding labels to train a binary classifier for anomaly detection. There are some supervised anomaly detection methods that have been proposed. Zhou et al. [39] proposed a spatial-temporal convolutional neural network (ST-CNN) for anomaly detection and localization in crowded scenes. The network uses both spatial and temporal information to extract features and classify the scenes as normal or abnormal. SABOKROU et al. [40] proposed a approach consists of multiple 3D convolutional neural networks (CNNs) cascaded to handle different scales of input data. Each cascaded network is trained on a specific subset of the data to optimize its performance. Miao et al. [41] proposed an abnormal event detection method based on Support Vector Machines (SVM) in video surveillance. The method first extracts motion features from video frames using background subtraction and optical flow, and then trains an SVM classifier to distinguish between normal and abnormal events. Many methods based on supervised learning perform well in video anomaly detection, so we present a supervised learning anomaly detection method.

2.3. Object Detection

In most cases, the abnormal situations in surveillance videos are usually moving objects or targets. However, the large area of background or stationary objects in the video makes the abnormal detection process become complex and computationally expensive. Additionally, a large amount of noise and redundant information makes feature extraction and behavior representation difficult, thereby greatly reducing the efficiency and quality of abnormal detection. Therefore, motion object detection is an indispensable step in intelligent abnormal detection systems. Traditional motion object detection methods include frame difference method, background subtraction method, and optical flow method. Background subtraction [44] is a widely used method for moving object detection in videos The basic principle of this method is to detect moving objects using the difference between the current and reference frames, that is, the background image or model. Several common background subtraction methods, such as mean and median filtering, bimodal backgrounds, long-term scene changes, and adaptive Gaussian mixture model, can be used. We select the most common method, that is, the adaptive Gaussian mixture model (AGMM). The adaptive Gaussian mixture method (AGMM) [45] and [46] can effectively deal with many component models. The use of many models is recommended because the range error of individual components decreases with the addition of models, thereby decreasing the net range of background values. This technique selects a suitable number of Gaussian distributions for each pixel, allowing preferable adaptation to scene changes and improving the robustness due to changes in brightness.

With the extension of deep learning in the field of video object extraction, target extraction technology has become more efficient and accurate, with a wider range of application prospects. YOLACT based on deep learning is one of the most commonly used real-time object detection and instance segmentation techniques, combines target detection and instance segmentation by using interactive convolutional networks (Interact Convolutional Networks). Specifically, YOLACT uses a loss function called Mask-IoU, which optimizes the performance of both target detection and instance segmentation. Mask-IoU loss uses intersection over union (IoU) to measure the model's performance, combines the results of target detection and instance segmentation, and minimizes the difference between the two. In addition, YOLACT uses a feature pyramid network to process feature maps of different scales, improving the model's ability to detect and segment objects of different scales. Overall, YOLACT achieves efficient object detection and instance segmentation by combining target detection and instance segmentation and using techniques such as Mask-IoU loss and feature pyramid network. Although advanced features have good features, low-level features often have advantages that high-level features do not have, such as invariance under illumination changes.

By leveraging the strengths of AGMM for effective foreground and background extraction and YOLACT for precise object detection and segmentation, this paper provides a robust and accurate solution that can handle complex scenes better than using either method alone.

2.4. Feature Extraction

Efficient extraction of appropriate features plays a crucial role in rapid and accurate discrimination of normal and abnormal behavior in video anomaly detection research. Researchers have proposed various methods for feature extraction and behavior representation. Manually designed features commonly used for video anomaly detection include texture features, color, MoSIFT (Motion Scale Invariant Feature Transform), optical flow features, trajectory features, and so on. In practice, optical flow features can be used in a wide range of applications. Optical flow is caused by the movement of the foreground object, motion of the camera, or both in a scene. Optical flow [47] is defined as the apparent motion of a single pixel on the image plane and can calculate the motion information of objects between adjacent frames. This motion usually serves as a good approximation of the real physical motion projected onto the image plane. Optical flow utilizes the change of pixels in the temporal domain and the correlation between adjacent frames to detect the corresponding relationship between the previous and current frames [48]. Several optical flow methods exist, such as the Horn–Schunck, pyramid Lucas–Kanade, and Gunnar Farneback techniques [53]. Bullinger et al. explored a different approach in [49]and reported that optical-flow-based tracking techniques perform well, especially when the target position in the image is also subject to camera motion in addition to the object's own motion. After obtaining the segmentation masks for the various instances present in the current frame, an optical flow method [50–52] is applied to predict the position and shape of each instance in the next frame.

Although hand-designed features extracted by manual design have many theoretical justifications, they are too subjective to objectively represent behavior. Additionally, features extracted in this way often rely on databases, meaning that hand-designed features may only perform well on certain databases and may not necessarily produce the same results on other databases. Features extracted by deep learning can automatically learn the distribution rules of data from massive datasets, extract more robust and high-level semantic features, and are insensitive to crowded scenes. As a result, deep learning-based feature extraction methods are gradually replacing traditional feature extraction algorithms. Deep learning-based method for motion estimation: using deep learning networks to extract image features, and then estimating object motion based on the changes in the features, representative methods include FlowNet and PWC-Net [54]. PWC-Net has been shown to outperform previous state-of-the-art methods on several optical flow benchmarks, and is widely used in computer vision applications such as video analysis. PWC-Net [4] is a deep learning network used for optical flow estimation and can be used for video frame anomaly detection. The structure of PWC-Net is simple and lightweight, making it suitable for real-time anomaly detection scenarios. By comparing the predicted flow with the actual flow, abnormal video frames can be detected. Wu et al. [5] proposed a real-time video anomaly detection method using PWC-Net and frame difference. The PWC-Net is used for optical flow estimation, and the frame difference is used to capture the temporal changes of the video. The proposed method achieves real-time performance and outperforms several state-of-the-art methods in terms of anomaly detection accuracy. Peng et al. [6] introduced a novel weighted loss function that considers both reconstruction error and motion difference to enhance the feature representation of normal frames. They also propose a two-stage detection method that combines frame-level and pixel-level anomaly scores to improve the detection accuracy.

In this paper, we used PWC-Net to extract optical flow fields between video frames and used these fields to represent motion information. Then, we fed these optical flow features into our proposed anomaly detection model for training and prediction. Through this method, we can effectively utilize the technology of deep learning to improve the accuracy and efficiency of video anomaly detection.

2.5. Detection Based Tracking

With the rapid development of deep learning, the object detection performance has been greatly improved, and the detection-based tracking (DBT) scheme was developed. It has quickly become the mainstream framework and greatly promotes the progress of the video behavior detection task. Meanwhile, the joint framework based on detection and tracking has attracted the attention of researchers. At present, the most widely used DBT techniques with good performance are YOLOv5 and DeepSORT.

YOLOv5 is a deep learning technique used for object detection, which can quickly detect the position and category of multiple objects in an image. The working principle of YOLOv5 is to extract features from the input image through a convolutional neural network, and then use anchor boxes to predict the pixels on the feature map to determine the location and category of objects. DeepSORT [42] is a commonly used method in multi-object tracking (MOT) [43], which can track objects in a video and generate the object's motion trajectory. The DeepSORT method is an improvement of the simple online and real-time tracking (SORT) method. Using Kalman filtering and the Hungarian method, SORT greatly improves the speed and accuracy of multi-target tracking. The Kalman filter technique is divided into two processes: prediction and update. In this method, the motion state of the target is defined as eight normally distributed vectors. In the prediction process, when the target moves, the parameters of the previous frame are used to predict the position and speed of the current frame. In the update process, to obtain the results predicted by the current model, the two normally distributed states of the predicted value and the observed value are linearly weighted. Finally, trajectory prediction is realized. In the MOT step, the similarity is calculated to obtain the similarity matrix of the current and reference frames. By solving the similarity matrix, the tracks predicted by the Kalman filter is matched with those detected in the current frame, the Hungarian method finally achieves the matching of the current and reference frames and solves the allocation problem. The most special characteristics of DeepSORT are the addition of appearance information, the application of the ReID domain model (an appearance model) to extract features, and the reduction of the number of ID switches.

YOLOv5 and DeepSORT can work in collaboration, with YOLOv5 used to detect objects in the video and output their position and category, and DeepSORT used to track these objects and generate their motion trajectory using the output from YOLOv5. In this way, we can better understand the behavior and relationship of objects in the video.

3. Proposed Method

In this paper, we take into account the appearance and dynamics of video surveillance scenes, as well as their spatial and temporal characteristics. Furthermore, deep features have stronger descriptive abilities compared to handcrafted features. Thus, in our proposed hybrid method that combines traditional and deep learning methods, we use deep features to replace handcrafted features and employ traditional machine learning methods to detect anomalies.

Usually, anomalies refer to an individual's activity that significantly deviates from its neighboring individuals in a densely crowded environment, or to a specific abnormal situation in the overall scene, while the activity of local individuals may be normal. Therefore, the identification of abnormal appearance and movement patterns is the key problem of anomaly detection. Supervised-learning-based feature extraction methods have been successful in other tasks. The proposed supervised method involves judging video abnormal behavior based on video frame features and classifiers to model normal and abnormal patterns, which includes the scene background, appearance, and motion of normal and abnormal activities. The flowchart of the proposed detection method is presented in Figure 2.

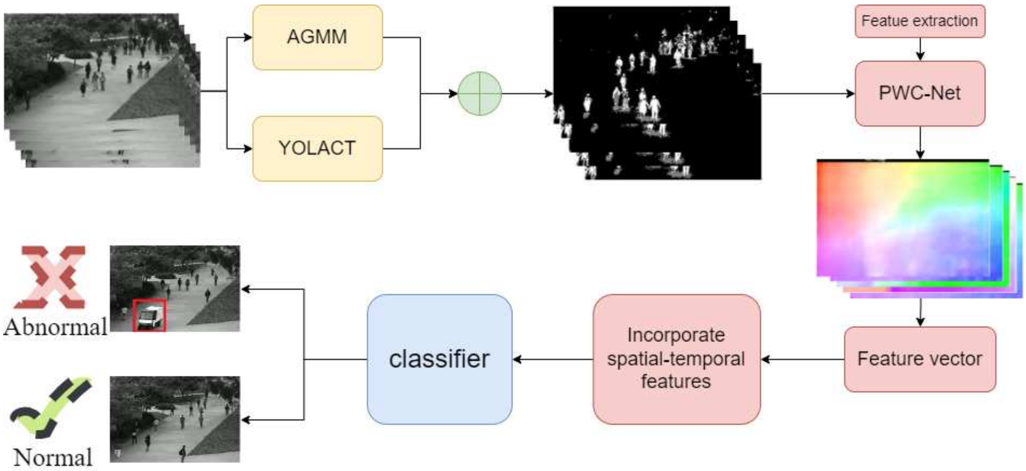

Figure 2. The flowchart of abnormal behavior detection and localization.

