Expert Systems With Applications 201 (2022) 117030

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A multi-stream deep neural network with late fuzzy fusion for real-world anomaly detection

Thakare Kamalakar Vijay [a], Nitin Sharma [a], Debi Prosad Dogra [a,*], Heeseung Choi [b], Ig-Jae Kim [b]

[a] Indian Institute of Technology Bhubaneswar, India
[b] Korea Institute of Science and Technology, South Korea

ARTICLE INFO

Keywords:
Multi-stream networks
Fuzzy fusion
Anomaly detection
Event classification
Multiple instance learning

ABSTRACT

Abnormal event detection in video is alternatively known as outlier detection, where machine learning can be highly effective. While testing an unknown video, the objective of such methods is to verify the video's category, e.g. normal or abnormal. This paper exploits visual information from normal as well as abnormal videos to train a deep multiple instance learning classifier for classification of videos. Existing multiple instance learning classifiers presume that the training videos contain only short-duration anomalous events. This assumption may not be valid for all real-world anomalies. Also, multiple occurrences of anomalies in training videos cannot be ruled out. This paper shows that by injecting temporal information in feature extraction, anomaly detection performance can be improved. To accomplish this, two spatio-temporal deep feature extractors have been applied in parallel on the training videos. These streams are then used to train a modified multiple instance learning-based classifier. Finally, a fuzzy aggregation is employed to fuse the anomaly scores. Additionally, two lightweight deep-learning classifiers have been used to substantiate the model's efficacy for classifying fire and accident events. To understand the reliability and performance of the proposed method, extensive experiments have been carried out using UCF-Crime video dataset containing 13 anomaly categories. The dataset has been restructured into five broad categories based on the severity of actions to study the robustness of the proposed method. The paper provides sufficient empirical evidence which proves that by incorporating temporal features in the pipeline, anomaly detection accuracy can be significantly improved. Moreover, the model helps to detect long-duration anomalies in videos, which was not possible using existing methods. The proposed end-to-end multi-stream architecture performs abnormal event detection with accuracy as high as 84.48%, which is better than the performance of existing video anomaly detection methods. Moreover, the class-wise detection accuracy has improved by 6%–14% across various broad categories.

1. Introduction

Anomaly detection with localization in surveillance videos is challenging for certain reasons, including the subjectivity in defining real-world anomalies and unavailability of annotated data. Moreover, anomalies are highly contextual and less frequently occurring than normal activities. Some anomalies like shooting and stealing often contain very subtle changes with low-level interactions between the objects. On the contrary, arson and explosion events contain drastic environmental changes that make it difficult to identify them manually in long duration videos.

Abnormal or anomalous events occur rarely and moreover visual patterns depicting anomalies are different from normal events. Also, universal grouping of normal patterns is challenging as the inter-class variations can be very high. Unary-classification, commonly known as one-class classification has already been exploited in Sparse Combination Learning (SCL) (Lu et al., 2013) and Dynamic Texture Modeling (Li et al., 2014). The main aim of one-class classification paradigm is to encode normal behavioral features extracted from non-anomalous videos and consider an input video to be abnormal if the features are different. However, due to the unavailability of all possible normal samples, it is challenging to encode the normal features universally. Some normal events may deviate from encoded features and may cause false alarms.

Over the past few years, some research works such as Multiple Instance Learning (MIL) (Sultani et al., 2018) and attention-aware

* Corresponding author.
E-mail addresses: tkv15@iitbbs.ac.in (K.V. Thakare), ns22@iitbbs.ac.in (N. Sharma), dpdogra@iitbbs.ac.in (D.P. Dogra), hschoi@kist.re.kr (H. Choi), drjay@kist.re.kr (I.-J. Kim).

https://doi.org/10.1016/j.eswa.2022.117030
Received 14 January 2021; Received in revised form 14 January 2022; Accepted 27 March 2022
Available online 7 April 2022

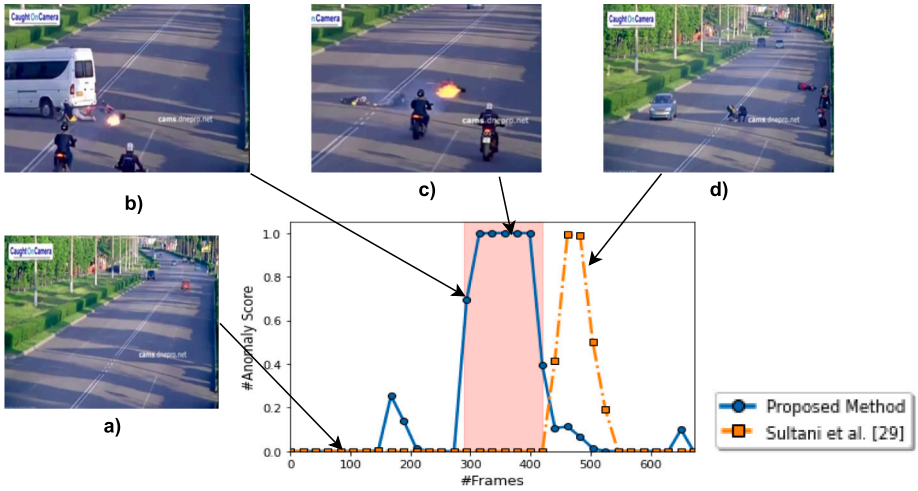

Fig. 1. An example of anomaly detection (Road Accident) taken from the UCF-Crime (Sultani et al., 2018). (a) A normal scenario, where cars and bikes are moving on the road. (b) Suddenly, an accident takes place. (c) The bike rider falls down and the bike caught into flames. (d) People rush to help the biker. Corresponding graphical representations of the anomaly scores predicted by the proposed method and the MIL-based method (Sultani et al., 2018). The shaded regions in the graph represent the ground truths. It may be observed that the anomaly scores generated by the proposed method are mostly agreeing with the ground truths. Though Sultani's approach predicts the anomalous event with a high score, however, it detects anomaly only when the people are running to help the biker after the accident event. This implies that the features and the fusion strategy used in this work perform better as compared to the work proposed in Sultani et al. (2018).

features (Zhu & Newsam, 2019) have been proposed to addressed this problem using binary classification paradigms, where feature encoding is done over normal and abnormal samples. The work (Sultani et al., 2018) follows a bag-of-segments method that divides the training dataset into positive (anomalous) and negative (normal) bags and train the model using Multiple Instance Learning (MIL)-inspired hinge-based loss. Following this, Zhu and Newsam (2019) have introduced attention-aware features into the MIL framework to improve anomaly detection performance. Though these approaches leverage both normal and abnormal feature encoding, the distinctive capabilities are not appealing. First, they are unable to capture the discriminating spatio-temporal features essential for detecting subtle anomalies like stealing, robbery, shooting, etc. Secondly, the boundary between anomalous and normal events is often unclear. The same behavior can be normal or anomalous under different circumstances. Therefore, proper utilization of visual features at video-level and robust feature encoding at training-level are needed for accurate anomaly detection. This paper focuses on how to obtain stronger visual features by intelligently fusing motion information during the training process.

In order to detect anomalous events more correctly, proper utilization of different types of inputs can be beneficial. Change in motion information (optical flow) can be highly informative in defining the behavior of an object. On the other hand, visual appearance (RGB features) can be useful to obtain spatial information. In each stream (pipeline), different type of input features are processed independently. In later stage, outputs of these multiple streams are fused together to produce final prediction (Simonyan & Zisserman, 2014). Therefore, this paper focus on how to obtain stronger visual features by intelligently fusing temporal information in a multi-stream network. Existing work such as graph convolution network (Zhong et al., 2019), MIL (Sultani et al., 2018), Temporal Segment Network (Wang et al., 2016) use either image-based features or features learned through deep learning to detect anomalies. It may be noted that irrespective of the depth of the networks and feature dimension, volume-based features with integrated motion information usually perform better than image-based features. This makes sense as multiple occurrences of irregular activity patterns and motion-aware features combined together can be more effective for detecting anomalous events. It is also desirable that the detection algorithm does not rely on any prior information about an event. Keeping these points in mind, this paper proposes a multi-stream network that exploits motion information to encode normal as well as abnormal features. Moreover, two additional light-weight binary classifiers have been introduced to detect fire and road accident events. An example is depicted in Fig. 1, where a road-accident is correctly being classified using the proposed fusion-based method. On the other hand, MIL-based method (Sultani et al., 2018) fails to localize the event correctly. An example is depicted in Fig. 1, where a road-accident event is correctly being classified using the proposed fusion-based method, whereas the MIL-based method (Sultani et al., 2018) fails to localize the event correctly.

