Received: 15 June 2021 | Revised: 25 October 2021 | Accepted: 11 November 2021 | CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12068

ORIGINAL RESEARCH

Abnormal event detection by a weakly supervised temporal attention network

Xiangtao Zheng[1] | **Yichao Zhang**[1,2] | **Yunpeng Zheng**[1,2] | **Fulin Luo**[3] | **Xiaoqiang Lu**[1]

[1]Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, Shaanxi, China

[2]University of Chinese Academy of Sciences, Beijing, China

[3]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore

Correspondence

Xiangtao Zheng, Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, China.
Email: xiangtaoz@gmail.com

Funding information

National Natural Science Foundation of China, Grant/Award Numbers: 61772510, 61806193; National Science Fund for Distinguished Young Scholars, Grant/Award Number: 61925112; Innovation Capability Support Program of Shaanxi, Grant/Award Number: 2020KJXX-091; Key Research Program of Frontier Sciences, Chinese Academy of Sciences, Grant/Award Number: QYZDY-SSW-JSC044

Abstract

Abnormal event detection aims to automatically identify unusual events that do not comply with expectation. Recently, many methods have been proposed to obtain the temporal locations of abnormal events under various determined thresholds. However, the specific categories of abnormal events are mostly neglect, which are important to help in monitoring agents to make decisions. In this study, a *Temporal Attention Network* (TANet) is proposed to capture both the specific categories and temporal locations of abnormal events in a weakly supervised manner. The TANet learns the anomaly score and specific category for each video segment with only video-level abnormal event labels. An event recognition module is exploited to predict the event scores for each video segment while a temporal attention module is proposed to learn a temporal attention value. Finally, to learn anomaly scores and specific categories, three constraints are considered: event category constraint, event separation constraint and temporal smoothness constraint. Experiments on the University of Central Florida Crime dataset demonstrate the effectiveness of the proposed method.

KEYWORDS

human detection, video analysis

1 | INTRODUCTION

Abnormal event detection has attracted extraordinary attention in the computer vision community [1] due to its critical applications such as video surveillance [2], human computer interfaces [3], violence alerting [4], evidence investigation [5] etc. Generally speaking, abnormal events mean patterns or motions that rarely occur in videos and are different from existing events. The core objective of abnormal event detection is to automatically identify abnormal events from surveillance videos [6]. However, the low quality of videos, shadows, occlusions, illuminations and complex backgrounds make abnormal event detection a very challenging task.

Abnormal events are extremely complicated in real-world environments, and the definitions of these events always have some extent of ambiguity according to different application scenarios [7]. Most previous methods learn patterns of abnormal events by modelling normal events in the given training datasets [8]. The patterns that never appeared in training videos are all treated as abnormal events [9–11]. These methods can be considered as unsupervised methods because there are no straightforwardly supervised information of

abnormal events in the training stage. It is obvious that the modelling of normal events is of great importance for unsupervised abnormal event detection methods [12]. To be more specific, unsupervised methods can be divided into two groups according to the development timeline, that is, hand-crafted feature methods and deep learning methods. Early methods belong to the first group, which models normal events by designing semantic video representations with hand-crafted features. For example, Reddy et al. [9] proposed to model normal events in videos by extracting local features of motion, size and texture in non-overlapping cells of foreground frames. Zhao et al. [10] obtained normal event representation by dynamic sparse coding on *Histogram of Gradient* and *Histogram of Optical Flow* (HOF) from local spatial–temporal interest patches.

Nowadays, with the blooming of deep learning [13], many deep neural networks have been proposed for abnormal event detection. For example, Xu et al. [11] designed three *Stacked Denoising AutoEncoders* to learn deep representations of appearance and motion as well as the pixel-level fusion of appearance and motion, respectively, and then combined them to model the pattern of normal events. Luo et al. [8] proposed to reconstruct normal events by mapping the *Temporally-coherent Sparse Coding* (TSC) with *stacked Recurrent Neural Network*, and then abnormal events can be detected based on the reconstruction error.

Most of the recent methods rely too much on the training datasets when they detect the anomaly events. These methods treat any behaviours different from the normal training datasets as anomaly ones. In real-world situations, it is unreasonable since the same behaviour may be anomalous or normal under different conditions [14]. As shown in Figure 1a, the behaviour of *car driving* is considered as an anomaly in the University of California, San Diego Pedestrian 2 (USCD ped2) dataset [15] but is considered as normal in the Street Scene dataset if the car is driving in the appropriate direction [16]. Nevertheless, monitoring agents not only care about the temporal locations of the abnormal events but also care about which kind of abnormal event occurs in practice. For example, as shown in Figure 1b, both 'running red lights' and 'violent clashes on the street' are abnormal events in real world, but they should be handled with different priorities apparently. Therefore, it is of great importance to exploit the specific categories of abnormal events. In this case, both anomaly scores and abnormal event categories can be obtained to help in monitoring agents to make decisions.

In practice, an abnormal event only occurs for a very short time in the long surveillance video. It is difficult to simultaneously detect and recognize the abnormal event, owing to the limitation that annotating the exact temporal duration and specific categories of abnormal events is very time-consuming and expensive. To address this limitation, *Temporal Attention Network* (TANet) is proposed in this study to learn the anomaly scores and specific categories of abnormal events with only the supervision of video-level abnormal event labels. More specifically, as abnormal events only occur for a very short time in a long video, it is obviously improper to directly treat a video as whole to predict the anomaly scores and abnormal event categories with the video-level abnormal event labels. Thus the strategy of *Multiple Instance Learning* (MIL) [14] is adopted in this study. A video containing abnormal events is treated as a positive bag, and a video not containing abnormal events is treated as a negative bag. Each bag contains multiple short video segments as instances. Then, three modules are included in the TANet to learn both anomaly scores and specific categories of abnormal events, that is, *Events*

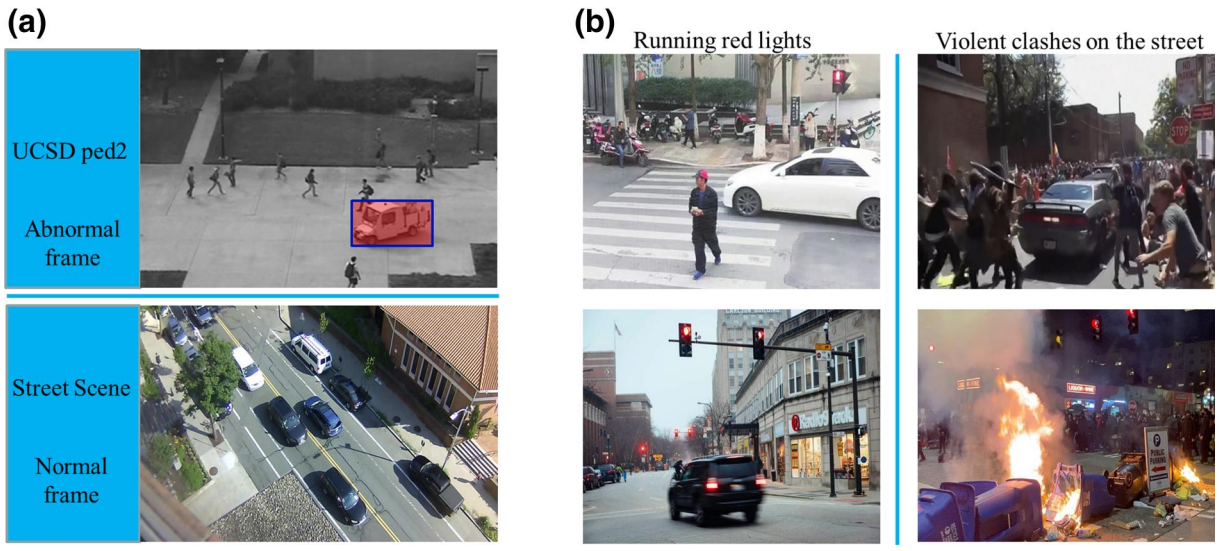

FIGURE 1 (a) The same event can be considered as both anomaly and normal in different application scenarios. For example, *car driving* is considered as anomaly in the University of California, San Diego Pedestrian 2 (UCSD ped2) dataset but is considered as normal in the Street Scene dataset if the car is driving in the appropriate direction. (b) Different abnormal events have different priorities. For example, both 'running a red light' and 'violent clashes on the street' are abnormal in real-world surveillance, but they should be handled with different priorities apparently