3.1. Object Detection

Traditional handcrafted features based on manually defined low-level visual features cannot represent complex behaviors and the extracted features are relatively simple, resulting in weak generalization ability. On the other hand, deep learning-based feature extraction can automatically learn the distribution rules of massive datasets and extract more robust high-level semantic features to better represent complex behaviors. However, low-level features often have some advantages that high-level features do not have, such as invariance under lighting changes, strong interpretability, and the ability to provide basic spatial, temporal, and frequency information. Therefore, this paper proposes to combine traditional background subtraction methods and the YOLACT network to detect and segment foreground images in video frames. Specifically, combining AGMM (Adaptive Gaussian Mixture Model) and YOLACT for video frame object extraction can enhance the accuracy and efficiency of the process. AGMM is a background subtraction method that models the background by a mixture of Gaussian distributions, and it adapts to the scene changes by adjusting the parameters of the distributions. AGMM can handle different types of scenes and achieve good results in complex environments with moving backgrounds. YOLACT, on the other hand, is a deep learning-based object detection and instance segmentation technique that combines target detection and instance segmentation using interactive convolutional networks (Interact Convolutional Networks) and can accurately detect and segment objects in real-time. By combining AGMM and YOLACT, we can effectively extract the background and foreground, which can not only use traditional methods to extract low-level features but also use deep learning to extract high-level semantic features, thus improving the accuracy and robustness of video anomaly detection. This method can effectively handle complex scenes and achieve better results than using only one method.

Overall, the combination of AGMM and YOLACT can provide a more robust and accurate solution for video frame object extraction. The specific implementation is as follows:

1. The YOLACT technique was used for object detection and instance segmentation to obtain the foreground map F_1.

YOLACT is a real-time object detection and instance segmentation technique based on deep learning techniques. Its main principle is to combine object detection and instance segmentation together by using Interact Convolutional Networks. Specifically, YOLACT uses a loss function called Mask-IoU loss, which optimizes the performance of both object detection and instance segmentation.

The Mask-IoU loss measures the model's performance using the Intersection over Union (IoU) metric, combining the results of object detection and instance segmentation and minimizing the difference between them. Additionally, YOLACT also uses a Feature Pyramid Network to process feature maps of different scales, improving the model's detection and segmentation capabilities for objects of different sizes. In summary, YOLACT achieves efficient object detection and instance segmentation by combining object detection and instance segmentation, using Mask-IoU loss and Feature Pyramid Network, among other techniques. The result of the YOLACT is shown in Figure 3.

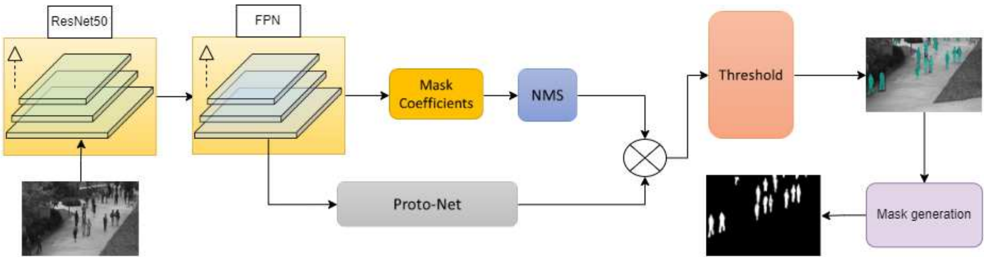

Figure 3. Flowchart of the improved YOLACT architecture.

From Figure 3, it can be seen that YOLACT generates foreground images, while what we need is a foreground mask similar to AGMM. By having a similar data type distribution for the foreground masks generated by YOLACT and AGMM, it is easier to fuse them together. Therefore, this paper proposes an improvement to the YOLACT network by adding a mask generation module. The YOLACT foreground map is a gray image where each pixel has a single value indicating whether the pixel belongs to the foreground object or not. In the YOLACT technique, a set of feature maps are obtained by performing convolution and feature extraction on the input image, and then a Mask Head network is used to process each feature map to obtain the corresponding foreground mask, which is the YOLACT foreground map. During the process of generating the foreground mask, each pixel is thresholded to classify it as foreground or background. Let the input image be I, the feature map extracted by the YOLACT model, and the foreground mask be M. For each position (x, y) in the feature map, the generation of the foreground mask can be expressed by the following formula:

$$M(x,y) = \begin{cases} 255, & \text{if } Q(x,y) > H; \\ 0, & \text{Otherwise.} \end{cases}. \tag{1}$$

Here, $Q(x, y)$ represents the foreground probability of the pixel corresponding to the position (x, y) in the feature map, and H is a preset threshold. Specifically, for each position (x, y) in the feature map, we can calculate its foreground probability $Q(x, y)$ using the Mask Head network:

$$Q(x,y) = \sigma(w^T f(x,y)). \tag{2}$$

Here, w is the parameter of the Mask Head network, $f(x, y)$ represents the feature vector corresponding to the position (x, y) in the feature map, and σ represents the sigmoid function. Finally, by setting the threshold H, we can convert the foreground probability $P(x, y)$ into a gray foreground mask value $M(x, y)$, thereby obtaining the YOLACT foreground map, the outcome of improved YOLACT is presented in Figure 3.

In this paper, the YOLACT network architecture is used for foreground extraction from video frames. The model is trained on the COCO dataset for 800,000 epochs with a 54-layer convolutional neural network.

2. AGMM is used to process the video frame and get the foreground map F_2, the foreground mask extraction process using AGMM is illustrated in Figure 4.

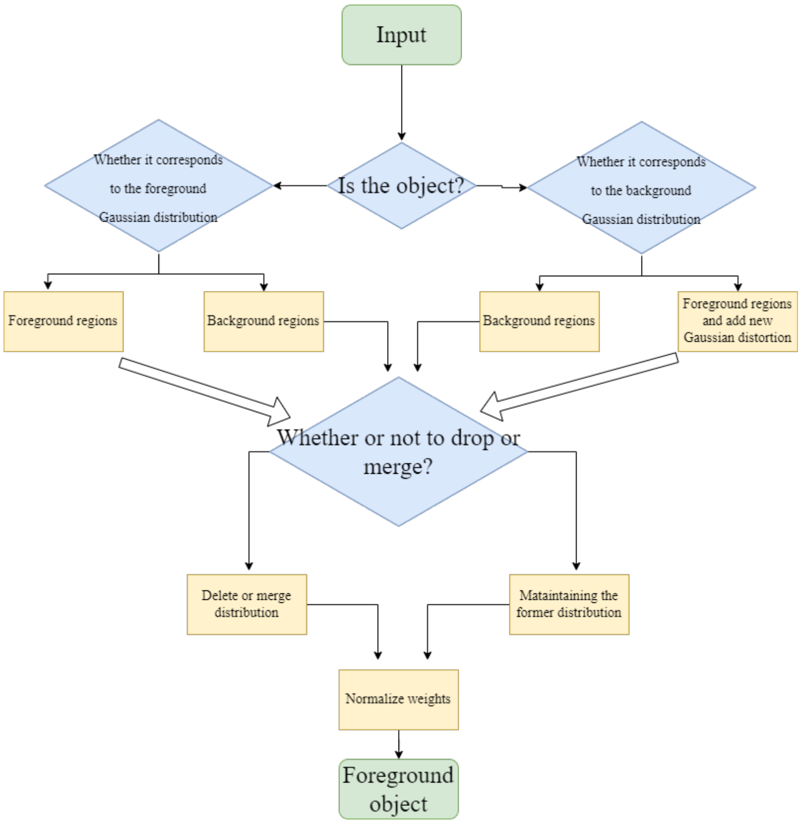
Input
Is the object?
Whether it corresponds to the foreground Gaussian distribution
Whether it corresponds to the background Gaussian distribution
Foreground regions
Background regions
Background regions
Foreground regions and add new Gaussian distortion
Whether or not to drop or merge?
Delete or merge distribution
Mataintaining the former distribution
Normalize weights
Foreground object

Figure 4. Foreground mask extraction process via AGMM.

Background subtraction is one of the main components of surveillance video behavior detection and used as the preprocessing of object classification in this paper. The background subtraction method based on AGMM [55], which has a good antijamming capability, especially the illumination change, is adopted. Thus, we select the AGMM background subtraction scheme to extract foreground images. A suitable time period T, and time t are assumed. x^t represents a sample, X^T=(x^t, x^{t-1}, ... , x^{t-T}). For each new data sample, we both update the model X^T and re-estimate the density $p(x_n|X^T, BG)$. These samples may contain values for the background (BG) and foreground (FG) objects, thus the density estimation $p(x_n^t|X^T, FG + BG)$. The Gaussian mixture model with K components is expressed as follows:

$$p(x^t|X^T, FG + BG) = \sum_{k=1}^{K} f_k \bullet G(x^t, \mu_k, \sigma_k^2 M). \tag{3}$$

$$G(x^t, \mu_k, \sigma_k^2 M) = \sum_{j}^{K} \frac{f_{j,t}}{(2\pi)^{d/2}|\sigma_{j,t}M|^{\frac{1}{2}}} e^{-\frac{1}{2}(x^t-\mu_{j,t})\sum_{j,t}^{-}1(x^t-\mu_{j,t})}. \tag{4}$$

Here f_k represents non-negative estimated mixing weights, and the k_{th} GMM is normalized at time t. μ_k and σ_k are the estimated mean value and variances of the Gaussain components, respectively. M is an identity matrix. Given a new data sample x^t at time t, the recursive update equations are as follows:

$$f_k = (1-\alpha)f_k + \alpha(o_m^t + c_T). \tag{5}$$

$$\mu_k = \mu_k + o_m^t(\alpha/f_k)\delta_k. \tag{6}$$

$$\sigma_m^2 = \sigma_m^2 + o_m^t(\alpha/f_k)(\delta_k^T\delta_k - \sigma_m^2). \tag{7}$$

Here $\delta_k = x^t - \mu_k$, α is the learning rate and the value of $1/T$ is set. For a new sample, the ownership o_m^t is set to 1 for the "close" component with the largest f_k and the others are set to zero. We define that a sample belongs to a component if its Mahalanobis distance from the component is less than a certain threshold, and the squared distance from the k_{th} component is calculated as follows:

$$D = \frac{\delta_k^T \delta_k}{\sigma_m^2}. \tag{8}$$

In turn, the algorithm will generate a new component $f_{k+1} = \alpha$, $\mu_{k+1} = x^t$ and $\sigma_{m+1} = \sigma_0$, here σ_0 is a initial value. This method shortens the processing time and improves the segmentation while providing highly specific image features for the next step of object detection. AGMM (Adaptive Gaussian Mixture Model) is a foreground extraction method based on Gaussian mixture model.