1.1. Contributions of the paper

Anomaly detection can be challenging for various reasons. Firstly, the definition of anomaly is dependent on how normality is modeled. Also, a robust detection system needs to adapt to the changing dynamics of a scene. Moreover, anomalous events are generally infrequent, sparse, and unpredictable. Thus, preparing training data becomes difficult. Therefore, in this work, weakly-labeled videos have been used in training. The videos are categorized as normal or anomalous before the training process. However, their frame-level locations are completely unknown. Moreover, the training dataset have been restructured into few subcategories to study the effectiveness of the inclusion of motion information. In accomplishing this, following technical contributions have been made:

- This paper presents end-to-end multi-stream network architecture that acts an expert system to perform anomaly detection in surveillance videos. Moreover, extensive experiments have been performed on particular events related to explosions and road accidents.
- The primary focus of this paper is to obtained stronger visual features by intelligently fusing the motion and spatial information. An improvement of MIL classifier has been proposed to accommodate feature variations.
- A fuzzy aggregation method has been proposed to combine the scores of various independent feature streams to improve the segment detection accuracy in long duration videos.
- Finally, a light-weight two-class classifier (TCC) has been used to detect anomalous segments based on severity of actions. This has been tested on two types of critical anomalies, namely road accident and fire-related incidents.

Table 1
Latest notable contributions in video anomaly detection.

References	Features	Method	Observation
Hasan et al. (2016)	HOG+HOF	Fully connected autoencoders	Leverage hand-crafted features.
Luo et al. (2017b)	TSC	Sparse-encoding using stacked RNN	Not suitable for longer videos.
Ionescu et al. (2017)	VGG-fc7	Unmasking abnormal videos	Leverage Frame annotations
Sultani et al. (2018)	C3D	Multiple Instance Learning (MIL)	Lack of motion information.
Sabokrou et al. (2018)	Pre-trained FCN	Deep-anomaly: FCN	Generates high false alarms.
Xu et al. (2018)	Foreground point	Stacked sparse coding	FIP features are computed block-wise.
Li and Chang (2019)	Gradient maps	Gaussian autoencoder (MGFC-AAE)	Depends on the autoencoder's error.
Zhong et al. (2019)	TSN	Graph convolution network	Lack of motion information.
Zhou et al. (2019)	ResNet-152	Iterative hard-thresholding	Not suitable for longer videos.
Pang et al. (2020)	ResNet-50	Self-trained ordinal regression	Method works on spatial features only.

The rest of the paper is organized as follows. Related work is presented in Section 2. The proposed multi-stream network and training process is explained in Section 3. Section 4 presents the datasets and experimental evaluations. Discussions are presented in Section 5. In Section 6, the conclusion and future scopes of the work has been discussed.

2. Related work

The ability to identify anomalous or irregular behavior has applications in detecting traffic violations, crime, harmful, and illegal activities. There are a few notable contributions on video surveillance guided identification of abnormal events including angry crowd (Mohammadi et al., 2016), human aggression (Kooij et al., 2016) and violence (Gao et al., 2016). Table 1 depicts several notable contributions in video anomaly detection task. The few unsupervised learning methods such as unmasking technique previously used in author verification task (Ionescu et al., 2017), anomaly detection without temporal ordering (Giorno et al., 2016) encode the normal pattern using only normal samples for training to classify an event as anomalous when the test pattern deviates from the encoded patterns. In contrast, weakly-supervised methods such as MIL (Sultani et al., 2018), attention-aware features (Zhu & Newsam, 2019), complementary inner bag loss method (Zhang et al., 2019) and graph-based convolution (Zhong et al., 2019) learn more robust features from normal and abnormal training samples. Existing weakly-supervised methods can be divided into two categories, i.e., classifier-centered methods and feature-centered methods. (i) Classifier-centered methods focus on training of the classifier. The authors in Sultani et al. (2018) have introduced an anomaly detection classifier based on MIL, where a deep anomaly ranking model can predict scores of segments for both anomaly and normal videos. Following their approach, Yi Zhu et al. have incorporated motion information to enhance the efficiency of the MIL model (Zhu & Newsam, 2019). Few researchers (Zhang et al., 2019) have also introduced inner-bag score gap regularization. (ii) Feature-centered methods train the classifier and the feature extractor alternatively. The authors in Zhong et al. (2019) have formulated the problem using learning by noisy labels and graph-based convolution neural networks. Similarly, Jiang et al. have proposed regularized DNN (rDNN) that exploits class and feature relationships in video categorization (Jiang et al., 2018). Some recent work focuses on potential usage of weak noisy-labels (Lu et al., 2017) to improve the quality of noisy labels by avoiding pixel-level annotation. The authors of Yao et al. (2017) have introduced the Contrastive-Additive Noise Networks (CAN) to reduce the noise-label effects. Though training the classifier models using noisy-labels can be appealing, the underlying structure of the classifier often ignores the quality of spatio-temporal features. Moreover, as reported in Diba et al. (2017), it is difficult to improve the accuracy further by only using volume-based features as it fails to capture the subtle changes in the anomaly scene. Hence integrating the volume-based features with motion features, better results can be obtained (Simonyan & Zisserman, 2014).

Table 2
Notations and symbols used in the text.

Mathematical notations	Descriptions
n	Total number of frames in a video
k	Total number of segments in a video
(n/k)	Size of each segment
N	Total number of instances per bag
X	Bag (Set) of total N instances
y	Segment label, 0 or 1
B_{ma}	Bag contains anomalous instances
B_{la}	Bag contains normal instances
λ	Temporal smoothness factor
ω_i	ith weight used in broadcasting the anomaly score

The proposed method also leverages weakly-supervised data similar to MIL-based anomaly detection (Sultani et al., 2018) and attention-aware feature encoding technique (Zhu & Newsam, 2019). However, three notable differences are: (i) The proposed method integrates motion information along with spatial information. The motion information is obtained using SelFlow (Liu et al., 2019). (ii) A multi-stream architecture is designed for complex feature integration. (iii) A novel fuzzy aggregation is used during the final fusion of the scores.

3. Proposed work

This section presents the details of contributions made by this paper. A new anomaly localization framework as depicted in Fig. 2, has been proposed. The method is divided into three stages, namely (i) feature extraction (ii) anomaly *detection* (iii) anomaly *localization* using fusion.

3.1. Mathematical notations

The paper uses various mathematical notations and symbols to explain the methods and algorithms. All important notations and symbols have been summarized in Table 2.

3.2. Multi-stream feature extraction

Authors in Sultani et al. (2018) and Zhu and Newsam (2019) have implemented a trainable single stream architecture that perfectly works with Conv3Net C3D features and MIL classifier with some modification. Instead of using one stream for learning the context as done in GCN-based method (Zhong et al., 2019) or MIL-based method (Sultani et al., 2018), this paper uses two parallel deep neural networks. One is used for learning the spatial context (pre-trained), and the other one is used to learn the motion context. Moreover, the proposed method leverages over *weakly labeled* data. This implies all videos are marked either as anomalous or normal, but the exact location of anomaly is not known. Due to this, it is referred to as a semi-supervised approach. Each video is divided into k segments. For example, let the number of frames in a video is n. The number of segments $k = 32$ (k can be set as required). Then, the size of each segment becomes (n/k) frames. All segments of a video constitute a bag. Since Conv3D uses 16 frames

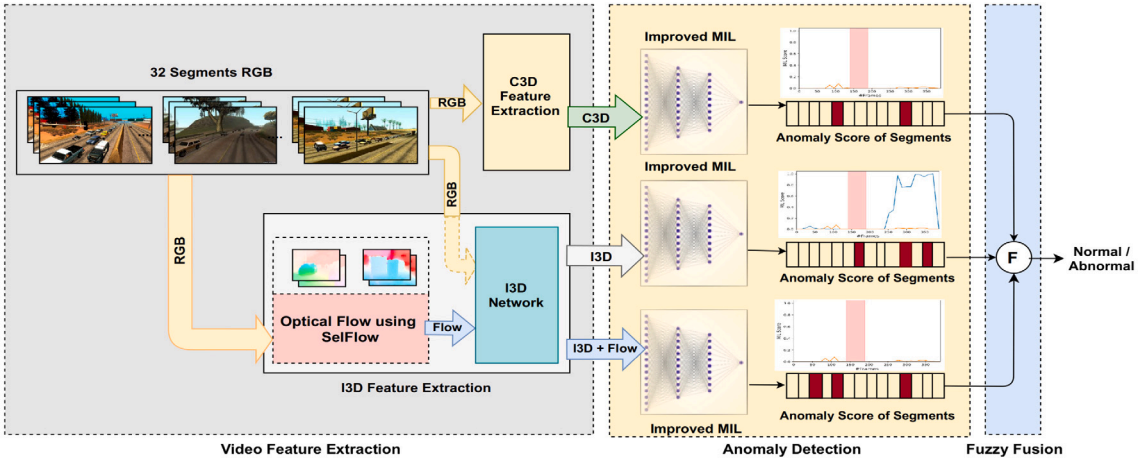

Fig. 2. The proposed end-to-end multi-stream anomaly localization and classification architecture. In the feature extraction stage, C3D (Tran et al., 2015) and I3D (Carreira & Zisserman, 2017) have been used to extract features independently. SelFlow (Liu et al., 2019) has been used to expedite the process of flow extraction in I3D. In the anomaly detection layer, MIL has been used with an improved loss function. This improved MIL then generates an anomaly score for each segment, the graph shown is segments (X-axis) and the corresponding anomaly score (Y-axis). The scores obtained from various streams are then fused using a fuzzy aggregation method and the anomalous segments are localized.