Recognition Module (ERM), *Temporal Attention Module* (TAM) and overall object function. As shown in Figure 2, first, the ERM is proposed to predict the event scores for all video segments; second, the TAM takes advantage of the event scores to generate a temporal attention value for each video segment by exploring the relationship of all abnormal events. In addition, each temporal attention value can reflect an attention degree that a corresponding video segment may contain an abnormal event. It is worth noting that based on the temporal attention values, both positive and negative instances can be localized. Finally, based on the located positive and negative instances, both anomaly scores and specific abnormal categories are learned. The proposed method is considered as a weakly supervised method since only specific video-level abnormal event labels are adopted as supervision. During testing, the anomaly scores of all video segments in a long untrimmed video can be predicted. Moreover, for videos that contain abnormal events, the specific label of abnormal events can also be predicted. In summary, the main contributions of this study can be summarized as follows:

1. A weakly supervised method TANet is proposed in this study, which can learn both anomaly scores and specific abnormal categories with only the supervision of abnormal event labels
2. The TAM can learn a temporal attention value of each video segment, thus abnormal segments in a training video can be more accurately located. The overall object function can guide TANet in learning more discriminative representations for abnormal event detection and recognition
3. Experiments conducted on the University of Central Florida (UCF) Crime dataset demonstrate the effectiveness of the proposed method, which can achieve better performance compared with other methods

The rest of this paper is organized as follows: Section 2 reviews related work. Section 3 describes the proposed method in detail. Section 4 presents the evaluation of the proposed method on the UCF Crime dataset. Finally, Section 5 presents concluding remarks of this paper.

2 | RELATED WORK

This section discusses the related studies of the proposed method. In particular, Subsection 2.1 reviews abnormal event detection. Subsection 2.2 reviews temporal action localization methods and discusses the relevance to this paper.

2.1 | Abnormal event detection

Abnormal events are extremely complex and diverse [17]; thus, most previous methods are designed as unsupervised methods due to the fact that listing all kinds of abnormal events is nearly impossible. Unsupervised methods are divided into hand-crafted feature methods [18, 19] and deep learning methods [20, 21] in this paper.

Hand-crafted feature methods primarily learn normal events by extracting hand-crafted features from trajectories [22, 23] or spatial–temporal local patches [17, 19]. For example, both Wu et al. [22] and Basharat et al. [18] proposed to represent normal events in videos by modelling trajectories with *Gaussian mixture models*. Cui et al. [24] proposed an interaction energy potential function to model human group activities by tracking spatial–temporal interesting points in videos. Anjum and Cavallaro [25] extracted multiple trajectory features of video objects and then adopted a clustering algorithm to model normal and rare object motion patterns. Adam et al. [19] proposed to model normal events in videos by estimating the probability distribution of optical flow at various scene locations in the videos. Xu et al. [26] extracted HOF of both the global scene and local spatial–temporal cells in videos to learn multi-level normal event patterns. Li et al. [17] proposed a hierarchical *Mixture of Dynamic Texture*s model to obtain a joint representation of videos with a background subtraction technique and a discriminant saliency method.

With the development of deep learning [27], many deep neural networks have been proposed for abnormal event detection. Deep learning methods refer to methods that apply to deep neural networks for abnormal event detection [13, 14, 21]. Deep auto-encoders are most used to learn normal

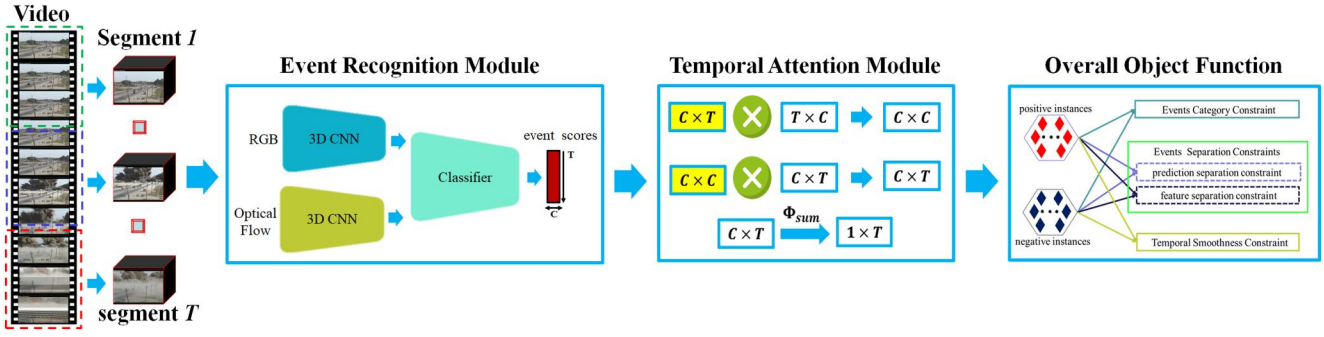

FIGURE 2 The overview of the proposed method. The definitions of symbols are explained in detail in Section 3. *T* video segments are sampled from each video. Then, three modules are proposed for abnormal event detection and recognition: first, *Events Recognition Module* is proposed to predict the event scores of all video segments; second, *Temproal Attention Module* is proposed to generate the abnormal confidence scores for each input video segments and finally, three constraints are considered to learn the anomaly scores and the specific categories of abnormal events

events in videos by reconstructing current frames [8, 21] or predicting feature frames [28, 29]. For example, Hasan et al. [30] proposed a fully connected auto-encoder and a fully convolutional feed-forward auto-encoder to reconstruct hand-crafted features and local image patches, respectively. Then, the abnormal events can be detected by the fusion of reconstruction costs. Ravanbakhsh et al. [21] attempted to learn spatial and motion patterns of normal frames by two condition *Generative Adversarial Networks* (GANs). Since their GANs are only trained with normal data, the magnitude of the difference between real and the generated frames can be used to detect abnormal events during testing. Different from [21], Liu et al. [29] adopted a Least Square GAN [13] and designed three constraints to predict future frames for abnormal event detection. After training, events with larger generation costs are considered as abnormal.

Besides unsupervised methods, a few methods have been proposed for abnormal event detection with more supervision recently. For example, as mentioned in Section 1, Sultani et al. [14] proposed a weakly supervised method to detect real-world abnormal events, in which a deep multiple instance ranking framework is adopted to learn representations of both normal and abnormal events with sparsity and temporal constraints. Besides, Zhu and Newsam [31] also proposed an attention-based temporal MIL framework with motion-aware features to detect potential abnormal events in untrimmed videos in a weakly supervised manner. In addition, Perez et al. [32] introduced a new dataset and proposed a fully supervised method to detect the temporal location of *Closed-Circuit TeleVision* fight events.

2.2 | Temporal action localization

Motivated by recent achievement in image object detection [33], a number of methods have been proposed to perform temporal action localization in untrimmed videos. In general, these methods can be briefly divided into two groups according to the degree of supervision, namely fully supervised methods [33, 34] and weakly supervised methods [35, 36]. The former mainly takes advantage of all useful clues such as frame labels, spatial bounding boxes and video labels to detect the temporal action location of actions. The latter commonly treats video-level labels or coarse video annotations as supervision.

More specifically, fully supervised methods tend to adopt proposals then classification frameworks to locate the start and end points of actions. For example, both Saha et al. [34] and Peng et al. [37] proposed to employ *Region Proposal Networks* [33] to generate temporal proposals for temporal action localization. Escorcia et al. [38] proposed to adopt *Long Short-Term Memory* cells to learn a set of temporal action proposals. Kalogeiton et al. [39] proposed a method called *ACtion Tubelet detector* (ACT-detector) to generate video tubelets, which are treated as spatial–temporal anchor action cuboids to detect the spatial–temporal action locations.