For each pixel at location (i, j), calculate the Mahalanobis distance (D) between the pixel's color and the mean color of each Gaussian component in the mixture model:

$$D(k) = ||X(i, j) - \mu(k)||^2 / \sigma(k). \tag{9}$$

where $X(i, j)$ is the pixel, $\mu(k)$ is the mean color of the k_{th} Gaussian component, and $\sigma(k)$ is the standard deviation of the k_{th} Gaussian component. Then, the pixel is classified as foreground if the minimum Mahalanobis distance is smaller than a threshold:

$$min_k(D(k)) < Th. \tag{10}$$

where Th is the threshold. Otherwise, the pixel is classified as background.

After initializing the model parameters, for each frame of the image, the foreground probability of each pixel is first calculated based on the current model parameters, and then the foreground and background pixels are segmented according to a threshold to obtain the foreground map. The foreground mask is detected clearly for the UCSD ped1 [56] benchmark dataset using AGMM.

3. For the fusion of two foreground graphs, weighted average can be adopted, as shown in the formula. The weight of the YOLACT foreground can be set to a larger value and the weight of the background subtracting foreground can be set to a smaller value to preserve the details of the YOLACT foreground.

$$F_t = w_1 \times F_1 + w_2 \times F_2. \tag{11}$$

where $w_1 = 0.6$, $w_2 = 0.4$, and F_t is the fused foreground map.

4. Finally, YOLACT technique can be used again for target detection and instance segmentation for the fused foreground, so as to further improve the accuracy and robustness of target detection and segmentation.

3.2. Feature Learning

After extracting the foreground masks from the video frames, the next step is to perform feature extraction on the foreground images. These features will be used to classify the frames as either normal or abnormal, providing valuable information for anomaly detection analysis. In video anomaly detection, the efficient detection of appropriate features plays an important role in the rapid and accurate identification of normal and abnormal behaviors. Feature extraction can performed in two ways. One is to manually extract features, and the other is to learn the original video frames to obtain deep features. Manual feature extraction methods, despite having numerous theoretical justifications, are often influenced by human factors and may not objectively represent behavior. Moreover, features extracted through this method often depend on the database, meaning that manually designed features may only perform well for certain databases and may not yield the same results for others. Traditional video behavior detection technology has low accuracy, and shallow learning cannot parse big data.

In contrast, deep learning can overcome these problems well. Deep feature extraction through direct learning from data requires only the design of feature extraction rules, the manual design of the network structure, and learning rules to obtain deep model parameters and extract deep features, thus improving the recognition accuracy and the robustness in the process of video behavior detection. The PWC-Net based on deep learning is a kind of optical flow estimation technology used to calculate motion information between adjacent image frames. When combined with classification methods, which can be used for image anomaly detection tasks. PWC-Net (Pyramid, Warping, and Cost Volume, with multi-scale and multi-stage architecture) is a state-of-the-art method for optical flow estimation, proposed by Sun et al. in 2018. It builds upon the FlowNet architecture and introduces several improvements, including: 1). Multi-scale processing: the input images are processed at multiple scales to better capture small-scale and large-scale motion. 2). Multi-stage processing: the network has multiple stages, each of which takes as input the output of the previous stage, allowing for a more fine-grained estimation of flow. 3). Cost volume: instead of concatenating the feature maps of the two input images, PWC-Net computes a cost volume by computing the dot product between each pair of feature vectors. This allows the network to compute the cost of different flow hypotheses, which is then used to estimate the final flow field.

The optical flow vectors can be used as features to construct a classifier, which is used to classify input image frames into normal and abnormal categories. The results of PWC-Net is shown in Figure 5. During the training of the classifier, image frames with typical motion patterns can be used as normal samples, while image frames with atypical motion patterns can be used as abnormal samples. Then, the trained classifier can be used to classify new input image frames and detect whether anomalies are present.

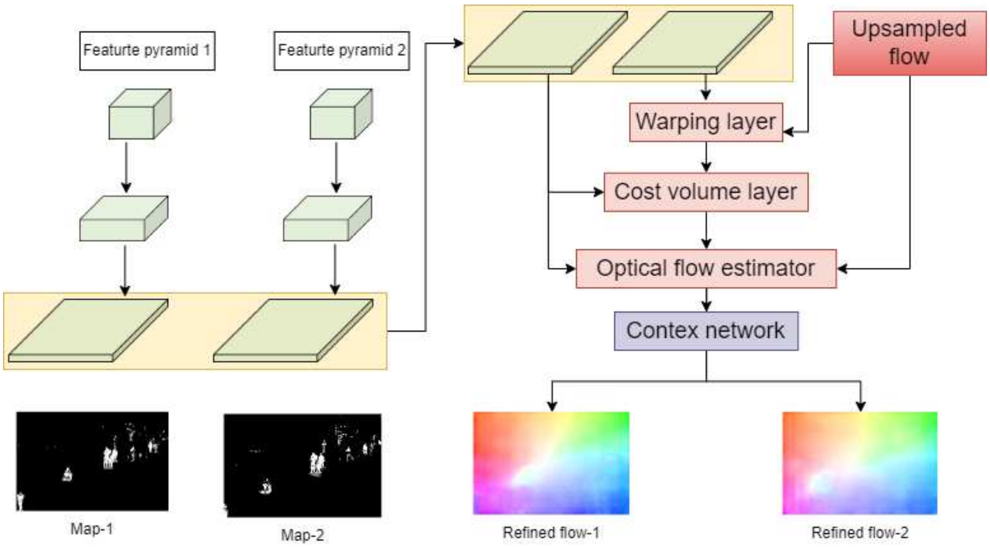

Figure 5. Flowchart of PWC-Net architecture.

3.3. *Abnormal Behavior Detection*

Given a new visual sequence, we select abnormal video frames by adopting the trained classifier model to compute images in the video. And we estimate the likelihood of $P(x, y, t)$ by verifying the validity of the corresponding spatio-temporal video.

$$P(x, y, t) = P(I_t, D_t, I_x, D_y). \tag{12}$$

where I_x is known feature vector, D_y is the location. I_t denotes the feature vector of the observed objection t, and D_t denotes its location. To model the relationship between x, z from the aspect of spatio-temporal appearance, we estimate the conditional probability $C(I_t|I_x)$ by the cosine similarity between I_t and I_x:

$$C(I_t|I_x) = \frac{\sum_{i=1}^{M} I_{xi} I_{ti}}{\sqrt{(\sum_{i=1}^{M} I_{xi}^2)(\sum_{k=i}^{M} I_{ti}^2)}}. \tag{13}$$

The location similarity between y and t is modeled using a Gaussian function:

$$G(D_t|D_y) = \alpha \cdot exp(-\frac{1}{2}(D_t - D_y)^T \times (D_z - D_y)). \tag{14}$$

Dimension of the feature vector is M, I_{ti} is the i_{th} element of t feature vector, exp denotes natural exponential function, and α is a constant. Assuming that the variables x and y are conditionally independent and have a uniform prior distribution, it suggests that there is no prior preference for valid samples in the set. Consequently, the joint likelihood of the observed object t and the hidden variables x and y can be factorized in the following way:

$$P(x, y, t) = \gamma \cdot C(I_t|I_x) \cdot G(D_t|D_y). \tag{15}$$

The constant γ is used to ensure that the maximum value of $P(x, y, t)$ is limited to a value smaller than 1. We aim to find the samples x and y that maximize the maximum a posteriori probability assignment. This can be achieved by using equation (15):

$$max\ P(x, y, t) = [max\ C(I_t|I_x)]_1 \cdot [max\ G(D_t|D_y)]_2. \tag{16}$$

The first term in equation (16) represents the max inference of spatio-temporal appearance, while the second term represents the max inference of spatial location. Based on the max inference, a sample that appeared only once in the normal fixation set is equally likely as samples that appeared multiple times. A large likelihood indicates that it is more likely to find x and y in the set to infer the anomaly object t in terms of spatio-temporal appearance and spatial location.

3.4. Tracking

In addition to anomaly detection, this paper also demonstrated the use of YOLOv5 and DeepSORT for object detection and tracking in videos. In this paper, firstly, YOLOv5 is used to detect the position and class information of each target in each frame of the video. When the target is detected using YOLOv5, its position and class information can be passed to DeepSORT, The DeepSORT parameters are listed in Table 1. DeepSORT uses a method based on appearance and motion information to determine the identity of the target, using Kalman filtering and the Hungarian technology for multi-object tracking, thereby obtaining the trajectory information of the target in the video. It compares the currently detected targets with the previously tracked targets to determine whether they are the same target and creates new target tracking when necessary. By combining YOLOv5 and DeepSORT, this paper achieves accurate detection and tracking of targets in videos, providing target position and ID information. The results of the object detection and tracking using YOLOv5 and DeepSORT are illustrated in Figure 14 and Figure 15.

Table 1. DeepSORT parameters setting.

DeepSORT parameters	Weight
MAX-DIST	0.2
MIN-CONFIDENCE	0.3
NMS-MAX-OVERLAP	0.5
MAX-IOU-DISTANCE	0.7
MAX-AGE	70
N-INIT	3
NN-BUDGET	100

In a word, we propose a method for detecting and locating abnormal behavior in a monitoring scene. Firstly, combining the AGMM and YOLACT technologys to obtain more accurate foreground information. Then we use the PWC-Net technology to extract features of the foreground images and input them into an anomaly classification model for classification, which further improves the accuracy of our method. Additionally, we employ YOLOv5 and DeepSORT networks for object detection and tracking in the video, allowing us to detect different objects present in the video and track them for better understanding of the scene. Overall, our proposed method effectively detects and locates abnormal behavior in the monitoring scene.