as the default depth, each video segments have been clipped into 16 frames. The C3D features for each segment are then calculated. Finally, an average is taken of all the clips to obtain the segment's features. A bag containing anomalous video segment is labeled as positive and a bag with normal video segment is labeled as negative. Now, using both positive (anomalous) and negative (normal) bags, the anomaly detection model is trained.

Following this (Sultani et al., 2018), the proposed framework extend the feature extraction module into a multi-stream network. In the present multi-streams architecture, C3D (Tran et al., 2015) and I3D (Carreira & Zisserman, 2017) are used for feature extraction. A homogeneous architecture with a small $3 \times 3 \times 3$ convolution kernels in all layers have been found to be the best performing architecture for 3D ConvNets. Similarly, I3D also uses a two-stream based 3D-ConvNet to utilize the training weights from the pre-trained 2D-inception networks. I3D can be used to encode both RGB frame-based vectors as well as optical-flow data, simultaneously. For the optical flow encoding, it requires pre-computed flow vectors in horizontal as well as vertical directions. The proposed architecture utilizes a tightly coupled SelFlow (Liu et al., 2019) with the I3D flow-based pre-trained module to speed up the whole process. In this work, the visual features are computed from the fully connected (FC) layer FC6 of the C3D network and mixed_5C layer of the I3D network. The resultant encoded vectors are of 4096 and 1024 dimensions, respectively. The addition of extracted and encoded flow information using SelFlow and I3D has helped the network to learn motion and appearance-invariant features. All single streams are independently trained with different features.

3.3. Anomaly detection

Binary classification is a supervised learning problem in which the model aims to find a model that predicts a value of a target variable, say $y \in \{0, 1\}$, for the given input variable, say $x \in \mathbb{R}$. However, MIL assumes that there is a bag of instances, $X = (x_1, x_2, \ldots, x_N)$, where N is total number of instances in the bag. There is no inter-dependency between these instances. There is also a binary label y associated with the bag. Moreover, a label exists for the instance within a bag such that $y_1, y_2, \ldots, y_N$, and $y_j \in \{0, 1\}$ for $j = 1, 2, \ldots, N$. The above formulation can be expressed using (1).

$$Y = \begin{cases} 0, & \text{iff } \sum_N y_j = 0 \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

Following weakly-supervised work such as attention-aware feature encoding (Zhu & Newsam, 2019), the segment-level annotations have been obtained. Finally, a ranking loss as given in (2) is used to train the model,

$$f(S_{ma}) > f(S_{la}) \tag{2}$$

where S_{ma} and S_{la} are more anomalous and less anomalous segments. $f(S_{ma})$ and $f(S_{la})$ represent the corresponding predicted anomaly scores consist of binary value between 0 and 1. However, only video-level annotations have been accessed. Therefore, the MIL ranking loss has been modified to (3),

$$\max_{i \in \mathcal{B}_{ma}} f(S^i_{ma}) > \max_{i \in \mathcal{B}_{la}} f(S^i_{la}) \tag{3}$$

where S^i_{ma} and S^i_{la} are ith segments of relatively more anomalous videos and less anomalous video, respectively. Also, max is taken over all video segments in each bag. Relatively anomalous video are represented as a positive bag $\mathcal{B}_{ma}$, where different temporal segments are individual instances in the bag e.g. $(ma_1, ma_2, \ldots, ma_m)$, where m is the number of instances in the bag. Similarly, a less anomalous video is kept in the negative bag $\mathcal{B}_{la}$, where different temporal segments are individual instances in the bag (e.g. $la_1, la_2, \ldots, la_m$). It has been assumed that at least one of these instances contain anomaly. This ranking works well since an instance in the positive bag with the highest anomaly scores usually ranks higher than the instance with the highest anomaly score in the negative bag. This happens as there is no anomaly in a negative bag. In order to keep a large margin between the positive and negative instances, the authors of MIL-based method (Sultani et al., 2018) have introduced a hinge-based ranking loss as given in (4).

$$l(\mathcal{B}_{ma}, \mathcal{B}_{la}) = \max\left(0, 1 - \max_{i \in \mathcal{B}_{ma}} f(S^i_{ma}) + \max_{i \in \mathcal{B}_{la}} f(S^i_{la})\right) \tag{4}$$

However, the relation given in (2) lacks to capture the underlying temporal structure in anomalous videos. This is due to the use of a simple *max* function. The anomaly score typically vary smoothly across the video segments. However, the scores of instances (or segments) in the anomalous bag can be small depending on the situations. Moreover, the anomaly may be present only in some segments. Thus, authors in Sultani et al. (2018) have introduced temporal smoothness and sparsity restriction between frames on the anomalous scores to mitigate the above problems. However, a video may contain several anomalous segments. It has been observed that putting any restriction on the temporal smoothness between adjacent segments can be misleading. To overcome this, an explicit addition of motion information during MIL training for encoding the temporal information has been introduce. In the revised MIL ranking loss estimation, the error that is back-propagated to the network, is nothing but the maximum-scored

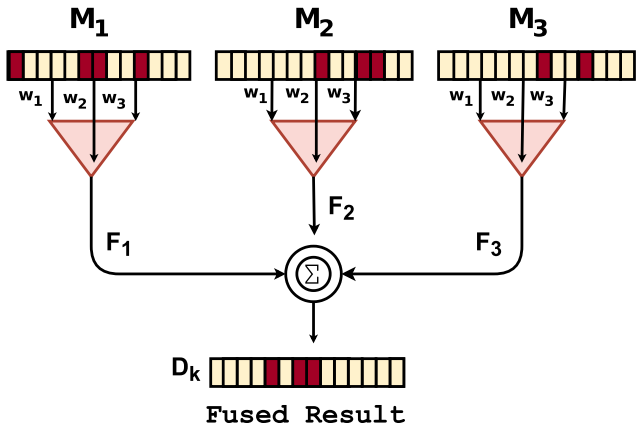

Fig. 3. M_1, M_2 and M_3 are segment-wise anomaly scores of a video from three different streams. D_k is the fused score.

video segments from both anomalous and non-anomalous video bags. Moreover, the proposed method utilizes the sparse variation of segment scores throughout the video as defined in (5),

$$\lambda \sum_{i}^{n} f\left(S_{ma}^{i}\right) \tag{5}$$

where S_{ma}^{i} is the ith segment of anomalous video and λ is the smoothness factor of the MIL ranking loss across all anomalous segments similar to the formulation presents in Sultani et al. (2018). Finally, the ranking loss is given in (6).

$$L\left(\mathcal{B}_{ma}, \mathcal{B}_{la}\right) = l\left(\mathcal{B}_{ma}, \mathcal{B}_{la}\right) + \lambda \sum_{i}^{n} f\left(S_{ma}^{i}\right) \tag{6}$$

3.4. Fusion of anomaly scores for localization

Anomalies often appear only for a short period in real-world situations. Therefore, scores of the instances (segments) in the anomalous bag should be sparse. Moreover, the anomaly score is expected to vary smoothly across nearby video segments. Hence, it is advisable to fuse different streams together with a normalized score. This section, discuss a fuzzy-based fusion process to combine the anomaly scores of the localized segments. As depicted in Fig. 2, the proposed architecture consists of three integrated streams of features. These features are processed through independent MIL classifiers. Thus, three anomaly scores are obtained for each segment. Finally, with the help of a fuzzy membership function, these scores are fused. In attention-aware feature encoding technique (Zhu & Newsam, 2019), all the frames within a segment are assigned (broadcast) with the same MIL score. However, this broadcasting of segment score to all the enclosed frames produces undesired effect in contrast with the real-life situations. Hence, the obtained segment scores have been normalized before assigning it to the respective frames. Suppose n denotes the total number of frames in a video and M_i represents the MIL score obtained from the ith stream. Since it is an array of anomaly scores, M_i contains 32 segment-wise prediction values between [0, 1]. At the end, M_i array consists of a total of n prediction scores, $M_i = \{S_0, S_1, \ldots, S_n\}$, where S_i is the anomaly score of the ith segment. The proposed fuzzy fusion is depicted in Fig. 3. A few conventional fuzzy membership functions have been modified by defining various intervals to obtain the anomaly score of the targeted frame in a surrounding neighborhood. Moreover, percentage-slice of total frames available in a segment is also introduced by a base limit of a, b and upper limit m, where $a, b < m$ and total frames occupied by a, m and b are $0 < a$; $a, b < m$ and $m < F_r$. It ensures that the interval (specific number of frames) ranges between 0 and $F_r - 1$, where F_r is the number of frames per segment, which depends on the length of the video. These three intervals contain a specific number of frames and their anomaly scores are distribute using (7).