Weakly supervised methods primarily aim to predict framewise action labels with video-level supervision, such as video-level action labels [36], movie scripts or subtitles, and the temporal order of video frames [40]. In particular, the video-level action labels are most commonly used [41]. For example, Wang et al. [35] proposed to models for action recognition and detection without detailed temporal annotations of action instances. Islam and Radke [42] proposed a balanced binary cross-entropy loss and a metric loss to learn discriminative action representations for weakly supervised temporal action localization.

In this study, the TANet is proposed for abnormal event detection and recognition with only the supervision of video-level abnormal event labels. From this perspective, methods for temporal action localization are related to this study. However, the TANet is different from methods for temporal action localization in the following two aspects. First, the datasets are quite different. More specifically, this study treats UCF Crime dataset [14] as the benchmark dataset, which consists of much longer untrimmed videos than datasets used in temporal action localization. Moreover, compared with datasets for action recognition, the UCF Crime dataset is more challenging. For example, the method of C3D [43] achieved 85.2% accuracy on the UCF101 dataset (a benchmark dataset for action recognition and temporal action localization) but only achieved 23.0% on the UCF Crime dataset [14]. Second, the performance metrics between the TANet and methods for temporal action localization are utterly contrasting [44]. Generally speaking, the performances of temporal action localization are measured by finding the temporal interval, which is overlapped with ground truth as much as possible, but the performances of abnormal event detection are measured by a series of frame-level performances under various determined thresholds.

3 | THE PROPOSED METHOD

This section introduces the proposed method in detail. Subsection 3.1 formalizes abnormal event detection problem. Subsection 3.2 introduces specific steps of ERM. Subsection 3.3 gives details of TAM. Subsection 3.4 introduces the overall object function. Subsection 3.5 describes how to predict anomaly scores and abnormal event categories with the proposed method during testing.

3.1 | Problem formulation

Previous unsupervised methods for abnormal event detection easily suffer from the limitation of generalization. Though a weakly supervised method proposed by Sultani et al. [14] can have better generalization with the supervision of video-level binary abnormal labels, it cannot predict the specific abnormal event categories. To learn both anomaly scores and specific categories of abnormal events, a weakly supervised method called TANet is proposed in this study. To be more specific, let $\mathcal{V}_{normal} = \{v_n\}_{n=1}^{N_{normal}}$ denote the normal video set which contains N_{normal} videos, where v_n is the nth video in

$\mathcal{V}_{normal}$, and let $\mathcal{V}_{anomaly} = \{v_a, y_a\}_{a=1}^{N_{anomaly}}$ denote the anomaly video set with $N_{anomaly}$ videos in which each video contains a specific kind of abnormal event. v_a is the ath video in $\mathcal{V}_{anomaly}$. $y_a = [y_{a;1}, y_{a;1}, \ldots, y_{a;C}] \in \mathbb{R}^C$, where C is the total number of abnormal event categories. If the v_a contains qth abnormal event category, where $q \in \{1, 2, \ldots, C\}$, $y_{a;q} = 1$, otherwise $y_{a;q} = 0$. Then, the TANet is trained on $\mathcal{V}_{normal}$ and $\mathcal{V}_{anomaly}$ with the guidance of the overall object function. After training TANet, the anomaly score can be predicted for each frame in a testing video. Moreover, the category of a specific abnormal event can be predicted for a testing video containing abnormal events.

In addition, during training each video in $\mathcal{V}_{normal}$ and $\mathcal{V}_{anomaly}$ is split into several non-overlapping video segments due to memory constraints. Note that videos in both $\mathcal{V}_{normal}$ and $\mathcal{V}_{anomaly}$ are processed the same in ERM and TAM, so we take video v_a as an example to describe the process in the ERM and TAM.

3.2 | Event recognition module

Let $v_a = \{s_{a,l}\}_{l=1}^{L_a}$ denote the split video segment set, where L_a is the total number of video segments in a video v_a. The video length in the $\mathcal{V}_{normal}$ and $\mathcal{V}_{anomaly}$ varies dramatically; thus, a fixed number of T video segments are sampled in each training video for the sake of unifying all training videos. Let $\tilde{v}_a = \{\tilde{s}_{a,t}\}_{t=1}^{T}$ denote the sampled video segment set of video v_a, in which each segment contains the same number of frames. As shown in Figure 3, the ERM is exploited to predict the event scores for each video segment $\tilde{s}_{a,t}$.

3.2.1 | Feature extraction

This study adopts a two-stream 3DCNN (I3D network [45]) as the backbone network, due to its outstanding performance on the video classification tasks. Given a video segment $\tilde{s}_{a,t}$, the I3D network extracts the spatial (RGB) features $x_{a,t}^{rgb} \in \mathbb{R}^D$ and temporal (optical flow) features $x_{a,t}^{flow} \in \mathbb{R}^D$ from RGB frames and optical flows, respectively. Then, the spatial features $x_{a,t}^{rgb}$ and temporal features $x_{a,t}^{flow}$ are connected together as the spatial–temporal feature $x_{a,t} \in \mathbb{R}^{2D}$. Afterwards, the concatenated spatial–temporal features of all video segments in $\tilde{v}_a$ are denoted as $X_a = [x_{a,1}, \ldots, x_{a,T}] \in \mathbb{R}^{2D \times T}$.

3.2.2 | Temporal fusion layer

To fuse the spatial and temporal features, a temporal fusion layer is adopted on X_a to obtain discriminative representations. The feature fusion layer $\Phi_{fusion}(\cdot; \theta_{fusion})$ consists of a 1×3 temporal convolutional layer and a ReLU activation function. The temporal fusion feature set F_a of video $\tilde{v}_a$ can be obtained as follows:

$$F_a = \Phi_{fusion}(X_a; \theta_{fusion}), \tag{1}$$

where $F_a = [f_{a,1}, \ldots, f_{a,T}] \in \mathbb{R}^{2D \times T}$, $f_{a,t}$ indicates the temporal fusion feature of $\tilde{s}_{a,t}$, and θ_{fusion} denotes the parameters.

3.2.3 | Classification layer

A classifier $\Phi_{cls}(\cdot; \theta_{cls})$ is presented on F_a to predict the event scores of abnormal events for all input video segments. The classifier $\Phi_{cls}(\cdot; \theta_{cls})$ is a fully connected layer and the event scores can be obtained as:

$$P_a = \Phi_{cls}(F_a; \theta_{cls}), \tag{2}$$

where $P_a = [p_{a,1}, \ldots, p_{a,T}] \in \mathbb{R}^{C \times T}$, $p_{a,t}$ is the event scores, C is the total number of abnormal event categories and θ_{cls} denotes the parameters.

3.3 | Temporal attention module

Abnormal events occur infrequently and only exist in a very limited time slot if occurred. There is no detailed supervision about the exact temporal extent of abnormal events in the training dataset, which makes the problem of learning anomaly scores and abnormal event categories simultaneously very

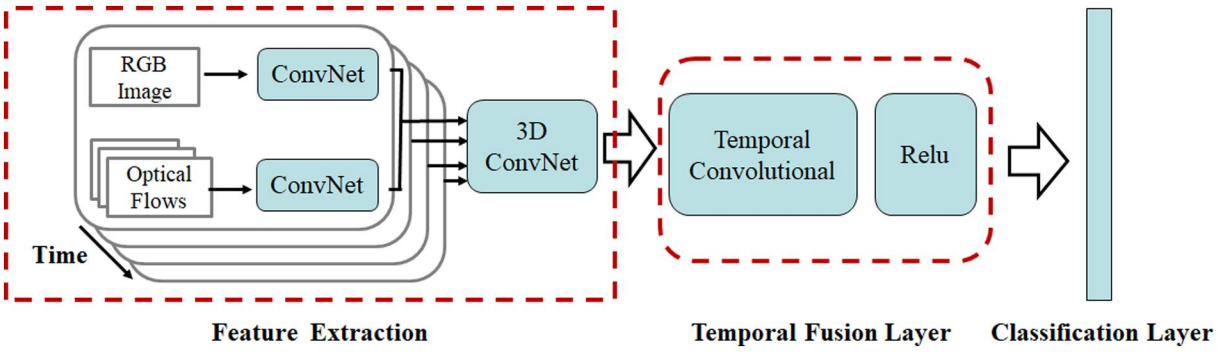

FIGURE 3 Illustration of Event Recognition Module: Feature extraction, temporal fusion layer and classification layer

challenging. To obtain reliable results, the TAM is proposed to localize video segments that abnormal events are most likely to exist in a training video.