4. Experiments

The proposed method is evaluated on the UCSD dataset [57], which means that it has labels and ground truth information for the task of abnormal detection and localization in crowded scenes, the specific experimental analysis is outlined as follows.

4.1. Experimental Basis

We test the performance of the proposed method on the UCSD anomaly detection dataset. The dataset was obtained using a fixed camera mounted at a certain height overlooking the walkway. The UCSD dataset contains two subsets, Ped1 and Ped2. Both subsets contain the training and test sets. Ped1 contains 34 training and 36 test videos, with a frame resolution of 238 × 158 pixels. Each video sequence has a frame length of 200. Ped2 contains 16 training videos and 12 test videos of pedestrian motion videos on the sidewalk with a frame resolution of 360 × 240 pixels. The frame length of each sequence varies from 120 to 170. Moreover, all the frames in the training set are normal frames, containing only pedestrians. In addition to normal objects, the testing set also contains abnormal objects. The normal object in each frame is the pedestrian walking, and the rest of the behavior is regarded as abnormal behavior, such as cyclists, skateboarders, and cars. The UCSD dataset is a challenging local anomaly dataset in the crowded field. The low resolution of the objects in Ped1 makes them difficult to identify, while the occlusion problem in Ped2 is serious. All the test videos are used in this paper to evaluate the scheme. The sample images from the UCSD dataset are shown in Figure 6.

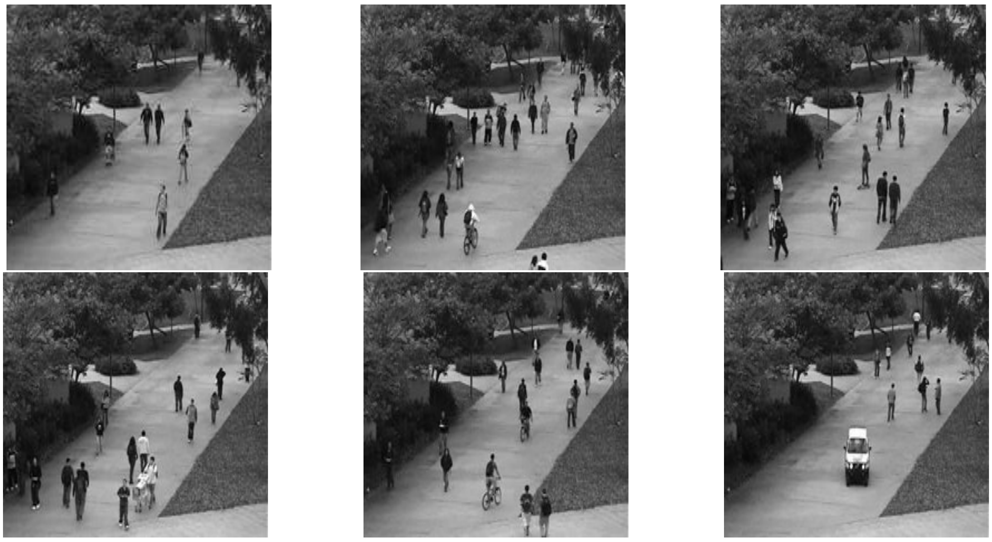

Figure 6. Examples of Ped1 dataset on UCSD.

For the anomaly detection of the video behavior on the UCSD dataset, we select the commonly accepted criteria for the evaluation of abnormal detection, including the equal error rate (EER) and the area under curve (AUC). Notably, the lower the EER and the higher the AUC are, the better the performance is. The two criteria are derived from the receiver operating characteristic curve (ROC). The ROC is composed of the true positive rate (TPR) and the false positive rate (FPR), where TPR refers to the correctly classified positive samples among all the positive samples during the test period, and FPR defines the number of false positive results among all the negative samples during the test period. TPR and FPR are as shown expressed as follows:

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{FP + TN}. \tag{17}$$

where true positive (TP) is the correctly labeled abnormal samples, true negative (TN) is the correctly labeled abnormal samples, false negative (FN) is the incorrectly labeled normal samples, and false positive (FP) is the incorrectly labeled abnormal samples. In this paper, we calculated the TPR and the FPR for generating the ROC curves.

4.2. *Moving Target Detection and Analysis*

In this paper, the Adaptive Gaussian Mixture Model (AGMM) and YOLACT are combined to extract video targets. To make AGMM more suitable for the scenario in this paper, we select appropriate experimental parameters. We mainly consider the following AGMM parameters for foreground extraction: detect shadows (DS), frames (F), learning rate (LR), threshold (T), the number of distributions (K), frame size (FZ), frame per second (FPS), and initial variance (IV). These parameters are selected to generate a relatively predictable mask for the further processing of object feature detection. For this paper, we perform appropriate adjustments and finally obtain the optimization result. To achieve the goal, the values of DS, F, LR, T, K, FZ, FPS, and IV are set to $Ture$, 200 frames, 0.005, 25, 5, 238 × 158, 30, and 15, respectively. The specific experimental results are presented in the third and fourth rows of the Figure 8.

In addition to traditional foreground extraction methods, this paper also employs deep learning networks to extract foreground images. The advantage of using YOLACT to extract foreground images from video frames lies in its strong real-time performance, which enables almost real-time foreground extraction. Moreover, YOLACT has high robustness and can adapt to different scenes and complex backgrounds. Using YOLACT for foreground extraction from video frames can produce high-quality foreground images, thus improving the accuracy and efficiency of subsequent processing. Based on the

doi:10.20944/preprints202305.1617.v3

ideas of Mask-RCNN and FCN, YOLACT achieves foreground image extraction through multi-level feature fusion and segmentation. Compared to traditional threshold-based segmentation methods, YOLACT performs better in handling complex scenes and motion blur, resulting in higher quality foreground images. The specific experimental results are presented in the fifth and sixth rows of the Figure 8. This paper fine-tunes the YOLACT network and introduces a mask generation module to meet the requirements of the proposed method. Due to the need for obtaining the foreground mask of grayscale images in this paper, a foreground mask generation module was added to the YOLACT network to meet the requirements of the technique. Additionally, Otsu foreground extraction and adaptive threshold methods were attempted to extract the foreground image generated by YOLACT, but the results were not satisfactory (as shown in the Figure 7). Therefore, a new foreground mask generation module was designed in this paper to generate more ideal mask images based on image features. The specific experimental results are presented in the seventh and eighth rows of the Figure 8.

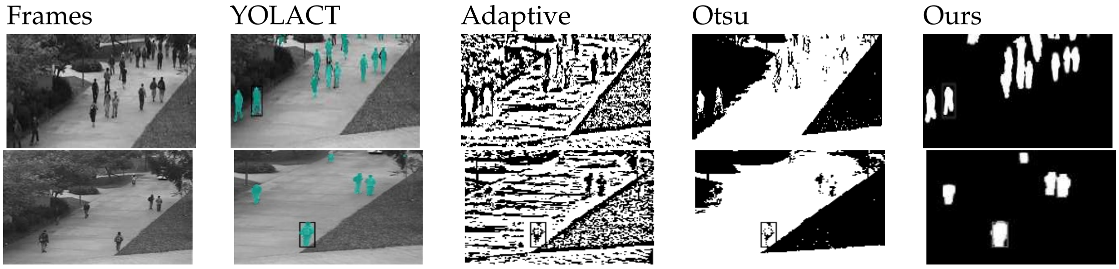

Figure 7. Foreground mask results display.

Why do we need to fuse these two types of features? Let's take a look at the content shown in Figure 9, both methods have their own limitations. The proposed method in this paper aims to fuse AGMM and YOLACT to obtain a more precise and comprehensive foreground map. From Figure 8 and Figure 9, it can be observed that the AGMM technique has robustness to changes in illumination and can effectively handle noise caused by such changes. Moreover, it performs very well for stationary objects or scenes with minimal changes. YOLACT technique utilizes deep learning techniques to achieve more accurate object detection and segmentation, resulting in high-quality foreground images with high precision. As shown in Figure 9, YOLACT can extract more detailed foreground information during object detection, which enhances its accuracy in foreground extraction. Meanwhile, it can be observed from Figure 8 that YOLACT has a lower recognition rate and may not even recognize low-light targets. However, although AGMM extracts a relatively rough foreground, it can identify every minor change in the target. Therefore, this paper proposes a fusion of target images extracted by YOLACT and AGMM to obtain a rich foreground image with high quality and accuracy. The specific experimental results are presented in the ninth and tenth rows of the Figure 8.

Experimental results demonstrate that the proposed method not only enhances the accuracy and completeness of foreground extraction, but also improves the processing speed and efficiency while ensuring real-time performance. In summary, the AGMM and YOLACT fusion method proposed in this paper leverages the strengths of both methods, improves the accuracy and efficiency of foreground extraction, and provides precise target features for video surveillance anomaly detection and localization.