Table 3
Three different anomaly score distribution strategies with % value of base (a, b) and upper m limits.

	Distribution type	Percent value		
		a	m	b
1.	Sparse	10	80	10
2.	Medium	25	50	25
3.	Broadcast (Sultani et al., 2018)	–	100	–

$$M_i\left(S_x\right) = \begin{cases} (S)/2 & x \le a \\ S & a < x \le m \\ (S)/2 & x \ge b \end{cases} \tag{7}$$

Here, M_i is the MIL score obtained from the ith stream, S_x is the anomaly score of xth frame, and S is the entire segment's anomaly score. Experiments have been carried out with 3 different value-set of a, b and m as shown in Table 3. Now, using a triangular window, namely $F_{j,k}^{i}$ the fused anomaly score of the ith stream and the kth frame is obtained. The prediction is done using (8). Here, w_1, w_2, and w_3 are relative weights empirically defined as $w_1 = w_3 = 0.5$ and $w_2 = 1.0$.

$$F_{j,k}^{i} = w_1 * M_i[j-1] + w_2 * M_i[j] + w_3 * M_i[j+1] \tag{8}$$

The values of j and k vary between 1 and $n - 1$. Finally, all three fused results are added using (9) to obtain the final score of abnormality. Here, $i = \{1, 2, 3\}$ and D_k represent the fused anomaly score of the kth frame. This completes the process of classifying a video as normal or anomaly.

$$D_k = \text{softmax}(F(i,k) + F(i+1,k) + F(i+2,k)) \tag{9}$$

3.5. Anomaly classification

As discussed in Section 3, in addition to anomaly *detection*, anomaly *classification* has been done on two types of anomalies, i.e., fire and accident. The method is presented in Fig. 4. Deep convolutional neural networks often make the underlying classification process slower (Chan et al., 2015; Perera & Patel, 2018). Therefore, a smaller network has been used to predict the scores quickly. Depth-wise convolutions neural network can be useful in this scenario (Howard et al., 2017). Usually, such networks consist of two separate convolutions layers, namely a depth-wise convolution layer and a point-wise convolution layer. Depth-wise separable convolution networks reduce the computational cost by factorizing the convolution process within the aforementioned layers. Therefore, two proposed light-weight classifiers have been proposed using depth-wise separable convolutions to make the classification process faster.

In order to perform the classification, only those segments that are anomalous (higher scores) are processed and remaining are ignored. The processing is done using two-class classifiers (TCC) as depicted in Fig. 4. TCC is a depth-wise convolution neural network. The parameters of the TCC model are represented in Table 4. Every frame of the anomalous segment is passed through two separate TCCs and *winner-takes-all* strategy has been adopted to decide the class of the segment. In this paper, only fire and road accident classes have been addressed. Classes like robbery-related (theft, shop lifting, vandalism, etc.) or human-related (assault, arrest, shooting, etc.) usually contain ambiguous object-level interactions. TCC may not be sufficient for classifying such anomalies.

4. Experiments and results

4.1. Dataset details

To analyze the performance of the proposed method, several experiments have been carried out on **UCF-Crime** dataset (Sultani et al.,

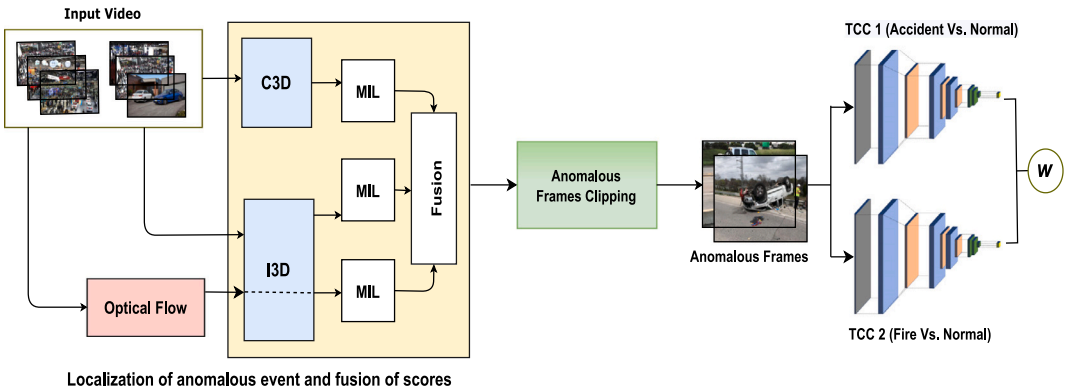

Fig. 4. The proposed model of event classification. After fusion of the MIL scores, the final segment scores and the corresponding video are taken as the inputs to the event classifier module. The frames within the anomalous segments are processed through the trained 2-class classifiers (TCCs). The *winner-takes-all* strategy has been adopted and the highest confidence score is taken as the final prediction.

2018). It contains 1900 real-world surveillance videos related to crime scenes and public place anomalies. It has 13 different anomaly classes. It has been divided into 1610 training videos and 290 testing videos. However, some of these 13 classes share similarities between each other in terms of action. In order to study the effectiveness of the motion information during the training process, UCF-Crime dataset has been partitioned into five distinct categories. For example, *assault* and *fighting* classes contain group of people brutally attacking each others. Similarly, *arson* and *explosion* classes contain fire and smoke-related incidents. Many of these categories fall under the crime category. Contextual similarity among the classes has been taken into account and grouped videos into five broad categories, namely *fire-related*, *human group-related*, *robbery*, *road accidents*, and *normal* based on the severity of anomaly and rate of change of action. As discussed earlier, the definition of anomaly is highly subjective. Therefore, these five categories are also subjective. A *rank* or *level* to each category has been assigned based on the severity and rate of change of activities. Normal videos are categorized as level 1 (no severity) and fire-related videos are grouped as level 5 (with the highest severity). The idea behind assigning levels to the categories is to characterize these videos into more generic levels. This characterization has helped us in the classification process. The dataset division is summarized in Table 5. The results using three well-known evaluation metrics, namely frame-level receiver operating characteristics (ROC) curve, area under the curve (AUC), and false alarm rate (FAR) have been presented.

4.2. Results of anomaly localization

This section presents the results of anomaly detection followed by specific anomaly type classification. The experiments have been conducted in two ways. The first approach uses a 2-stream architecture involving RGB and (C3D+I3D) streams. The second approach is an enhancement that supports an extra flow-stream used in a tightly-coupled (SelFlow+I3D) manner. Both setups use fuzzy-based fusion for aggregation of the anomaly scores. Several experiments have been carried out by training the architecture against each type of broad-level anomaly. In such experiments, it has been trained as normal vs. broad category. Such analysis has not been done in any of the earlier works. The original C3D-based implementation incorporated in MIL-based method (Sultani et al., 2018) has been utilized. Moreover, I3D, a fused C3D with I3D features, and the SelFlow combined with C3D and I3D features with improved MIL have been used for performance comparisons. Each type of the anomaly as presented in Table 5 is evaluated against the proposed model. The frame-level localization accuracy values (in terms of AUC) are presented in Table 6.

A few samples category-wise anomaly localization results are presented in the Fig. 5. These experiments reveal that the proposed fuzzy-based fusion model with SelFlow improves the localization accuracy,

Table 4
Parameters of the classifier architecture.