Inspired by the trilinear attention mechanism [46], the TAM is exploited to explore the relationships of different abnormal event categories in a video. Specifically, given the event scores P_a from the ERM, the TAM is formulated as:

$$z_a = \Phi_{sum}(\mathcal{N}(\mathcal{N}(P_a)P_a^T)P_a), \tag{3}$$

where $z_a \in \mathbb{R}^T$ indicates the temporal attention values for video $\tilde{v}_a$, $\mathcal{N}(\cdot)$ denotes the *softmax* normalization function over the second dimension of a matrix, and P_a^{T} is the transposition of P_a. As shown in Figure 4, $\mathcal{N}(P_a)$ is the temporal normalization that keeps the event score of each segment in a video within the same scale. $\mathcal{N}(P_a)P_a^{\mathrm{T}}$ indicates the temporal relationship among different abnormal events, and $\mathcal{N}(\mathcal{N}(P_a)P_a^{\mathrm{T}})$ is the relationship normalization which processed on different abnormal events. Then, the temporal relationships are integrated by conducting dot production over $\mathcal{N}(\mathcal{N}(P_a)P_a^{\mathrm{T}})$ and P_a. Φ_{sum} is a sum pooling function over the first dimension of a matrix. Each value in z_a denotes an attention degree of a corresponding video segment containing abnormal events in $\tilde{v}_a$.

3.4 | Overall object function

To detect the abnormal events with only video-level labels, MIL is used to train the proposed method with positive (abnormal event) and negative (normal event) training bags. All bags are training videos with the same number of instances (video segments). Moreover, z_a can be adopted to find more exact positive and negative training instances from a video $\tilde{v}_a$. To be more specific, for a training video $\tilde{v}_a$, video segments with k^{top} largest temporal attention values in z_a are treated as positive instances, whose position index set is denoted as Ω_a^{top}. Further, video segments with k^{bot} smallest temporal attention values in z_a are treated as negative instances, whose position index set is denoted as Ω_a^{bot}. Based on Ω_a^{top} and Ω_a^{bot}, the positive and negative fusion features in $\tilde{v}_a$ can be obtained as:

$$f_a^{top} = mean(F_a[:, \Omega_a^{top}]), f_a^{bot} = mean(F_a[:, \Omega_a^{bot}]), \tag{4}$$

where $f_a^{top} \in \mathbb{R}^D$ and $f_a^{bot} \in \mathbb{R}^D$ indicate the positive and negative fusion features, respectively. $mean(\cdot)$ is the average function over the second dimension of a matrix.

Besides f_a^{top} and f_a^{bot}, the average event scores of positive and negative segments in $\tilde{v}_a$ can also be obtained as follows:

$$\begin{aligned} \hat{y}_a^{top} &= softmax(mean(P_a[:, \Omega_a^{top}])), \\ \hat{y}_a^{bot} &= softmax(mean(P_a[:, \Omega_a^{bot}])), \end{aligned} \tag{5}$$

where $\hat{y}_a^{top} \in \mathbb{R}^C$ and $\hat{y}_a^{bot} \in \mathbb{R}^C$ indicate the average event scores of positive and negative segments in $\tilde{v}_a$, respectively. Afterwards, based on f_a^{top}, f_a^{bot}, $\hat{y}_a^{top}$, $\hat{y}_a^{bot}$ of each video $\tilde{v}_a$ sampled in $\mathcal{V}_{anomaly}$ (positive instances), and f_n^{top}, f_n^{bot}, $\hat{y}_n^{top}$, $\hat{y}_n^{bot}$ of each video $\tilde{v}_n$ sampled in $\mathcal{V}_{normal}$ (negative instances), the overall object function contains three constraints: event category constraint, events separation constraint and temporal smoothness constraint. The details of these constraints are introduced as follows.

3.4.1 | Event category constraint

Event category constraint tries to learn the specific category of abnormal events from videos in $\mathcal{V}_{anomaly}$. Event category constraint $\mathcal{L}_{EEC}$ is formalized as:

$$\begin{aligned} \mathcal{L}_{EEC} = &-\frac{1}{N_{anomaly}} \sum_{u=1}^{N_{anomaly}} \sum_{c=1}^{C} y_{a;c} \log \hat{y}_{a;c}^{top} \\ &-\frac{1}{N_{anomaly} C} \sum_{u=1}^{N_{anomaly}} \sum_{c=1}^{C} \hat{y}_{a;c}^{bot}, \end{aligned} \tag{6}$$

where $y_{a;c}$ indicates the cth value in the ground truth y_a. $\hat{y}_{a;c}^{top}$ and $\hat{y}_{a;c}^{bot}$ indicate the cth value in $\hat{y}_a^{top}$ and $\hat{y}_a^{bot}$, respectively. Note that the first term in $\mathcal{L}_{AEC}$ is binary cross-entropy loss, which is adopted in this study to recognize the abnormal events. The second term can prevent the segments containing normal events from obtaining high event scores by forcing

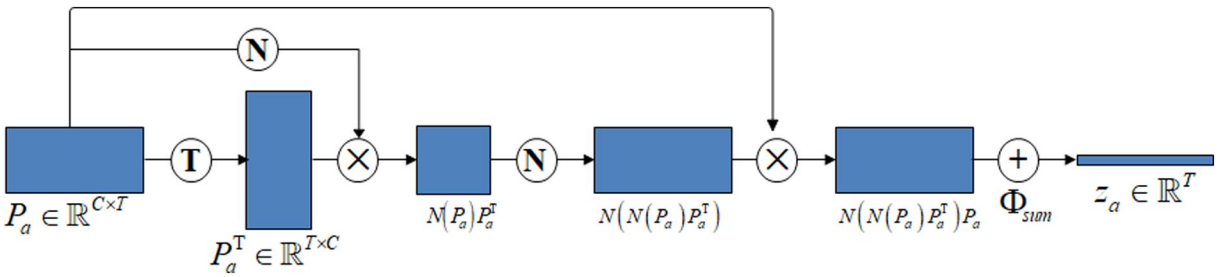

FIGURE 4 Illustration of Temporal Attention Module (TAM). The TAM inputs the event scores P_a and outputs an attention degree z_a. Each box indicates the shape of the matrix. T is matrix transposition, N is the normalization function, × is dot product and + is sum pooling function Φ_{sum} over the first dimension of a matrix

them in a uniform probability distribution for all abnormal event categories.