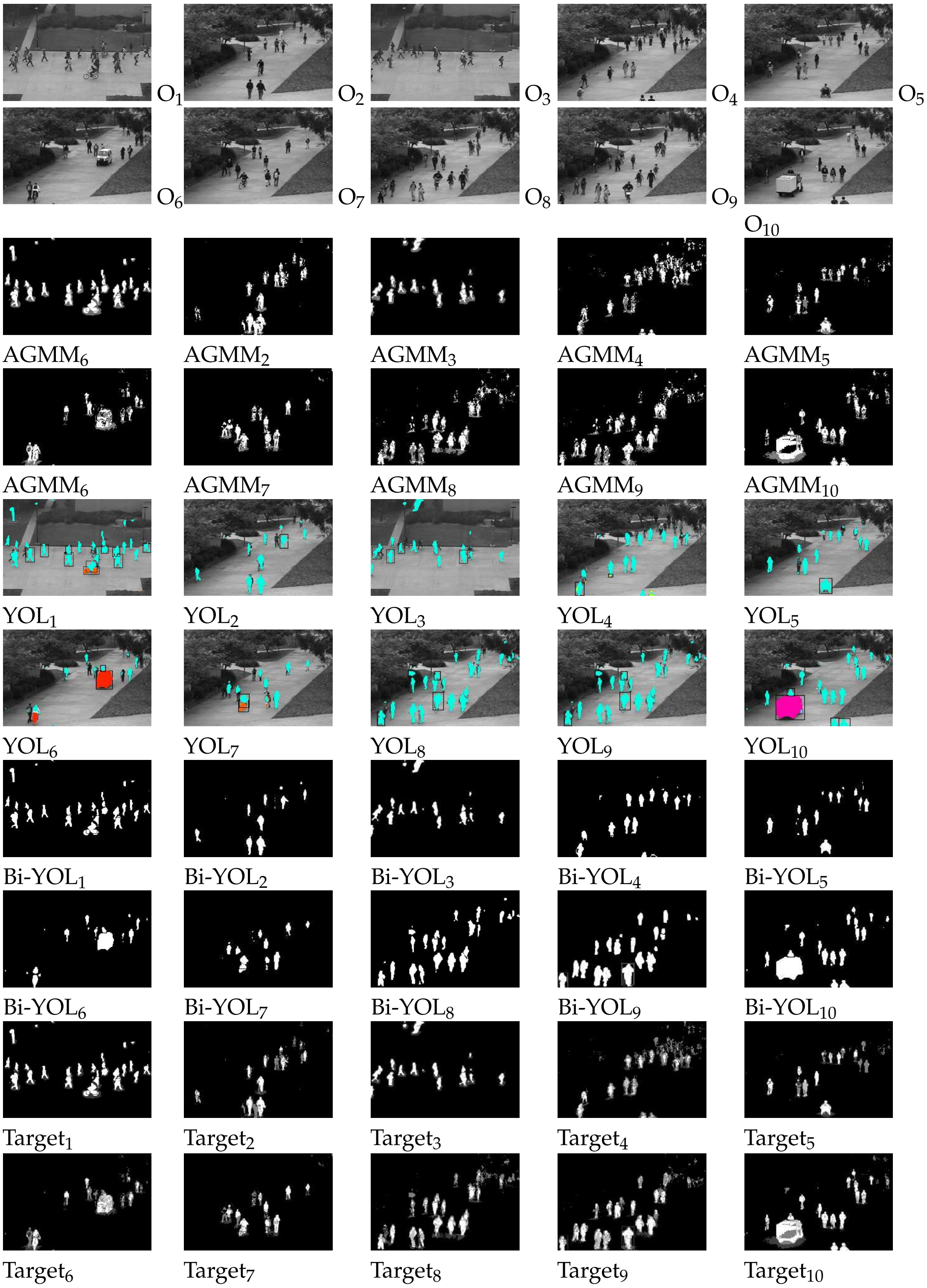

Figure 8. Display of the extracted foreground maps from video frames.

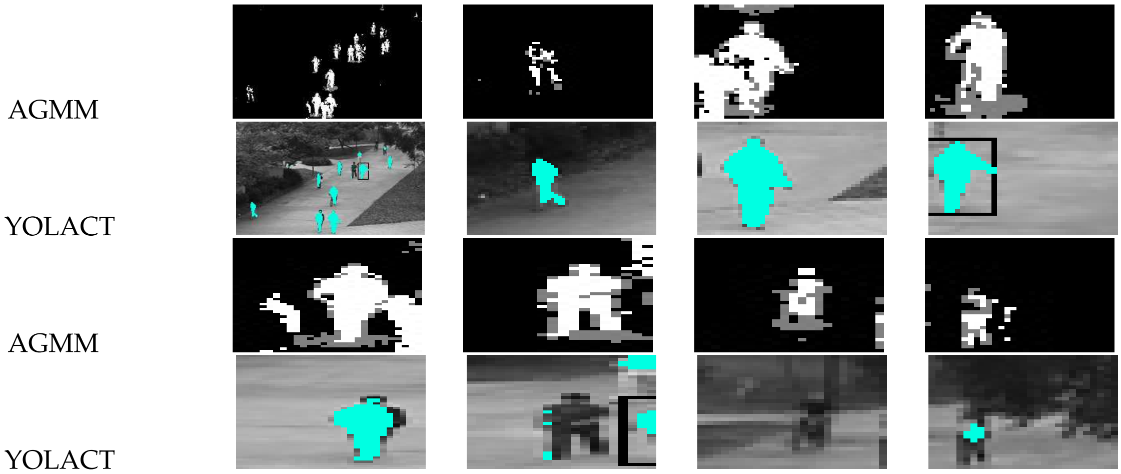

Figure 9. Comparison of object extraction results using YOLACT and AGMM.

4.3. Target Feature Analysis

The deep learning model based on PWC-Net optical flow estimation has the characteristics of strong robustness, high accuracy, and fast speed. The design of PWC-Net follows the principles of simplicity and completeness: pyramid processing, image warping using optical flow estimation, and the use of cost volume. Projected onto a learnable feature pyramid, PWC-Net warps the convolutional neural network features of the second image using the current optical flow estimate. It then constructs a cost volume using the warped features of the first image and the features, which is processed by the CNN to estimate the optical flow. In this paper, PWC-Net is used to extract features from the foreground map of video frames, which can obtain more accurate and robust feature representations, thereby improving the accuracy and efficiency of the abnormal classification model. The experimental results are shown in Figure 10.

Based on the results shown in Figure 10, PWC-Net demonstrates strong robustness against factors such as changes in lighting, occlusion, and motion blur. Additionally, PWC-Net uses a pyramid structure and flow propagation to improve the accuracy of feature matching and flow estimation. In cases where there is little change between video frames, such as in Video15, the target feature map remains almost unchanged. However, when abnormal objects appear in the video, the feature transformation of moving targets becomes particularly evident, as seen in Video11, Video13, Video14, and Video17. The advantage of such features is that they provide clear feature transformations for the input classifier to make judgments on abnormal frames, thus providing good information to the classifier and improving the classification accuracy.

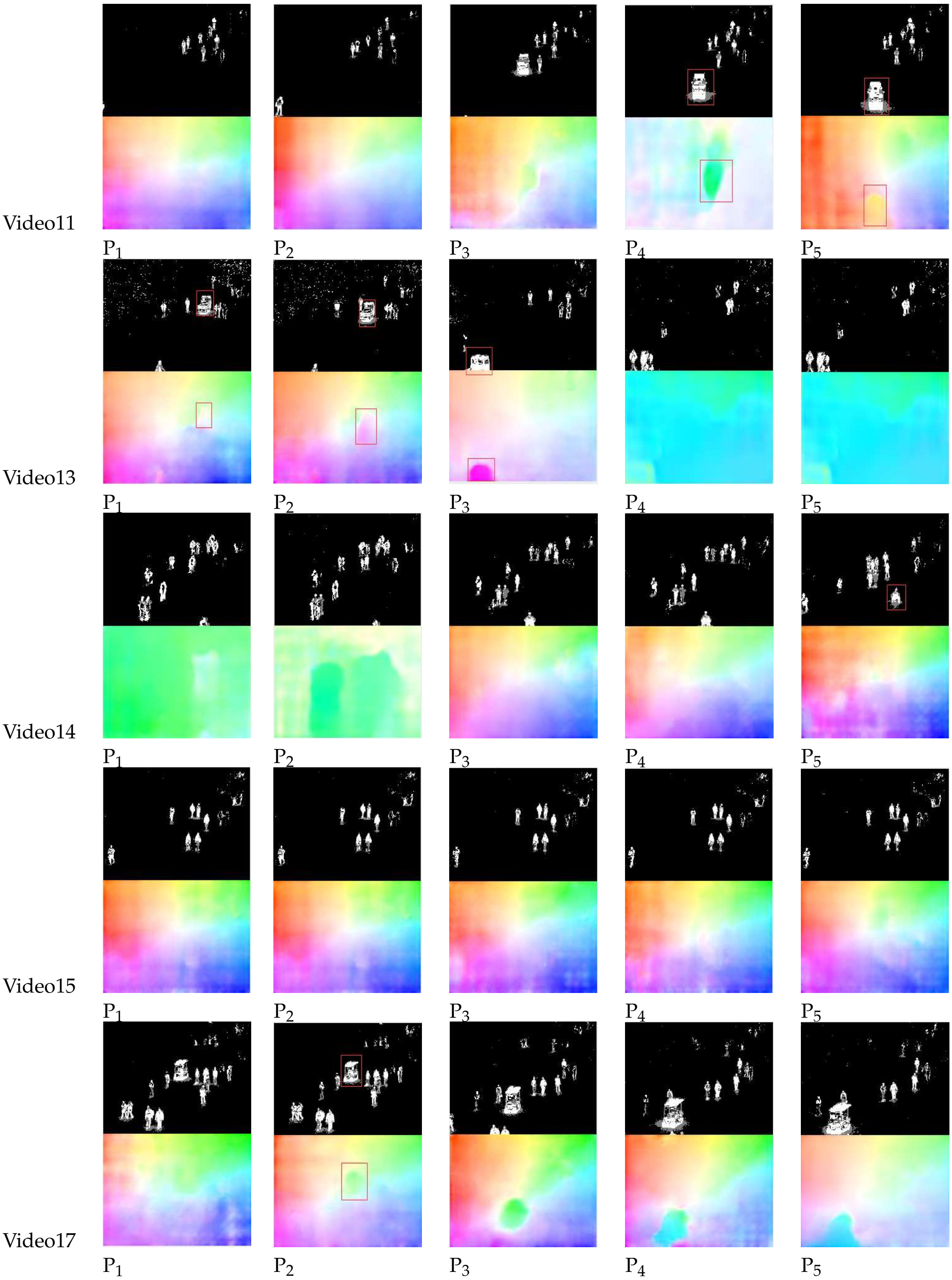

Figure 10. PWC-Net results on the UCSC dataset.

4.4. Anomaly Detection Analysis

In surveillance video anomaly detection, the UCSD dataset is used, where all anomalies occur naturally and are not staged for dataset assembly. The dataset is divided into two subsets, each corresponding to a different scene. Video clips recorded from each scene are segmented into various segments of approximately 200 frames. In this study, pedestrians walking normally are defined as normal, while abnormal targets include bicyclists, skateboarders, cars, wheelchairs, and small carts, as shown in Figure 11. To detect anomalous video frames, a ground truth is also created, which includes a binary label for each frame indicating whether an anomaly is present. Specifically, 36 subsets of Peds1 are chosen to provide hand-generated pixel-level binary codes that identify regions containing anomalies, as shown in Figure 12. This is done to enable performance evaluation of the method's ability to locate anomalies.