Layer type	Filter shape	Output shape	Param #
separable_conv2d_1	7 × 7 × 16	128 × 128 × 16	211
separable_conv2d_2	3 × 3 × 32	64 × 64 × 32	688
separable_conv2d_3a	3 × 3 × 64	32 × 32 × 64	2400
separable_conv2d_3b	3 × 3 × 64	32 × 32 × 64	4736
max_pooling2d	2 × 2 × 1	16 × 16 × 64	–
flatten	–	16 384	–
dense	–	128	2 097 280
activation_norm × 2	–	128	512
dropout_1	–	128	0
dense	–	2	258
classifier	–	2	0

particularly on the videos that are under Level 4 or *road accident* and Level-3 or *human-related* categories. The inclusion of flow certainly helps the architecture to learn object-level interactions in a better way for these categories. The ROC/AUC plots for Level-3 and Level-4 are presented in Fig. 8. It can be concluded from the results that the proposed architecture with the inclusion of SelFlow, C3D, and I3D performs better than the method proposed by MIL-based method (Sultani et al., 2018). The increase in AUC across Level-2 and level-5 can also be observed in Figs. 6–9. Though the gain is lesser as compared to Level-3 and Level-4, however, fusion of multiple streams has certainly improved the AUC values. However, this lesser gain as compared to Level-3 and Level-4 is due to the nature of events under these categories. For example, in Level-2 (robbery-related), interactions can happen between humans (child lifting, fighting, etc.) as well as human-to-static objects (burglary or shop lifting). Distinguishing these events can be difficult, particularly in crowded environments. Events in Level-5 (fire-related) category usually contain sudden change in the scene. This can happen in normal events (e.g. rapid change in a scene due to fast moving objects like crowd moving in uncertain direction), thus getting relatively lower AUC values.

The performance of the proposed method for normal vs anomaly (binary) classification are also compared with six recently published research work such as the *dictionary-based approach* (Luo et al., 2017a) to encode the normal behavior from videos and used reconstruction error to identify anomalies. The authors (Hasan et al., 2016) have used *deep auto-encoder approach* to extract relevant features to train the classifier. In MIL-based method (Sultani et al., 2018), C3D features have been extracted from the videos. These features are used to train an improved MIL model with Adagrad optimizer using a learning rate of 0.001. Following the work proposed in Sultani et al. (2018), the authors in Zhu and Newsam (2019) have integrated the motion-aware features obtained by *attention-block approach* into the pre-computed C3D features. They have recorded a significant increase in the accuracy

Table 5
Level-based division of the videos into various categories used for model training and testing. Higher the level, the more the severity of the event.

Category	Type of anomalies	Training videos	Testing videos	Level
Normal	Normal behavior	800	150	1
Robbery related	Burglary, Shoplifting, Stealing, Robbery, Vandalism	401	49	2
Human related	Abuse, Assault, Arrest, Fighting, Shooting	212	38	3
Road accident	Road accident	127	23	4
Fire related	Arson, Explosion	70	30	5

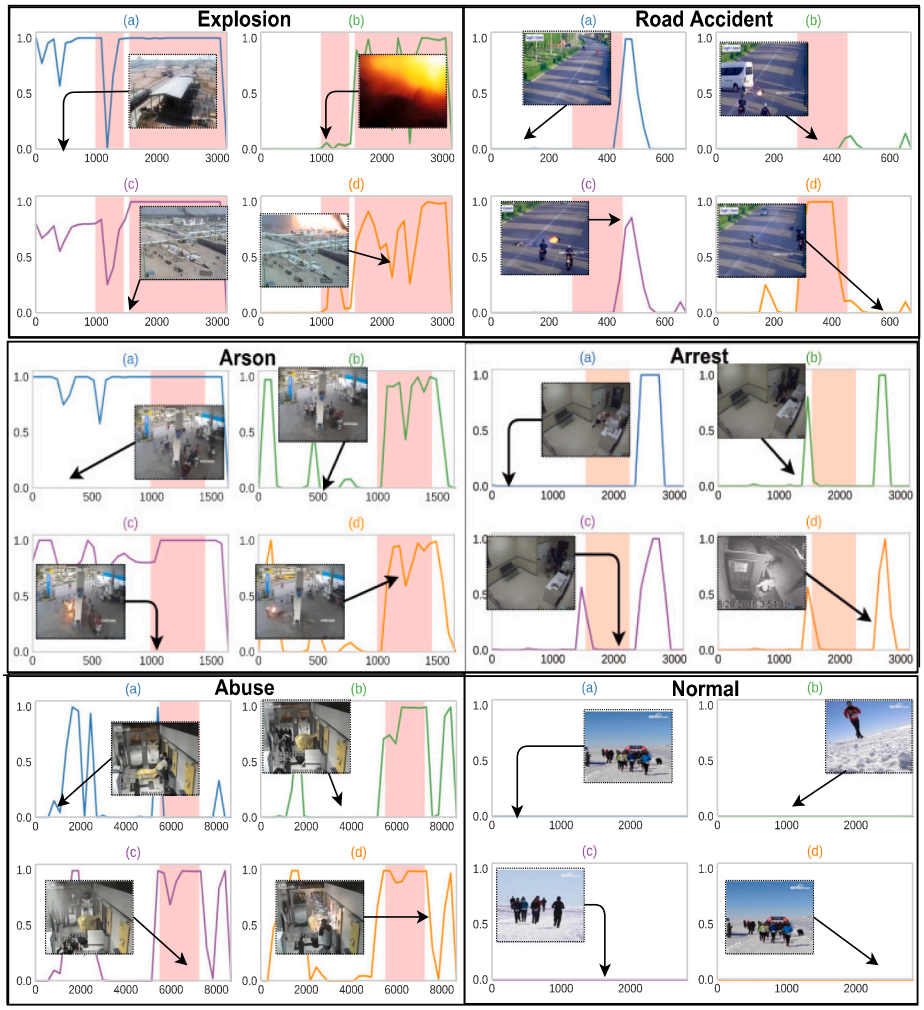

Fig. 5. Quantitative results of the proposed model and other methods. Frame numbers are given along the X-axis and Y-axis presents the anomaly scores. The shaded regions represent the ground truths. (a) Represents anomaly scores obtained using the method proposed by Sultani et al. (2018) (b) Presents the results of the proposed model trained on I3D features only (Carreira & Zisserman, 2017). (c) and (d) Represent the results obtained by fusion of various streams (C3D+I3D) and (C3D+I3D with SelFlow Liu et al., 2019), respectively. Visual results reveal that the proposed fuzzy-based fusion model with SelFlow improves the localization accuracy.

Table 6
Comparisons among various models on the original UCF-Crime dataset. The rows represent categories and columns represent the ROC values. Each cell gives the frame-level localization accuracy (using AUC in %) on that particular category. The second column presents the results (Sultani et al., 2018).

	Baseline MIL with C3D features	Improved MIL with I3D features	Proposed fusion model C3D + I3D	Proposed fusion model (C3D + I3D) with SelFlow
Level 1	99.84	97.97	99.64	99.65
Level 2	50.54	55.37	54.46	**56.45**
Level 3	61.58	64.91	69.05	**70.48**
Level 4	50.08	56.39	52.26	**64.80**
Level 5	48.66	53.22	52.89	**56.44**

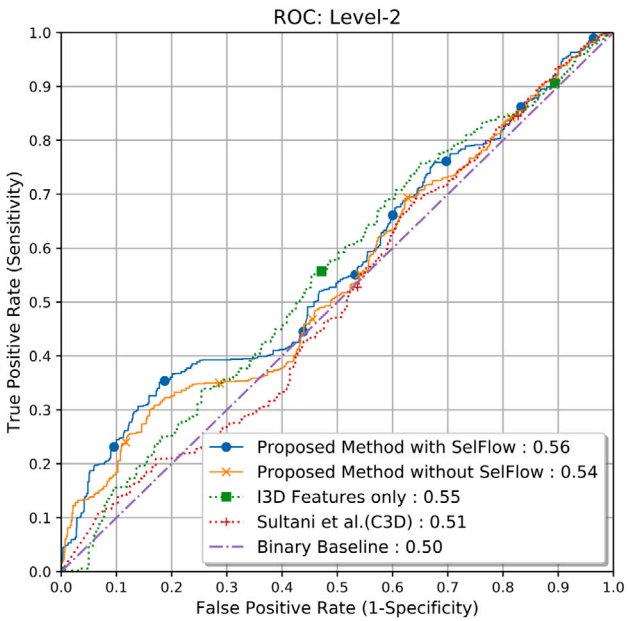

Fig. 6. AUC/ROC curves of experiments (Normal vs. Level 2) using the UCF-Crime dataset. Level-2 category contains challenging crime scenes like stealing, vandalism. etc. These crime scenes usually take place in congested environment making them difficult to track using shallow details about the scene. It may be observed that no model is reliable enough. However, the proposed fused architecture performs relatively better than other methods.