3.4.2 | Events separation constraint

During training, video segments in both $\mathcal{V}_{normal}$ and $\mathcal{V}_{anomaly}$ are treated as input. There are no abnormal events that occur in normal videos; thus, for a normal video $\tilde{v}_n$ in $\mathcal{V}_{normal}$, video segments with k^{top} largest temporal attention values in z_n are treated as hard negative instances. In addition, $\hat{y}_n^{top}$ and f_n^{top} are treated as hard negative event scores and fusion feature, respectively. Both $\hat{y}_n^{top}$ and f_n^{top} should be constrained carefully to avoid normal events being identified as abnormal events. Specifically, there are two sub-constraints: prediction separation constraint and feature separation constraint. Prediction separation constraint is designed to separate the event scores of abnormal events from normal events. Feature separation constraint is designed to separate the features of abnormal events from normal events. Concretely, prediction separation constraint $\mathcal{L}_{PSC}$ is formalized as:

$$\mathcal{L}_{PSC} = \frac{1}{N_{mix}} \sum_{a=1}^{N_{anomaly}} \sum_{n=1}^{N_{normal}} \max\Big(0, m_1 - \max(\hat{y}_a^{top}) + \max([\hat{y}_n^{top}; \hat{y}_a^{bot}])\Big), \tag{7}$$

where $N_{mix} = N_{normal} \times N_{anomaly}$, $\max(\cdot)$ can output the largest value of a matrix or vector. m_1 is a constant to make the maximal event scores of normal segments much lower than abnormal segments. $\mathcal{L}_{PSC}$ can make the event scores of positive instances in an abnormal video $\tilde{v}_a$ not only higher than the negative instances in $\tilde{v}_a$ but also higher than hard negative instances in a normal video $\tilde{v}_n$.

Besides, feature separation constraint $\mathcal{L}_{FSC}$ is formalized as:

$$\mathcal{L}_{FSC} = \frac{1}{N_{anomaly}} \sum_{a=1}^{N_{anomaly}} \max(0, m_2 - \|f_a^{top}\| + \|f_a^{bot}\|) + \frac{1}{N_{normal}} \sum_{n=1}^{N_{normal}} \|f_n^{top}\|, \tag{8}$$

where $\|\cdot\|$ indicates the L_2 normalization, m_2 is a constant to make the L_2 norm of abnormal fusion features much larger than normal fusion features. $\mathcal{L}_{nerc}$ can increase the dissimilarity between positive and negative instances in a video $\tilde{v}_a$ and increase the similarity between negative instances in $\tilde{v}_a$ and hard negative instances in $\tilde{v}_n$ simultaneously. Moreover, after adopting $\mathcal{L}_{FSC}$, for a video segment in $\tilde{v}_a$, the fusion feature magnitude of L_2 norm can be related to the confidence of anomaly, that is, the larger fusion feature norm indicating larger confidence with this segment being abnormal. In summary, events separation constraint $\mathcal{L}_{FSC}$ is expressed as follows:

$$\mathcal{L}_{ESCs} = \mathcal{L}_{PSC} + \alpha \mathcal{L}_{FSC}, \tag{9}$$

where α is the weight coefficient of $\mathcal{L}_{FSC}$.

3.4.3 | Temporal smoothness constraint

In surveillance videos, the temporal order of video segments is very important for abnormal event detection [14]. The adjacent video segments should vary smoothly. The temporal smoothness is enforced by minimizing the feature difference between adjacent video segments. Concretely, temporal smoothness constraint $\mathcal{L}_{TSC}$ is formalized as:

$$\mathcal{L}_{TSC} = \frac{1}{N_{anomaly}(T-1)} \sum_{a=1}^{N_{anomaly}} \sum_{t=1}^{T-1} (\|f_{a,t}\| - \|f_{a,t+1}\|)^2 + \frac{1}{N_{normal}(T-1)} \sum_{n=1}^{N_{normal}} \sum_{t=1}^{T-1} (\|f_{n,t}\| - \|f_{n,t+1}\|)^2, \tag{10}$$

where $\|\|$ indicate L_2 norm. $f_{a,t}$ indicates the feature of the tth video segment feature in an anomaly video, while $\|f_{v,t}\|$ indicates the feature of the tth video segment in a normal video.

In summary, the training objective $\mathcal{L}_{AL}$ (anomaly learning) is given by combing the above constraints together:

$$\mathcal{L}_{AL} = \mathcal{L}_{ECC} + \mathcal{L}_{ESCs} + \beta \mathcal{L}_{TSC}, \tag{11}$$

where β is the weight coefficient of $\mathcal{L}_{TSC}$.

3.5 | Inference

After training the TANet with the $\mathcal{L}_{AL}$, the anomaly scores and abnormal event categories of videos in the testing dataset can both be predicted. More specifically, let v_t denote a testing surveillance video that we know an abnormal event occurs in this video but do not know the specific category of the abnormal event. The event scores $P_t = [p_{t,1}, \ldots, p_{t,L_t}] \in \mathbb{R}^{C \times L_t}$ can be obtained by the ERM first, in which L_t is the number of video segments in v_t. Then, P_t is the input to the TAM to obtain the position index set of positive instances Ω_t^{top}. Afterwards, the average event scores of positive instances in v_t can be obtained as $\hat{y}_t^{top} = softmax(mean(P_t[:, \Omega_t^{top}]))$. Hence, the abnormal event category can be predicted as:

$$c_t = \arg \max_{1 \le c \le C} \hat{y}_t^{top}[c], \tag{12}$$

where c_t is the predicted abnormal event label.

In addition, let v_x denote a test surveillance video that we do not know whether there are abnormal events that occur in it. The anomaly scores of frames in v_x can be predicted as follows. First, the temporal fusion features $F_x = [f_{x,1}, \ldots, f_{x,L_x}]$

$\in \mathbb{R}^{C\times L_x}$ and event scores $P_x = [p_{x,1}, \ldots, p_{x,L_x}] \in \mathbb{R}^{C\times L_x}$ can both be obtained by the ERM, in which L_x is the number of video segments in v_x. Second, the norm of fusion features and maximal event scores for all video segments in v_x can be obtained as follows:

$$A_{x,1} = [\|f_{x,1}\|, \ldots, \|f_{x,L_x}\|], \tag{13}$$

$$A_{x,2} = [\Phi_{pick}(p_{x,1}), \ldots, \Phi_{pick}(p_{x,L_x})], \tag{14}$$

where $\Phi_{pick} = \max(softmax(\,))$ in this study. Φ_{pick} can pick the highest normalized value from the event scores of each video segment. Then, the anomaly score of lth video segment in v_x can be obtained as $score_{x,l} = A_{x,1}[l] * A_{x,2}[l]$. In addition, all frames in the same video segment share the same anomaly score. Based on the obtained anomaly scores of all frames in the test dataset, frames containing abnormal events can be predicted.

4 | EXPERIMENTS

Three parts are included in this section to explain the experiment settings in detail and prove the effectiveness of the proposed method. Subsection 4.1 introduces the dataset, implementation details, and evaluation metrics of the proposed method. Subsection 4.2 compares the proposed method with other methods. Subsection 4.3 displays the ablation study of the proposed method.

4.1 | Experimental settings

4.1.1 | Dataset

In this study, the UCF Crime dataset [14] is employed to evaluate the proposed method as described in Section 2. Compared with other datasets such as UCSD Peds [17], Avenue [47] and Subway Exit [19] datasets, the UCF Crime dataset is a larger and more realistic dataset; thus, it can be generalized well for real-world applications. To be more specific, the UCF Crime dataset consists of 1900 untrimmed real-world surveillance videos with 950 normal videos and 950 abnormal videos. Moreover, it has 13 types of real-world abnormal events, that is, abuse, arrest, arson, assault, accident, burglary, explosion, fighting, robbery, shooting, stealing, shoplifting, and vandalism. Figure 5 shows four frames of an example video from each abnormal event. The whole dataset is split into a training dataset and a testing dataset. The training set contains 810 abnormal videos and 800 normal videos. Besides, the testing set contains 150 normal videos and 140 videos that contain abnormal events. Furthermore, the training set only has video-level abnormal event labels while the testing set has both video-level labels and temporal annotations. Most videos in the UCF Crime dataset are very long and contain different scenarios, which make it very challenging for abnormal event detection and recognition.

4.1.2 | Implementation details

The I3D network [45] pretrained on the Kinetics dataset [45] is adopted as the backbone network to extract features in the proposed method. The TVL1 algorithm [48] is employed for generating optical flows of video frames in this study. Each video segment is designed to contain 24 frames. During training, T is set to 750, k^{top} and k^{bot} are set as $T/8$ and $T/6$, respectively. All hyperparameters are empirically determined by the grid search: $\alpha = 0.001$, $\alpha = 0.1$, $m_1 = 100$ and $m_2 = 0.5$. The Adam optimizer is used when training TANet, and the learning rate is set as 0.0001.