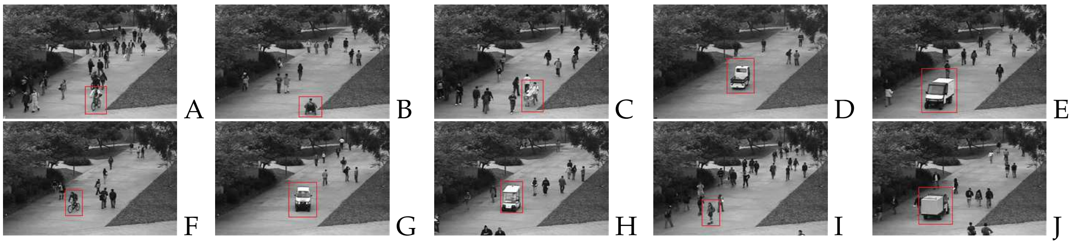

Figure 11. Abnormal samples on UCSD dataset.

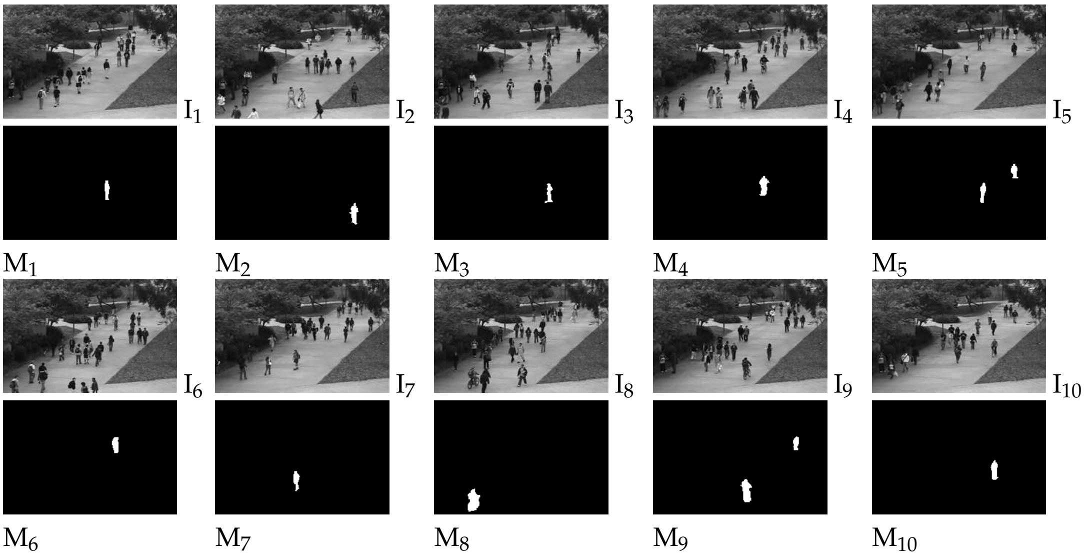

Figure 12. Displaying samples and their corresponding ground truth.

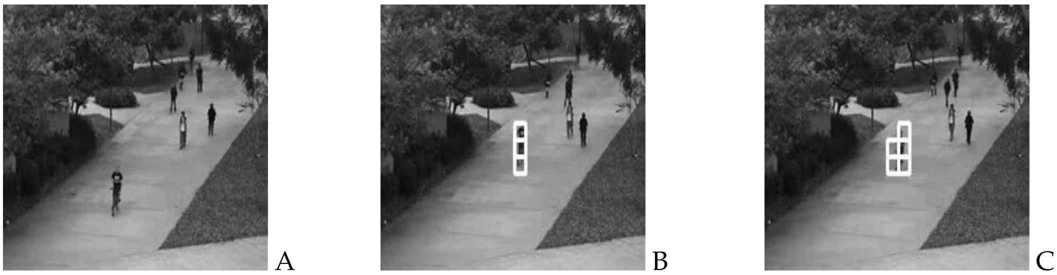

Figure 13. Abnormal behavior is marked with a white box. A. Normal video frame. B. Abnormal behavior detection with white box at time t. C. Abnormal behavior detection with white box at time $t+1$.

For video anomaly detection, a binary classifier was trained in this study to detect anomalies by using supervised learning to classify behaviors as normal or abnormal and assigning corresponding labels. Before modeling, all normal and abnormal data must be labeled and tagged using supervised classification methods. To improve the accuracy and generalization ability of the video anomaly detection model and enhance its prediction capability, this study used 107,306 datasets to train a classification model. Since videos contain temporal and spatial information, a neural network capable of extracting spatiotemporal features is required for anomaly detection. In this study, PWC-Net was

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2023 doi:10.20944/preprints202305.1617.v3

used to extract features for each detection target and assign corresponding labels, which were marked as $'0'$ for abnormal behavior and $'1'$ for normal targets. The anomaly detection results for the UCSD dataset are shown in Figure 13.

The detection and tracking of moving objects. YOLOv5 focuses on inference speed and accuracy. Considering the excellent performance of YOLOv5 in the object detection task, YOLOv5 is used as the object detection model of the network. The experimental results of YOLOv5-DeepSORT are shown in the fourth row in the Figures 14 and 15, the fourth row presents the abnormal detection with tracking.

4.5. Qualitative and Quantitative Analysis

Qualitative analysis. Figures 14 and 15 show some of the qualitative results of the detection of abnormal scenes. The first row of Figures 14 and 15 are the original video sequence, the second row of Figures 14 and 15 are foreground mask of raw video frames, the third row of Figures 14 and 15 are anomaly frames detected by the classifier and marks them with a white boxes. Figure 14 shows a case of a truck driving on a sidewalk that the system determines as abnormal behavior. Figure 15 shows that that a cyclist riding on the sidewalk is also considered an abnormal behavior because they are not allowed to use the sidewalk. As we can see from images in the fourth row, the scheme can track all pedestrians with high accuracy.

The scenes in the UCSD dataset are very diverse, and the people in the scenes have different body shapes and viewpoints, increasing the difficulty of the anomaly detection task. The defined anomaly measure allows us to identify multiple instances of several abnormal events that occur both individually and concurrently with other normal activities in the image. The objects causing the anomaly are marked with white boxes for identification. The experimental results demonstrate that our proposed method achieves high accuracy in classifying various scenes. It can effectively differentiate between normal pedestrian behavior and abnormal behaviors, such as those exhibited by cars, bicycles, wheelchairs, and skateboards.

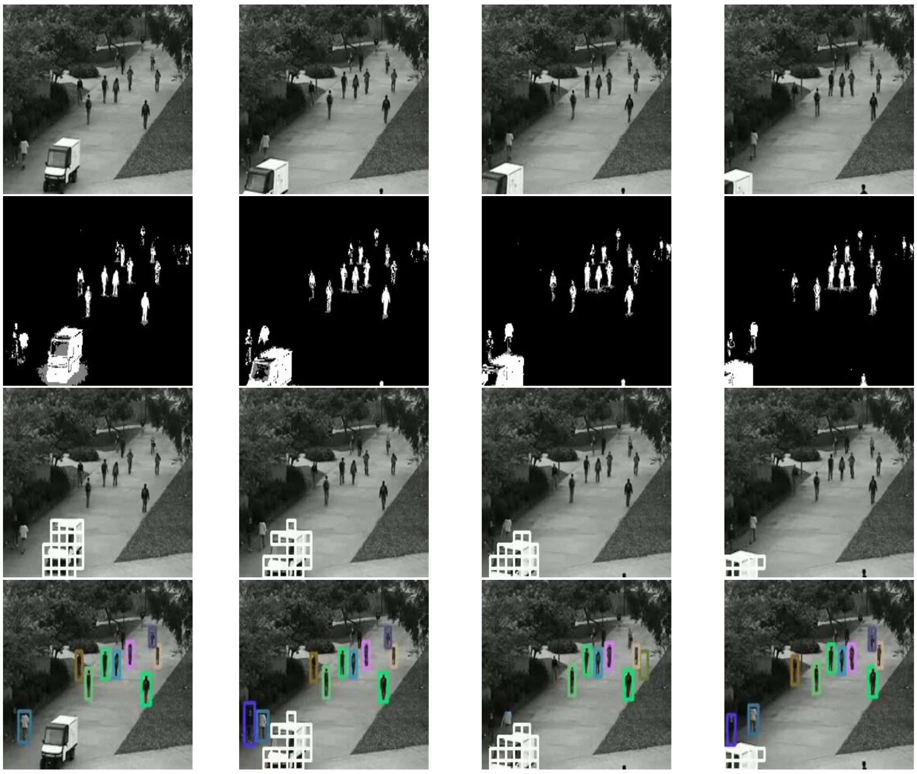

Figure 14. Example1 : Experimental results of the proposed method.

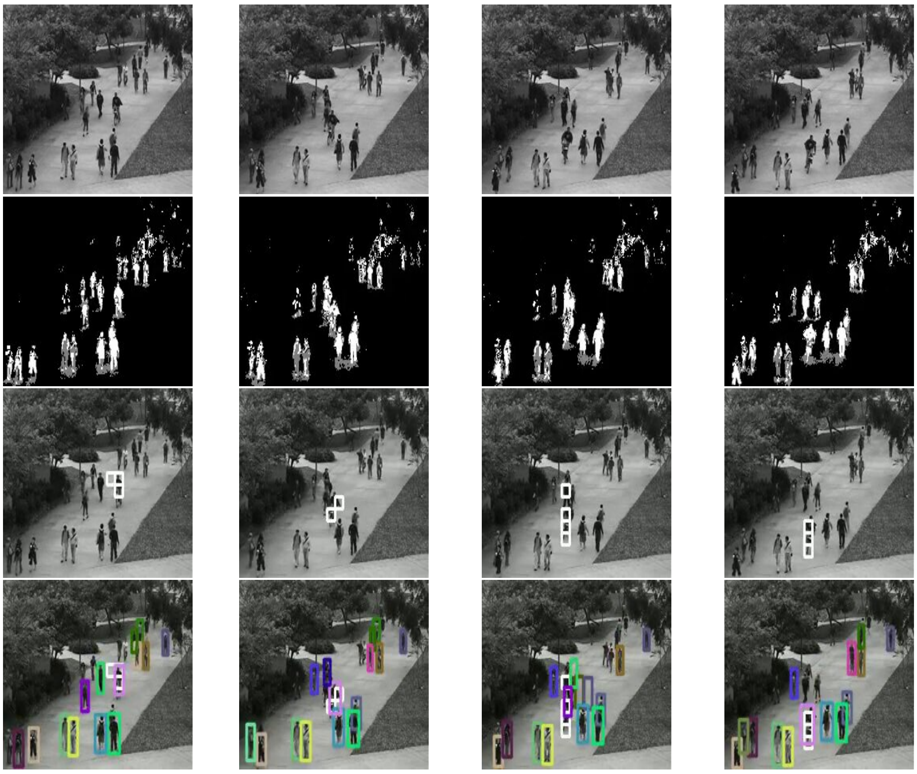

Figure 15. Example2 : Experimental results of the proposed method.