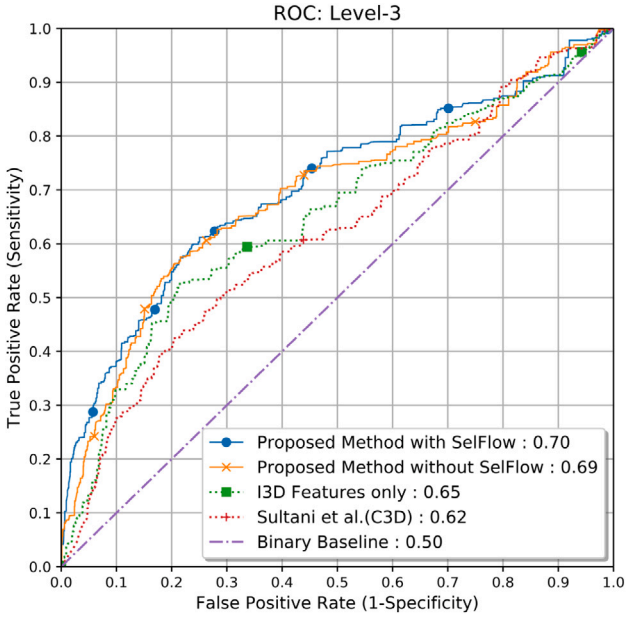

Fig. 7. AUC/ROC curves of experiments (Normal vs. Level 3) using the UCF-Crime dataset. Level-3 category contains human-related anomalies like fighting, abusing, and arresting involving movements of group of people. Rapid people movements are well captured with the help of fusion of C3D (Tran et al., 2015) and I3D (Carreira & Zisserman, 2017) features.

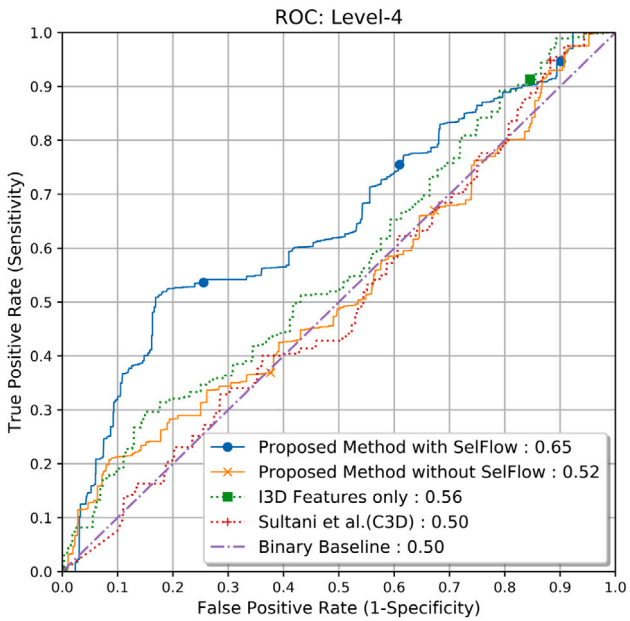

Fig. 8. AUC/ROC curves of experiments (Normal vs. Level 4) using the UCF-Crime dataset. Road accident anomaly contains motion of pedestrians, accelerating cars, crashing of cars, etc. In such scenarios, additional flow information certainly helps to localize the events more accurately. In both cases, the proposed architecture with the inclusion of SelFlow (Liu et al., 2019), C3D (Tran et al., 2015), and I3D (Carreira & Zisserman, 2017) performs better than MIL-based method (Sultani et al., 2018).

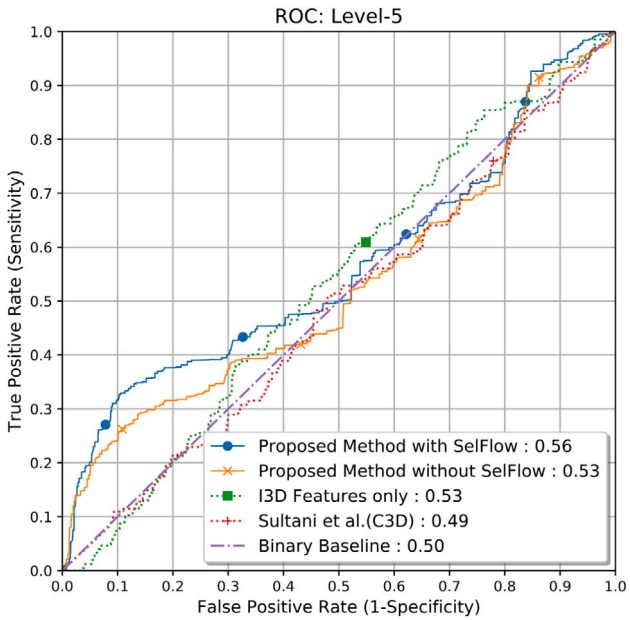

Fig. 9. AUC/ROC curves of experiments (Normal vs. Level 5) using the UCF-Crime dataset. Fire-related category contains severe crime scenes like arson and explosion. In this category, when an anomaly takes place, the entire scene gets filled with flames and smoke making it difficult to understand the objects and their interactions. It may be observed that the proposed model localizes such events with better accuracy.

as compared to Sultani et al. (2018). Following the approach proposed by Sultani et al. (2018) and Zhu and Newsam (2019), the graph-based method (Zhong et al., 2019) reconstructs the problem and introduces a separate *graph convolution network (GCN)* to correct the noisy labels. Based on the temporal information, GCN propagates the guided information from higher-confidence anomaly segments to lower-confidence segments. The quantitative comparisons of AUC values are shown in Table 7. Results demonstrate that the proposed method performs better than deep encoder-based method (Hasan et al., 2016), dictionary-based approach (Luo et al., 2017a), MIL method (Sultani et al., 2018), attention-aware feature encoding (Zhu & Newsam, 2019), graph convolution network (Zhong et al., 2019) with C3D, and TSN optical flow (Zhong et al., 2019) when AUC and FAR metrics are taken into considerations. It can be observed that the proposed method produces

Table 7
Results of normal vs anomaly classification using UCF-Crime dataset.

Method	AUC (%)	FAR (%)
Binary classifier	50.0	–
Deep feature encoder (Hasan et al., 2016)	50.6	27.2
Dictionary-based encoding (Luo et al., 2017a)	65.6	3.1
C3D features + MIL (Sultani et al., 2018)	73.6	1.7
Motion-aware features (Zhu & Newsam, 2019)	79.0	–
GCN + TSN optical flow (Zhong et al., 2019)	78.0	1.1
Temporal CN + Inner bag loss (Zhang et al., 2019)	78.66	–
FC network + Clustering loss (Zaheer et al., 2020)	79.54	–
GCN + TSN C3D (Zhong et al., 2019)	81.0	2.8
GCN + TSN RGB + Filter (Zhong et al., 2019)	82.1	0.1
Proposed method (C3D + I3D) with SelFlow	**84.48**	0.2

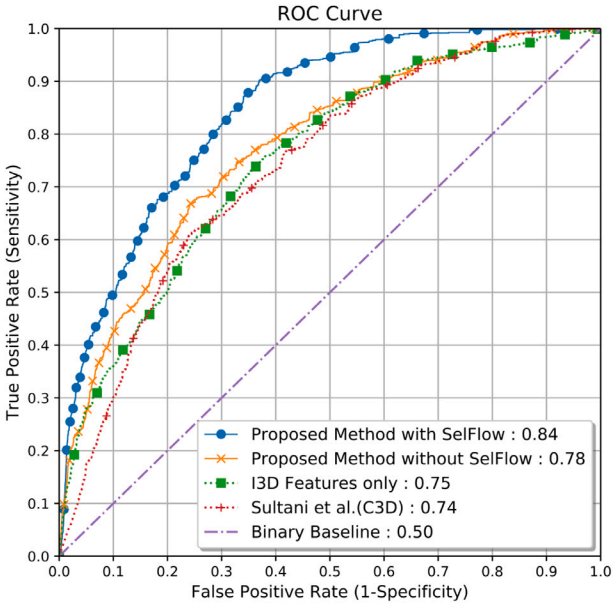

Fig. 10. AUC/ROC curves of experiments using the UCF-Crime dataset (normal v/s anomalous) videos. Proposed method without SelFlow (Liu et al., 2019) that fuses the scores obtained using C3D (Tran et al., 2015) and I3D (Carreira & Zisserman, 2017) features. Proposed model with SelFlow (Liu et al., 2019) that uses the fusion of scores obtained using C3D and I3D (with SelFlow) features.

the best AUC as compared to all methods summarized in Table 7 including the best method available (Zhong et al., 2019). Since the proposed method has improved AUC (84.48%) by 2% from the best method available, it is more likely to detect the true positive anomaly cases, which is very important in many surveillance contexts as depicted in Fig. 10. Even though the FAR is slightly higher (0.2) in the proposed method as compared to Zhong's method (Zhong et al., 2019), it can detect critical events at a higher rate that is more important as compared to the marginal increase in the false positive rates. For example, it is more important to send the rescue team to the on-road accident spots for quick action. A better detection accuracy is generally preferred over lower FAR, particularly in emergency situations.