4.1.3 | Evaluation metrics

Following previous methods for abnormal event detection [14, 45], AUC, that is, the corresponding area under the ROC curve is employed to evaluate the performance of abnormal event detection. The recognition accuracy [45] is employed for the evaluation of the abnormal recognition.

4.2 | Comparison with other methods

4.2.1 | Abnormal event detection

Table 1 summarizes the abnormal event detection performances of the proposed method on the UCF Crime dataset compared with other methods. Hasan et al. [30] proposed a temporal regularity to learn normal events by reconstructing hand-crafted features and video frames with different deep auto-encoders. Lu et al. [47] proposed a sparse dictionary learning to reconstruct the 3D gradient features of frame patches with a combination of sparse bases. Both Hasan et al. [30] and Lu et al. [47] are unsupervised methods. Sultani et al. [14] proposed a multiple instance ranking framework for weakly supervised abnormal event detection. Zhu et al. [31] proposed a temporal augmented network as an auto-encoder to learn a compact feature of multiple optical flow maps for abnormal event detection. Dubey et al. [49] proposed a 3D deep MIL with ResNet to predict the abnormality score at the video segment level. Kamoona et al. [50] proposed a temporal encoding–decoding network to capture the temporal information between video instances.

As can be seen in Table 1, the proposed method achieves an AUC score of 76.5%, which outperforms all other methods listed in this table. More specifically, most videos in the UCF Crime dataset are very long and untrimmed in which abnormal events only occurred in a very short period.

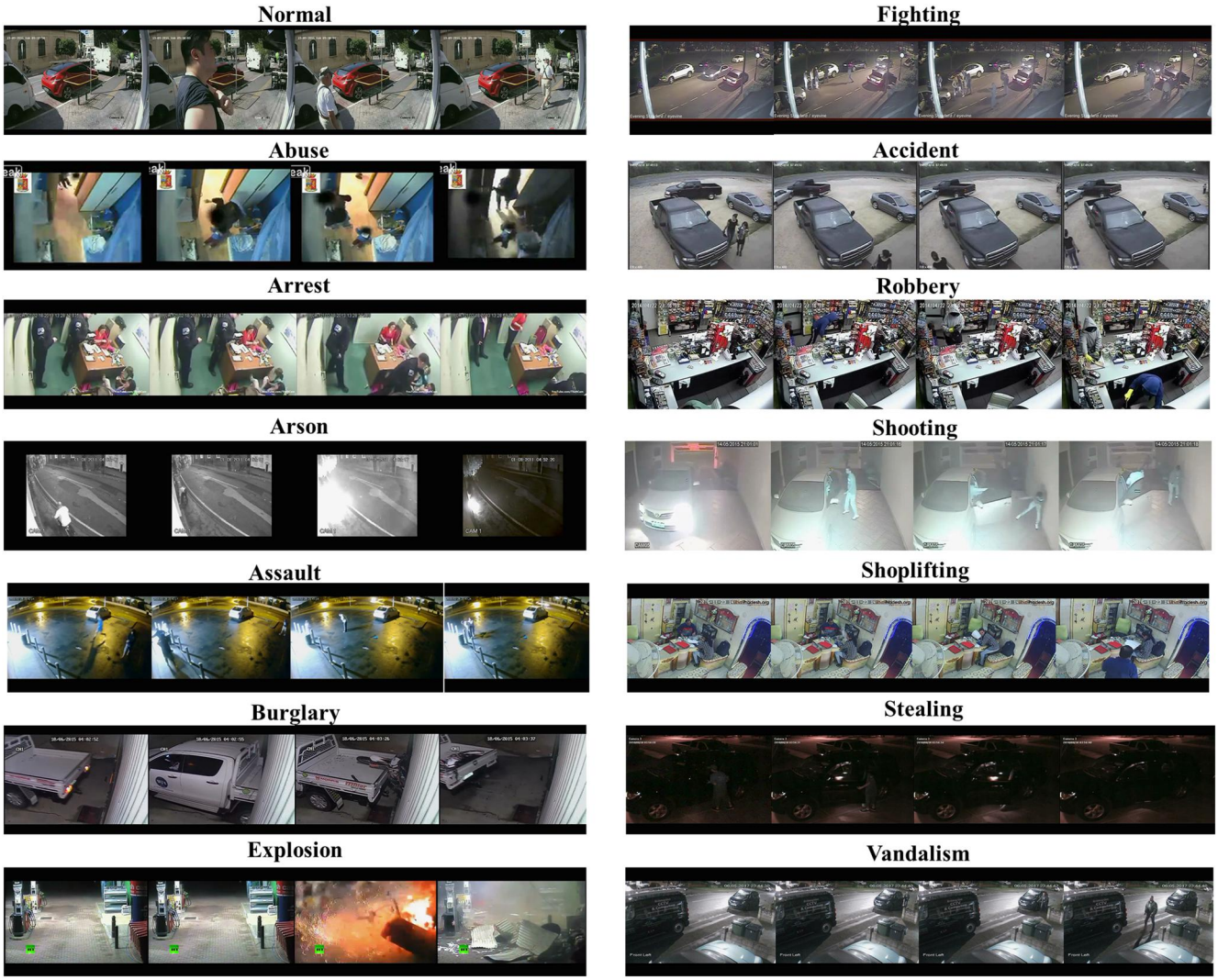

FIGURE 5 Examples of different abnormal events from the UCF Crime dataset. UGF, University of Central Florida

TABLE 1 Comparison with other methods on the abnormal event detection task

Methods	Supervision	AUC (%)
Hassan et al. [30]	Unsupervised	50.6
Lu et al. [47]	Unsupervised	65.5
Sultani et al. [14]	Weakly supervised	75.4
Zhu et al. [31]	Weakly supervised	72.1
Dubey et al. [49]	Weakly supervised	75.6
Kamoona et al. [50]	Weakly supervised	76.4
Our TANet	Weakly supervised	76.5

Moreover, different from other datasets, the background scenes in different videos of the UCF Crime dataset are quite various. Therefore, the hand-crafted features used in [30] are not discriminative enough to distinguish abnormal and normal events in the UCF Crime dataset. The dictionary learned in Ref. [47] could be limited to reconstructing normal and abnormal events in long untrimmed videos with small and large errors, respectively. MIL ranking [14] detected abnormal events by only the binary prediction scores. In Ref. [31], only information from optical flows is used to generate the motion-aware feature and features of RGB frames are ignored. However, visual information is also important for the detection of real-world abnormal events. The proposed method uses both the feature norm and the prediction scores for the abnormal event detection and thus can take advantage of more information in videos.