Quantitative evaluation. In the field of abnormal behavior detection, the performance of method is usually qualitatively evaluated by drawing receiver operating characteristic (ROC) curves with different thresholds on anomaly scores or probabilities, and quantitatively evaluated by metrics such as recognition accuracy (ACC), area under the ROC curve (AUC), and equal error rate (EER). ROC curves use the false positive rate (FPR) as the horizontal axis and the true positive rate (TPR) as the vertical axis. The FPR refers to the probability of predicting a positive sample among all actual negative samples, while the TPR refers to the probability of predicting a positive sample among all actual positive samples. Therefore, the closer the ROC curve is to the upper left corner, the smaller the EER, and the larger the AUC, indicating better algorithm performance.

In the field of abnormal behavior detection, it is necessary to detect the time and spatial location of abnormal behaviors, which are usually evaluated at two levels: frame-level and pixel-level. In the frame-level criterion, if any pixel in a frame is detected as abnormal, the frame is considered as an abnormal frame, regardless of whether the localization of the abnormal region is accurate or not. In contrast, the pixel-level criterion takes into account the spatial localization accuracy, and only when the detected abnormal pixels exceed the threshold of the true abnormal label, is the abnormal behavior considered to have occurred. Table 2 and Table 3 show the AUC and EER of different anomaly detection methods on the UCSD dataset under different evaluation criteria.

Table 2. Frame-level performance comparison of different methods on UCSD Ped1.

Methods	[39]	[18]	[19]	[20]	[22]	[21]	[23]	Ours
AUC	085	0.916	0.85	0.895	0.974	null	0.8382	0.9616
EER	0.24	0.148	0.20	null	8	0.143	0.223	0.15

Table 3. Pixel-level performance comparison of different methods on UCSD Ped1.

Methods	[39]	[18]	[19]	[20]	[22]	[21]	[23]	Ours
AUC	0.87	0.687	0.726	null	0.703	0.9425	null	0.8959
EER	null	0.357	null	null	0.35	null	null	0.39

As shown in Tables 2 and 3, we find that our supervised methods perform much better than the existing schemes. Figure 16 plots the ROC curves of the various methods for comparison. By varying the threshold parameter, we can obtain a series of anomaly detection results and their corresponding FPRs and TPRs. Thus, the ROC curve can be plotted by the series of coordinate points composed of FPRs and TPRs. The ROC curves illustrate that our method outperforms the state-of-the-art methods in [57–61] in detecting anomalous events on the frame-level evaluation criteria.

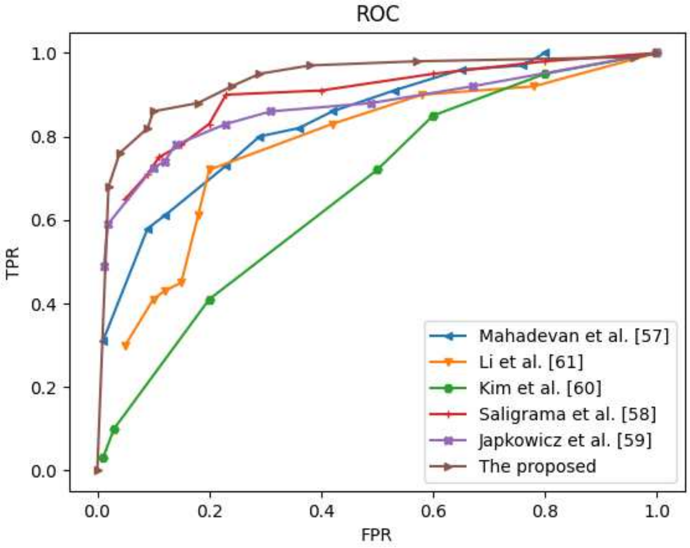

Figure 16. ROC curves of frame-level on Ped1.

On the basis of the ROC evaluation results, which reflect the accuracy of anomaly localization, our method outperforms all the comparing schemes. In addition to the ROC curves, the evaluation criteria also include two numerical indices, the AUC and EER in the frame-level, and the results are presented in Table 2 and Table 3. In terms of detection, the AUC values have increased, indicating that the method can locate anomalies with high accuracy. To evaluate the performance of the proposed anomaly representation, we compared it with five other recently proposed schemes [57–61]. Below are the performance results compared with those of other state-of-the-art schemes. Tables 2 and 3 show the comparison results of different algorithms on the UCSD Ped1 dataset at the frame level. The proposed method achieves a large frame-level AUC, which is better than those of the other comparison methods.

In [60] is just based on temporal anomaly detector is presented. Sikdar et al. [61] only considered the temporal and lack the spatial feature, thus, the accuracy of abnormal behavior detection is relatively low. Mahadevan et al. [57] obtained the errors made by the different detector components because anomalies are, by definition, difficult to define a priori, and normal events are either unusual or occur in unusual scenes. The method in [58] is simpler and requires very little parameter tuning compared with the other methods. The method's generalization ability is relatively poor, but the problem is relatively easy to solve. As shown in Figure 16 , we find that our supervised methods perform much better than the existing unsupervised schemes.

5. Conclusion

In this paper, we proposed a method for detecting and locating abnormal behavior in a monitoring scene by combining AGMM and YOLACT techniques to obtain more accurate foreground information. The PWC-Net algorithm is then used to extract features of the foreground images, which are fed into an anomaly classification model for classification, resulting in improved accuracy. Additionally, YOLOv5 and DeepSORT networks are employed for object detection and tracking in the video, respectively. The YOLOv5 network is able to detect different objects present in the video, while the DeepSORT network allows for better understanding of the scene in the video. Experimental results on the UCSD benchmark dataset demonstrate the effectiveness of our proposed method and its superiority over state-of-the-art schemes.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, W.H. and B.R.; methodology, W.H.; software, W.H.; validation, W.H., W.H. and B.R.; formal analysis, W.H.; investigation, W.H.; resources, L.L.; data curation, W.H.; writing—original draft preparation, W.H.; writing—review and editing, W.H.; visualization, L.L.; supervision, W.H.; project administration, W.H.; funding acquisition, W.H. All authors have read and agreed to the published version of the manuscript."

Funding: This research was funded by General Scientific Research Project of Zhejiang Education Department (No.Y202147609) and Humanities and Social Sciences Project of the Ministry of Education (17JDSZ3044).

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Data in this research paper will be shared upon request to the corresponding author.

Acknowledgments: Not Applicable.

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

Sample Availability: Not Applicable.

Abbreviations

The following abbreviations are used in this manuscript:

YOLOv5	You only look once version 5
GMM	Gaussian mixture model
OP	Optical flow
STT	Spatio-temporal technique
CNN	Convolutional neural networks
BoW	Bag-of-words
MRF	Markov random field
HMOFP	Histogram of maximal optical flow projection
AGMM	Adaptive gaussian mixture mode
DNN	Deep neural network
DeepSORT	Deep simple online and realtime tracking
MOT	Multiple object tracking
R-CNN	Region-based CNN
SSD	Single shot multiBox detector
ReID	Person Re-identification
FG	Foreground
BG	Background
UCSD	University of California San Diego

ROC	Receiver operating characteristic curve
AUC	Area under curve
EER	Equal error rate
FPS	Frame per second
TPR	True positive rate
FPR	False positive rate
TP	True positive
TN	True negative
FN	False negative
FP	False positive
PWC	Pyramid, Warping, and Cost Volume
YOLACT	You Only Look At CoefficienTs
SVM	Support vector machine
HOG	Histogram of oriented gradients
S-CNN	Slicing CNN
LDA	Linear discriminant analysis
LSTM	Long short-term memory
ST-CNN	Spatial-temporal convolutional neural network
IoU	Intersection over union
MoSIFT	Motion scale invariant feature transform
DBT	detection-based tracking
COCO dataset	Microsoft Common Objects in Context
S-CNN	Slicing-convolutional neural networks

References

1. Ren, J.; Xia, F.; Liu, Y.; I, Lee. Deep Video Anomaly Detection: Opportunities and Challenges. *2021 International Conference on Data Mining Workshops (ICDMW)* **2021**, 959—966.
2. L, Wan.; Y, Sun.; I, Lee.; W, Zhao.; F, Xia. Industrial pollution areas detection and location via satellite-based IIoT. *IEEE Transactions on Industrial Informatics* **2020**, vol. 17, no. 3, pp. 1785–1794.
3. Xu, S.; Zhao, M.; Huang, K. Real-time Video Anomaly Detection with Region Proposal Network and YOLO-Act. *IEEE Transactions on Industrial Informatics* **2021**, 17 (9), 6446-6454. DOI: 10.1109/TII.2021.3063356.
4. Sun, D.; Yang, X.; Liu, M. Y.; Kautz, J. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. *Proceedings of the IEEE conference on computer vision and pattern recognition* **2018**, pp. 8934-8943.
5. Wu, H.; Xiong, Y.; Yu, J. Real-Time Video Anomaly Detection Using PWC-Net and Frame Difference. *IEEE Access* **2021**, 9, 64281-64289.
6. Peng, Y.; Wei, X.; Liu, X. Real-time video anomaly detection based on improved PWC-Net. *Journal of Ambient Intelligence and Humanized Computing* **2020**, 11 (1), 177-184.
7. S, Varadarajan.; H, Wang.; P, Miller.; H, Zhou. Fast convergence of regularized Region-based Mixture of Gaussians for dynamic background modelling. *Computer Vision and Image Understanding* **2015**, 136 : 45–58.
8. R, Azzam.; M.S, Kemouche.; N, Aouf.; M, Richardson. Efficient visual object detection with spatially global Gaussian mixture models and uncertainties. *Journal of Visual Communication and Image Representation* **2016**, 36 : 90–106.
9. Z, Ji.; Y, Huang.; Y, Xia.; Y, Zheng. A robust modified Gaussian mixture model with rough set for image segmentation. *Neurocomputing* **2017**, 266 : 550–565.
10. M, Sabokrou.; M, Fathy.; M, Hoseini.; R, Klette. Real-time anomaly detection and localization in crowded scenes. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Hynes Convention Center in Boston, Massachusetts* **2015**, 56–62.
11. R, Leyva.;, V. Sanchez, C.T. Li. Video anomaly detection with compact feature sets for online performance. *IEEE Trans. Image Process* **2017**, 26(7):3463–3478.
12. T, Lu.; L, Wu.; X, Ma.; P, Shivakumara.; C.L, Tan. Anomaly detection through spatio-temporal context modeling in crowded scenes. *22nd International Conference on Pattern Recognition* **2014**, 2203–2208.
13. M, Marsden.; K, McGuinness.; S, Little.; N.E.O., Connor. Holistic features for real time crowd behaviour anomaly detection. *IEEE International Conference on Image Processing, Phoenix*, **2016**, 918–922.

14. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. Acm* **2017**, 60, 84–90.
15. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 23–28 June **2014**, pp. 580–587.
16. Afiq, A.A.; Zakariya, M.A.; Saad, M.N.; Nurfarzana, A.A.; Khir, M.H.M.; Fadzil, A.F.; Jale, A.; Gunawan, W.; Izuddin, Z.A.A.; Faizari, M. A review on classifying abnormal behavior in crowd scene. *Vis. Commun. Image Represent* **2019**, 58, 285–303.
17. Xu, D.; Yan, Y.;Ricci, E. Detecting anomalous events in videos by learning deep representations ofappearance and motion. *Computer Vision andImage Understanding* **2017**, 156: 117-127.
18. Tran, H.; Hogg, D. Anomaly detection using a convolu-tional winner-take-all autoencoder. *Proc of the BritishMachine Vision Conference* **2017**: 1-13.
19. Li, J.; Chang,L. Video anomaly detection andlocalization via multivariate Gaussian fully convolution adversarial auto encoder. *Neuro computing* **2019**, 369: 92-105.
20. Ribeiro, M.; Lazzaretti, A.; E, Lopes H S. A study of deep convolutional auto-encoders for anomaly detection in videos. *Pattern Recognition Letters* **2018**, 105: 13-22.
21. Wang, T.;Qiao, M N.; Lin, Z W. Generative neural networks for anomaly detection in crowded scenes. *IEEE Transactions on Information Forensics and Security* **2019**, 14 (5) : 1390-1399.
22. Ravanbakhsh, M.; Nabi, M.; Sangineto, E. Abnormal event detection in videos using generative adversarial nets. *2017 IEEE InternationalConference on Image Processing (ICIP)* **2017**, 2017: 1577-1581.
23. Li, Y Y.; Cai, Y H.; Liu, J Q. Spatio-temporal unity networking for video anomaly detection. *2017 IEEE Access* **2019**, 7: 172425-172432.
24. Y, Hu.; H, Chang.; F, Nian.; Y, Wang.; T, Li. Dense crowd counting from still images with convolutional neural networks. *Journal of Visual Communication & Image Representation* **2016**, 38 (C): 530-539.
25. J, Shao.; C.C., Loy.; K, Kang.; X, Wang. Slicing convolutional neural network for crowd video understanding. *IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas* **2016**: 5620–5628.
26. A, D.; Chevrolet, J. C.; Chevret, S. et al. Two-Stream Convolutional Networks for Action Recognition in Videos. *Advances in Neural Information Processing Systems* **2014**, 1 (4): 568-576.
27. C, Feichtenhofer.; Pinz, A.; Zisserman, A. Convolutional Two-stream Network Fusion for Video Action Recognition. *IEEE Conference on Computer Vision and Pattern Recognition* **2016**: 1933-1941.
28. S, Yi.; H, Li.; X, Wang. Pedestrian behavior understanding and prediction with deep neural networks. *European Conference on Computer Vision, Amsterdam, The Netherlands*, **2016**: 263–279.
29. Luo et al. Crowd abnormal behavior recognition based on deep learning and sparse optical flow. *Computer Engineering* **2020**, 46 (4): 287-293, 300.
30. Gong, M G.; Zeng, H M.; Xie, Y. et al. Local distinguish ability aggrandizing network for human anomaly detection. *Neural Net works* **2020**, 122: 364-373.
31. Zou, Y F. Recognition and research about abnormalbehavior of human based on video. *Kunming :Yunnan University* **2019**.
32. D, Fortun.; P, Bouthemy.; C, Kervrann. Optical flow modeling and computation: a survey. *Computer Vision and Image Understanding* **2015**, 134: 1-21.
33. Y, Yuan.; Y, Feng.; X, Lu. Statistical hypothesis detector for abnormal event detection in crowded scenes. *IEEE Transactions on Cybernetics* **2016**, 99: 1-12.
34. Q, Wang.; Q, Ma.; C.H., Luo.; H.Y., Liu.; C.L., Zhang. Hybrid histogram of oriented optical flow for abnormal behavior detection in crowd scenes. *International Journal of Pattern Recognition and Artificial Intelligence* **2016**, 30 (02): 14.
35. Li, A.; Miao, Z.; Cen, Y.; Wang, T.; Voronin, V. Histogram of maximal optical flow projection for abnormal events detection in crowded scenes. *International Journal of Distributed Sensor Networks* **2015**, 11 (11): 406941.
36. Larsen, M.L.; Schönhuber, M. Identification and Characterization of an Anomaly in Two-Dimensional Video Disdrometer Data. *Atmosphere* **2018**, 9, 315.
37. Kumar, K.; Kumar, A.; Bahuguna, A. D-CAD: Deep and crowded anomaly detection. *Proceedings of the 7th international conference on computer and communication technology* **2017**: 100-105.
38. T, Lu.; L, Wu.; X, Ma.; P, Shivakumara.; C.L., Tan. Anomaly detection through spatio-temporal context modeling in crowded scenes[C]. *22nd International Conference on Pattern Recognition* **2014**, 2203–2208.

39. Zhou, S F.; Shen, W.; Zeng, D. et al. Spatial-temporal convolu tional neural networks for anomaly detection and localization in crowded scenes. *Signal Processing:Image Communication* **2016**, 47 (9): 358-368.
40. Sabokrou, M.; Fayyaz, M.; Fathy, M. et al. Deep-cascade: cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing* **2017**, 26 (4): 1992-2004.
41. Miao, Y Y.; Song, J X.; Abnormal event detection based on SVM in video surveillance. *IEEE Workshop on Advanced Research and Technology in Industry Applications, Ottawa, Canada* **2014**, 1379-1383.
42. Wojke, N.; Bewley, A.; Paulus, D. Simple Online and Realtime Tracking with a Deep Association Metric. *arXiv e-prints*, **2017**.
43. Ciaparrone, Gioele. et al. Deep learning in video multi-object tracking: A survey. *Neurocomputing* **2020**: 61-88.
44. Piccardi, M. Background subtraction techniques: a review. *Systems, Man and Cybernetics. IEEE* **2004**.
45. Zoran, Zivkovic.; Ferdinand, van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters* **2006** 27 (7): 773–780.
46. Zoran, Zivkovic. Improved adaptive gaussian mixture model for background subtraction. *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference* **2004**, volume 2, pages 28–31.
47. O'donovan, Peter J. Optical Flow : Techniques and Applications. **2005**.
48. Guizilini, V.; Lee, K. H.; Ambrus, R. et al. Learning Optical Flow, Depth, and Scene Flow without Real-World Labels. **2022**.
49. Sebastian Bullinger, Christoph Bodensteiner.; Michael, Arens. Instance flow based online multiple object tracking. *2017 IEEE International Conference on Image Processing (ICIP)* **2017**, pages: 785–789.
50. Gunnar, Farnebäck. Two-frame motion estimation based on polynomial expansion. *Scandinavian conference on Image analysis. Springer* **2003**, pages; 363–370.
51. Jerome, Revaud.; Philippe, Weinzaepfel.; Zaid, Harchaoui.; Cordelia, Schmid. Deepmatching: Hierarchical deformable dense matching. *International Journal of Computer Vision* **2016**, 120 (3): 300–323.
52. Yinlin, Hu.; Rui, Song.; Yunsong, Li. Efficient coarse-to-fine patchmatch for large displacement optical flow. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* **2016**, pages: 5704–5712.
53. G, Farneback. Two-frame motion estimation based on polynomial expansion. *Image Analysis. Springer* **2003**, pages: 363–370.
54. D, Sun.; X, Yang.; M, Y Liu.; J, Kautz. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA* **2018**, pp. 8934-8943, doi: 10.1109/CVPR.2018.00931.
55. Z, Zivkovic. Improved adaptive Gaussian mixture model for background subtraction. *Proceedings of the 17th International Conference on Pattern Recognition ICPR* **2004**, pp. 28-31 Vol.2, doi: 10.1109/ICPR.2004.1333992.
56. Mahadevan, V.; Li, W.; Bhalodia, V.; V, asconcelos, N. Anomaly detection in crowded scenes. *CVPR* **2010**, pp. 1975–1981.
57. W, Li.; V, Mahadevan.; N, Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, 36 (1): 18– 32, jan.
58. Mahadevan, V.; Li, W.; Bhalodia, V.; Vasconcelos, N. Anomaly detection in crowded scenes. *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2010**, San Francisco, CA, USA, 13–18, pp. 1975–1981.
59. Japkowicz, N.; Myers, C.; Gluck, M. A. A Novelty Detection Approach to Classification. *Proceedings of the 14th International Conference on Artificial Intelligence (IJCAI-95)* **1995**, pp. 518–523.
60. J, Kim.; K, Grauman. Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates. *IEEE Conf. Computer Vision and Pattern Recognition* **2009**, 1, 2,4, 8, 9.
61. Sikdar, A.; Chowdhury, A. S. An adaptive training-less system for anomaly detection in crowd scenes. *ArXiv preprint arXiv:1906.00705,* **2019**, 36 (1): 18-32.

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