4.3. Anomaly localization results vs. Distribution of scores

The results of the fusion of anomaly scores as per the strategies discussed earlier is presented in this subsection. Table 3 presents these strategies. Several experiments have been carried out to check the performance of these strategies as shown in Table 8. As discussed earlier, in MIL-based method (Sultani et al., 2018) and attention-aware feature encoding method (Zhu & Newsam, 2019), all the frames within a segment are assigned only one MIL score. Other strategies by

Table 8
AUC results using various combinations of features including SelFlow (Liu et al., 2019).

Sr. No.	Distribution type	C3D + I3D	Fusion with SelFlow
1.	Sparse	77.32	**79.21**
2.	Medium	78.09	**81.40**
3.	Broadcast	78.45	**84.48**

distributing the MIL score across the frames within a segment are also tried. In one such strategy (sparse distribution), a fraction of MIL score is assigned to the first and last 10% of the frames present in a segment. Remaining frames are assigned with the actual MIL score. In a medium distribution strategy, 50% value of the MIL score is assigned to the first and the last 25% of the frames within the segment and a full score is assigned to the remaining frames. The broadcasting strategy that has been reported in Zhu and Newsam (2019) broadcasts the MIL score of a segment to all the frames present in that segment.

It can be inferred from the results presented in Table 8 that a simple broadcasting of the MIL score to all the frames of a segment is not a good choice as anomaly periods are often shorter in duration. The scores in the anomalous bag should rather be sparse. Hence incorporating medium or sparse distribution of segment score is better particularly for detecting rare and short occurrences of anomaly events.

4.4. Effect of smoothness factor and weights in fuzzy fusion

As discussed earlier, the MIL does not consider the temporal smoothness constraint. Inclusion of motion information during the training eliminates the need of temporal smoothness. However, in the proposed improved MIL, the smoothness constraint (λ) as defined in Eq. (5) has been employed. It ensures that the anomaly scores follow sparse distribution across the frames. Moreover, default value of λ is set to 8×10^{-5} in the work proposed in Sultani et al. (2018). Table 9 shows the effectiveness of values of λ at various scale. However, no significance performance change can be observed.

A separate experiment has been carried out to understand the effect of weight defined in Eq. (9). The initial values of $w_1 = 1.0$, $w_2 = 0.5$, and $w_3 = 0.5$ are empirically set using a triangular fuzzy membership function presented in Eq. (7). Trapezoidal and Gaussian membership functions have also been tested with. A trapezoidal membership functions defined into four intervals with weight values $w_1 = 0.2$, $w_2 = 0.8$, $w_3 = 0.8$, and $w_4 = 0.2$. For Gaussian membership function, a standard Gaussian distribution has been implemented to assign anomaly score for each video frame. Table 10 presents the performance comparisons using various fuzzy membership functions.

4.5. Results of anomaly classification

Since MIL acts as a binary classifier, it only tells whether a video is normal or anomalous. Then classification stage classify the segments

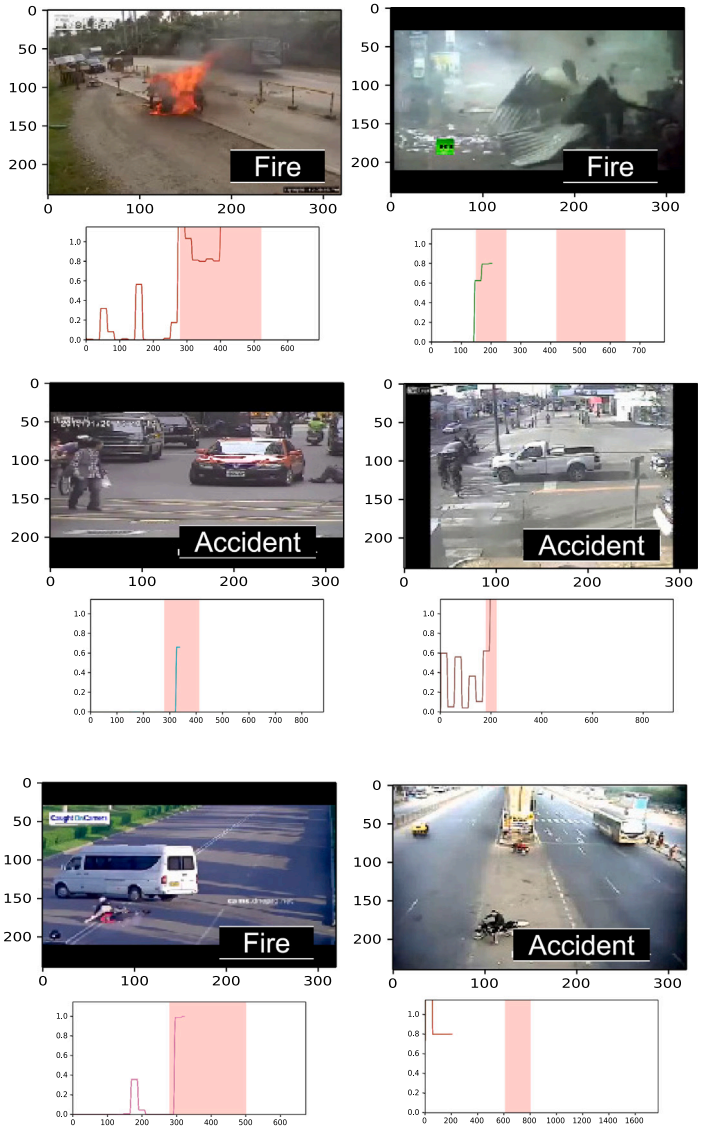

Fig. 11. Classification results on fire-related and road accident-related categories. Shaded regions represent the ground truths, whereas the given annotation is the prediction of one of the winners. The last two results are false positives produced by both models. In the fifth example, the model predicts the fire incident as an anomaly but it is actually a road accident. The last result is a failure of anomaly detection itself.

Table 9
AUC results with different values of smoothness factor λ.

Sr. No.	λ value	C3D + I3D	Fusion with SelFlow
1.	8×10^{-3}	77.89	84.06
2.	8×10^{-4}	78.20	84.32
3.	8×10^{-5} (Sultani et al., 2018)	**78.45**	**84.48**

Table 10
AUC results with different fuzzy membership functions used in the anomaly score distribution stage.

Sr. No.	Fuzzy membership function	C3D + I3D	Fusion with SelFlow
1.	Trapezoidal	75.90	80.17
2.	Gaussian	78.22	84.31
3.	Triangular	**78.45**	**84.48**

containing anomaly without identifying their types as shown in Fig. 4. To identify the type of an event, TCCs (Two-class classifiers) have been trained over normal and anomalous images related to anomalous events. At present, it has been tested on two types of videos, namely *fire-related* and *road accident-related*. A large dataset comprising fire images (1600 samples) and road accident images (2500 samples) have been created. The publicly available 8-scene dataset (Oliva & Torralba, 2001) that contains 2600 images of 8 outdoor scene categories, namely coast, mountain, forest, open country, street, inside the city, tall buildings, and highways has been used. During training, the model treated fire and road accident images as anomalous and images from the 8-scene dataset as normal. Two separate TCC models, one for the *fire vs normal* and another for the *road accident vs normal* have been trained with a learning rate of 0.01 and keeping the number of epochs to 70. Finally, *winner takes all* strategy has been employed to predict the type of anomaly present in the segment. Table 11 summarizes the results of classification. Some classification results are presented in Fig. 11. Performance of the classification can be analyzed from the results shown in Figs. 12 and 13. The classifier that predicts more **YES**, its base class is assumed to be the true category of the segment. This can be verified from the frames corresponding to the event as shown in the figures.