Moreover, the qualitative results of the proposed method on three videos are shown in Figure 6. Two major observations can be made from this figure. First, the proposed method can really provide high anomaly scores for abnormal events. Specifically, (b) is a video containing a shooting event, and the proposed method successfully predicts high anomaly scores for shooting frames. (c) is a video containing only normal frames, and all frames in (c) are predicted with low anomaly scores. Second, the exact start and end points

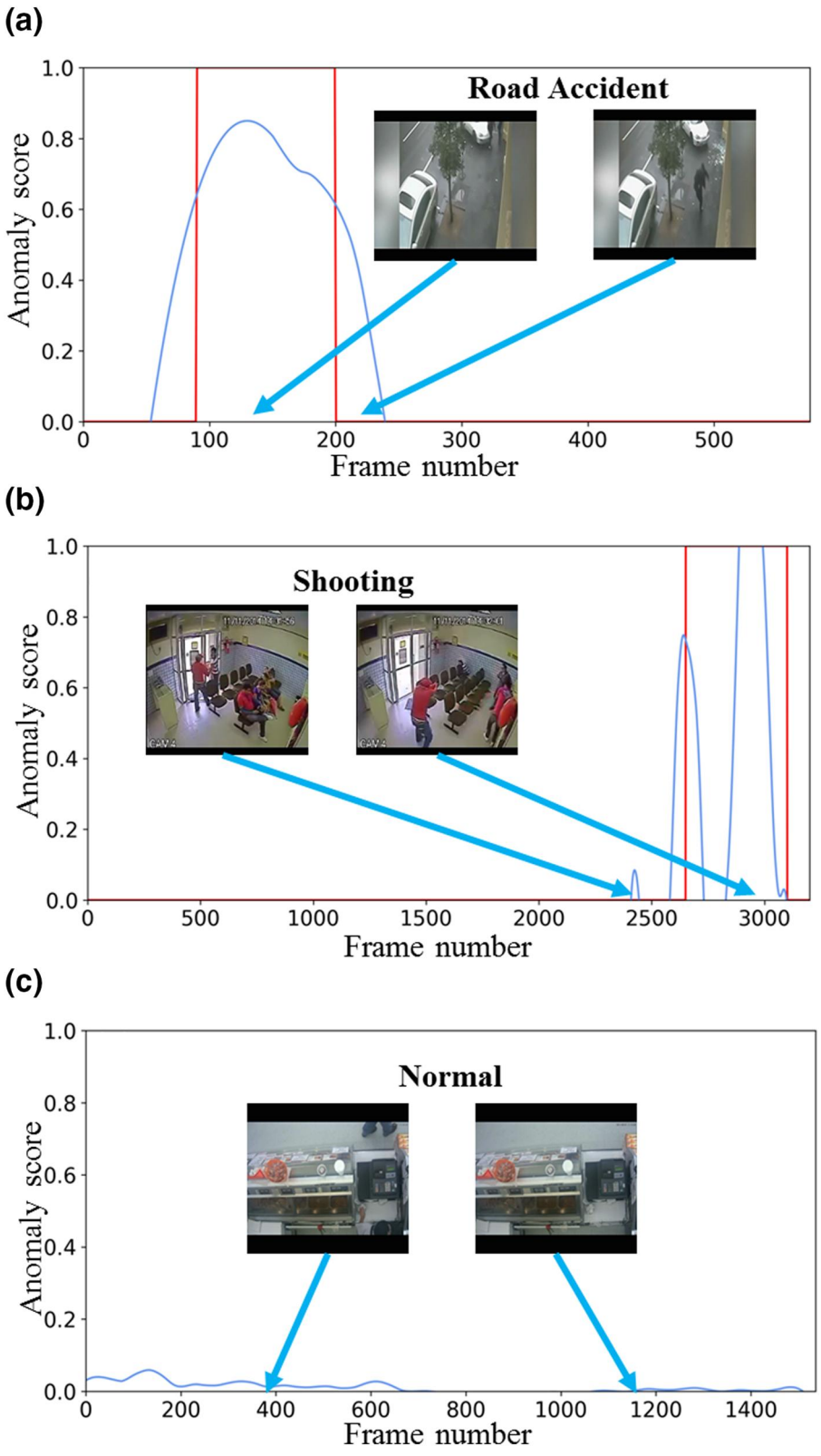

FIGURE 6 Qualitative abnormal event detection results of the proposed method on three testing videos. Red curves indicate the ground truth temporal regions of abnormal events. Blue curves indicate the anomaly scores obtained by the proposed method, which are smoothed in this figure for better visualization

still cannot be predicted for the reason of weakly supervised learning. For example, (a) is a video containing a car hitting a person, but the proposed method predicts high anomaly scores for both 'car hitting a person' and 'a person running after being hit by a car'.

4.2.2 | Abnormal event recognition

Table 2 summaries the abnormal event recognition performances of the proposed method compared with other methods on the UCF Crime dataset. In this table, 'C3D'

denotes the method of Tran et al. [43], which designed a 3D convolutional neural network for the video classification task. 'T-CNN' [51] denotes the method proposed for action detection in videos based on 3D convolution features. 'I3D' denotes the method of Carreira and Zisserman [45], which employed a two-stream 3D convolutional neural network for action recognition in videos and achieved state-of-the-art performance. Two observations can be made from Table 2. First, the proposed method can achieve better performance than other methods listed in this table, which proves the effectiveness of the proposed method. Second, all methods listed in this table are not able to achieve very promising performance on the UCF Crime dataset; this may be because most videos in the UCF Crime dataset are untrimmed and the backgrounds in videos of this dataset are too complex.

4.3 | Ablation study

To further study the influences of the TAM and the overall object function proposed in this study, experiments are performed on the UCF Crime dataset. The results are shown in Table 3. More specifically, when only $\mathcal{L}_{EEC}$ is adopted to train the proposed method with the TAM, the performance increases 0.9% AUC score compared without TAM, which can prove the effectiveness of the TAM. When $\mathcal{L}_{ESC}$ and $\mathcal{L}_{TSC}$ joined with $\mathcal{L}_{EEC}$, the performance of the proposed method can achieve better performances, which proves the effectiveness of the $\mathcal{L}_{ESC}$ and $\mathcal{L}_{TSC}$. Compared with $\mathcal{L}_{TSC}$, $\mathcal{L}_{ESC}$ can make a better performance. Moreover, when all loss functions are adopted, the performance of the proposed method can achieve an improvement of 2.0% AUC score compared with the method that only adopt $\mathcal{L}_{EEC}$ as constraint, which proves the effectiveness of the $\mathcal{L}_{ESC}$ and $\mathcal{L}_{TSC}$ again.

TABLE 2 Comparison with other methods on the abnormal event recognition task

Methods	Accuracy (%)
C3D [43]	23.0
T-CNN [51]	28.4
Motion-aware feature [31]	20.2
I3D [45]	29.2
Our TANet	31.5

Abbreviation: TANet, Temporal Attention Network.

TABLE 3 Ablation studies of the proposed method on the UCF Crime dataset

Methods	AUC (%)
$\mathcal{L}_{EEC}$	73.6
TAM + $\mathcal{L}_{EEC}$	74.5
TAM + $\mathcal{L}_{EEC}$ + $\mathcal{L}_{ESC}$	76.1
TAM + $\mathcal{L}_{EEC}$ + $\mathcal{L}_{TSC}$	75.6
TAM + $\mathcal{L}_{AL}$	76.5

Abbreviations: TAM, Temporal Attention Module; UGF, University of Central Florida.

5 | CONCLUSION

In this study, a method named as TANet is proposed to explore the idea of weakly supervised learning for abnormal event detection. Different from previous methods, the proposed method tries to learn both anomaly scores and specific abnormal event categories with only the supervision of video-level abnormal event labels, which can improve the generalization ability and save detailed human annotations. More specifically, the TAM in the TANet can learn abnormal attention degrees of all input video segments, which are subsequently used to locate abnormal and normal segments in training videos. In addition, an overall object function is proposed to guide TANet to learn discriminative representations and anomaly scores for abnormal event detection and recognition efficiently. Experiments on the UCF Crime dataset demonstrate that the proposed method can obtain outstanding performance compared with other methods on both abnormal event detection and classification tasks.

ACKNOWLEDGEMENTS

This work was supported in part by the National Science Fund for Distinguished Young Scholars under grant no. 61925112, in part by the National Natural Science Foundation of China under grant no. 61806193 and grant no. 61772510, in part by the Innovation Capability Support Program of Shaanxi under grant no. 2020KJXX-091, and in part by the Key Research Program of Frontier Sciences, Chinese Academy of Sciences under grant no. QYZDY-SSW-JSC044.

CONFLICT OF INTEREST

The authors declared that they have no conflicts of interest to this work.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in reference number [17].

ORCID

Xiangtao Zheng https://orcid.org/0000-0002-8398-6324

REFERENCES

1. Sodemann, A.A., Ross, M.P., Borghetti, B.J.: A review of anomaly detection in automated surveillance. IEEE Trans. Syst. Man Cybern. Syst. 42(6), 1257–1272 (2012)
2. Jiang, R., et al.: Flow-assisted visual tracking using event cameras. CAAI Trans. Intell. Technol. 6(2), 192–202 (2021). https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cit2.12005
3. Xing, Y., Zhu, J.: Deep learning-based action recognition with 3d skeleton: a survey. CAAI Trans. Intell. Technol. 6(1), 80–92 (2021)
4. Hassner, T., Itcher, Y., Kliper-Gross, O.: Violent flows: real-time detection of violent crowd behavior. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–6. (2012)

5. Yang, J., et al.: A two-branch network with pyramid-based local and spatial attention global feature learning for vehicle re-identification. CAAI Trans. Intell. Technol. 6(1), 46–54 (2021)
6. Yuan, Y., Feng, Y., Lu, X.: Structured dictionary learning for abnormal event detection in crowded scenes. Pattern Recogn. 73, 99–110 (2018)
7. Oluwatoyin, P.P., Wang, K.: Video-based abnormal human behavior recognition - a review. IEEE Trans. Syst. Man Cybern. Syst. 42(6), 865–878 (2012)
8. Luo, W., Liu, W., Gao, S.: A revisit of sparse coding based anomaly detection in stacked RNN framework. In: IEEE International Conference on Computer Vision, pp. 341–349. (2017)
9. Reddy, V., Sanderson, C., Lovell, B.C.: Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 55–61. (2011)
10. Zhao, B., Li, F., Xing, E.P.: Online detection of unusual events in videos via dynamic sparse coding. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3313–3320. (2011)
11. Xu, D., et al.: Learning deep representations of appearance and motion for anomalous event detection. In: British Machine Vision Conference, pp. 81–812. (2015)
12. Taha, A., Hadi, A.S.: Anomaly detection methods for categorical data: a review. ACM Comput. Surv. 52(2), 381–3835. (2019)
13. Mao, X., et al.: Least squares generative adversarial networks. In: IEEE International Conference on Computer Vision, pp. 2813–2821. (2017)
14. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6479–6488. (2018)
15. Mahadevan, V., et al.: Anomaly detection in crowded scenes. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, pp. 1975–1981. (2010)
16. Ramachandra, B., Jones, M.J.: Street scene: a new dataset and evaluation protocol for video anomaly detection. In: IEEE Winter Conference on Applications of Computer Vision, pp. 2558–2567. (2020)
17. Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization in crowded scenes. IEEE Trans. Pattern Anal. Mach. Intell. 36(1), 18–32 (2014)
18. Basharat, A., Gritai, A., Shah, M.: Learning object motion patterns for anomaly detection and improved object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. (2008)
19. Adam, A., et al.: Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Trans. Pattern Anal. Mach. Intell. 30(3), 555–560 (2008)
20. Feng, Y., Yuan, Y., Lu, X.: Deep representation for abnormal event detection in crowded scenes. In: ACM Conference on Multimedia Conference, pp. 591–595. (2016)
21. Ravanbakhsh, M., et al.: Abnormal event detection in videos using generative adversarial nets. In: IEEE International Conference on Image Processing, pp. 1577–1581. (2017)
22. Wu, S., Moore, B.E., Shah, M.: Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2054–2060. (2010)
23. Yuan, Y., Feng, Y., Lu, X.: Statistical hypothesis detector for abnormal event detection in crowded scenes. IEEE Trans. Cybern. 47(11), 3597–3608 (2017)
24. Cui, X., et al.: Abnormal detection using interaction energy potentials. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3161–3167. (2011)
25. Anjum, N., Cavallaro, A.: Multifeature object trajectory clustering for video analysis. IEEE Trans. Circ. Syst. Video Technol. 18(11), 1555–1564 (2008)
26. Xu, D., et al.: Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts. Neurocomputing. 143, 144–152 (2014)
27. Xu, Y., Qiu, T.T.: Human activity recognition and embedded application based on convolutional neural network. J. Artif. Intell. Res. 1(1), 51–60 (2020). https://ojs.istp-press.com/jait/article/view/6
28. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. In: International Conference on Learning Representations (2016)
29. Liu, W., et al.: Future frame prediction for anomaly detection - a new baseline. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6536–6545. (2018)
30. Hasan, M., et al.: Learning temporal regularity in video sequences. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 733–742. (2016)
31. Zhu, Y., Newsam, S.D.: Motion-aware feature for improved video anomaly detection. In: British Machine Vision Conference, pp. 270–282. (2019)
32. Perez, M., Kot, A.C., Rocha, A.: Detection of real-world fights in surveillance videos. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2662–2666. (2019)
33. Ren, S., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39(6), 1137–1149 (2017)
34. Saha, S., et al.: Deep learning for detecting multiple space-time action tubes in videos. In: Wilson, R.C., Hancock, E.R., Smith, W.A.P. (eds.) Proceedings of the British Machine Vision Conference, pp. 581–5813. (2016)
35. Wang, L., et al.: Untrimmednets for weakly supervised action recognition and detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6402–6411. (2017)
36. Lee, P., Uh, Y., Byun, H.: Background suppression network for weakly-supervised temporal action localization. In: AAAI Conference on Artificial Intelligence, pp. 11320–11327. (2020)
37. Peng, X., Schmid, C.: Multi-region two-stream R-CNN for action detection. In: Leibe, B., et al. (eds.) European Conference on Computer Vision, vol. 9908, pp. 744–759. (2016)
38. Escorcia, V., et al.: Daps: deep action proposals for action understanding. In: Leibe, B., et al. (eds.) European Conference on Computer Vision, vol. 9907, pp. 768–784. (2016)
39. Kalogeiton, V., et al.: Action tubelet detector for spatio-temporal action localization. In: IEEE International Conference on Computer Vision, pp. 4415–4423. (2017)
40. Richard, A., Kuehne, H., Gall, J.: Weakly supervised action learning with RNN based fine-to-coarse modeling. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1273–1282. (2017)
41. Nguyen, P., et al.: Weakly supervised action localization by sparse temporal pooling network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6752–6761. (2018)
42. Islam, A., Radke, R.J.: Weakly supervised temporal action localization using deep metric learning. In: IEEE Winter Conference on Applications of Computer Vision. IEEE, pp. 536–545. (2020)
43. Tran, D., et al.: Learning spatiotemporal features with 3d convolutional networks. In: IEEE International Conference on Computer Vision, pp. 4489–4497. (2015)
44. Zhong, J., et al.: Graph convolutional label noise cleaner: train a plug-and-play action classifier for anomaly detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1237–1246. (2019)
45. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4724–4733. (2017)
46. Zheng, H., et al.: Looking for the devil in the details: learning trilinear attention sampling network for fine-grained image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5012–5021. (2019)
47. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 FPS in MATLAB. In: Proc. IEEE International Conference on Computer Vision, pp. 2720–2727. (2013)

48. Wedel, A., et al.: An improved algorithm for tv-l1 optical flow. In: Statistical and Geometrical Approaches to Visual Motion Analysis, ser. Lecture Notes in Computer Science, vol. 5604, pp. 23–45. (2009)
49. Dubey, S., Boragule, A., Jeon, M.: 3d resnet with ranking loss function for abnormal activity detection in videos. In: 2019 International Conference on Control, Automation and Information Sciences (ICCAIS). IEEE, pp. 1–6. (2019)
50. Kamoona, A.M., et al.: Multiple instance-based video anomaly detection using deep temporal encoding-decoding. arXiv preprint arXiv:2007.01548 (2020)
51. Hou, R., Chen, C., Shah, M.: Tube convolutional neural network (T-CNN) for action detection in videos. In: International Conference on Computer Vision, pp. 5823–5832. (2017)

How to cite this article: Zheng, X., et al.: Abnormal event detection by a weakly supervised temporal attention network. CAAI Trans. Intell. Technol. 7(3), 419–431 (2022). https://doi.org/10.1049/cit2.12068