5. Discussion of results

The results reveal that the addition of the optical flow (SelFlow Liu et al., 2019) increases the prediction accuracy. Also, the SelFlow speeds

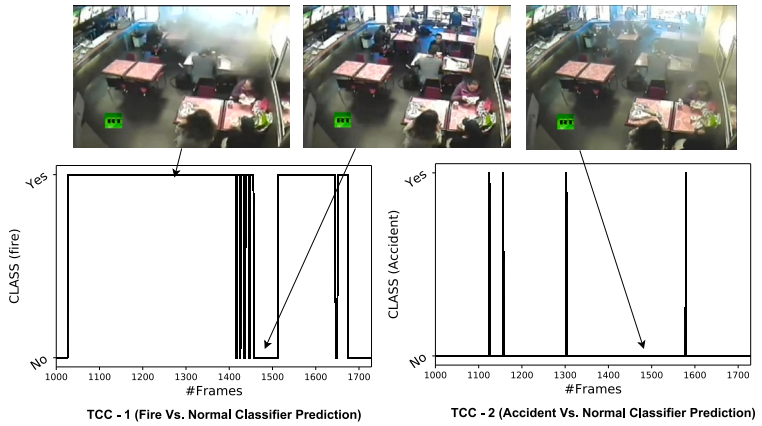

Fig. 12. An anomalous segment (fire-related) being classified by the 2-class classifiers (TCCs). The top row shows some of the representative frames of the incident. The bottom row presents how the TCC-1 (fire vs. non-fire) and TCC-2 (accident vs non-accident) predict the class of individual frames within the segment. Finally, TCC-1 wins as more number of frames of the segment is being classified as fire-related.

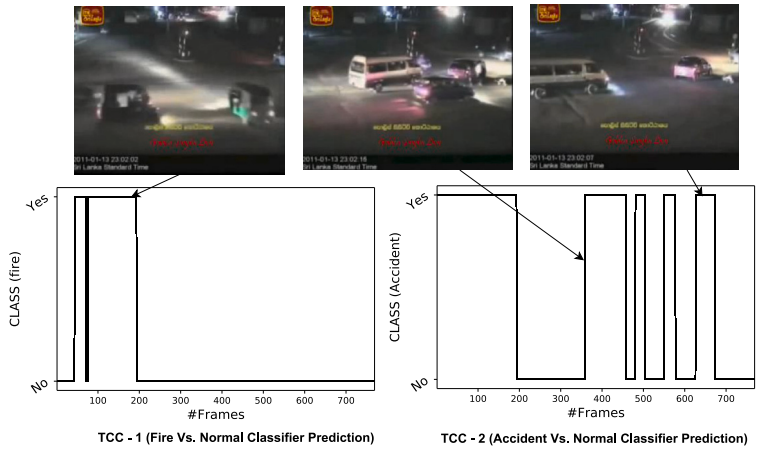

Fig. 13. An anomalous segment (road accident) being classified by the 2-class classifiers (TCCs). The top row shows some of the representative frames of the incident. The bottom row presents how the TCC-1 (fire vs. non-fire) and TCC-2 (accident vs non-accident) predict the class of individual frames within the segment. Finally, TCC-2 wins as more number of frames of the segment is being classified as accident-related. However, toward the beginning of the segment, some frames are wrongly classified as fire-related. This happens due to the presence of a strong headlight of the cars on the road.

Table 11
Results of fire and accident classification.

Models	Classes	Precision	Recall	f1-score
1.	Normal	0.86	0.97	0.91
	Fire	0.94	0.71	0.81
2.	Normal	0.83	0.99	0.90
	Accident	0.93	0.77	0.62

up the process of flow calculation. It operates at 3–4 times faster as compared to the original input stream that operates only at 30 fps. The I3D flow module acts as encoder that compresses the flow feature segment to a 1024-byte vector. Both strategy (with or without SelFlow) can maintain a lesser false-positive rate over the normal group. The fuzzy aggregation step helps to improve the overall accuracy score. Moreover, inclusion of SelFlow into the fusion has certainly helped to improve the localization AUC, particularly for *road accident* category events. The inclusion of flow helps to learn the behavior of moving vehicles and distinguish them from others. The increase in accuracy across other groups, though relatively lesser as compared to Level-4, is also notable. Table 6 suggests that category-wise classification results need further improvement as accuracy is hovering around 60% for Level-2, Level-3, and Level-5 categories. This may also be due to the data insufficiency in some of these classes against the normal class.

Here, some results have been shown to demonstrate that an event such as *running* can be normal in one context and anomalous in another context. As depicted in Fig. 14, in the first video, a criminal is running and the cops are following inside a railway station. In the second video, a group of athletes is running in a Marathon competition. Both videos contain running events. However, the criminal running segment in the first video has been detected as anomalous by the proposed architecture, whereas the running of athletes in the second video has been correctly detected as normal. This reveals that the proposed architecture can decode the context of an event, which is very important for accurate classification.

The final observations of the paper are as follows: (i) It is difficult to identify true class of an video using the anomaly detection process. However, experiments reveal that some post-processing using binary classifier can be highly useful as reported in Table 11. (ii) As anomalies in are rare in occurrence, detection methods should emphasize on minimizing the false alarms. (iii) Multi-stream architectures can be more accurate as compared to single-stream networks for video anomaly localization as explained in Table 7.

6. Conclusions

In this paper, a few critical problems related to video anomaly detection, namely higher false alarms, lack of scene understanding, and

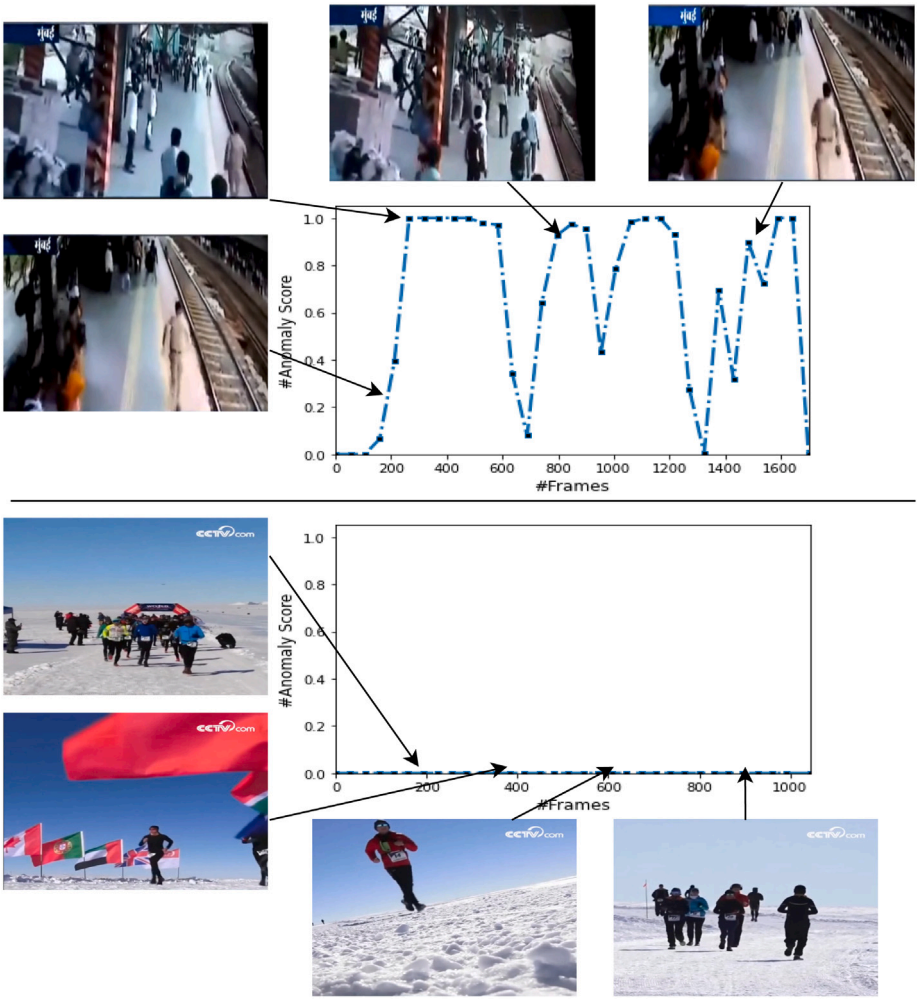

Fig. 14. An anomalous event in one context can be completely normal in another context. In the first video, a criminal is running and the cops are following inside a railway station. In the second video, a group of athletes are running in a Marathon. Both videos contain running events. However, criminal running video has been detected as anomalous by the proposed model, whereas the running of athletes in the second video has been correctly detected as normal.

poor detection performance have been addressed. Two key observations are: (i) Hand-crafted features are not sufficient to characterize complex anomaly events. A robust detection method should exploit appearance as well as motion features. (ii) A normal event can be abnormal in certain condition and vice-versa. This often leads to higher false alarm if the detection method follows only normal-encoding. The proposed multi-stream architecture with post-fusion strategy outperforms the current state-of-the-art methods by a notable margin. It exploits weakly-supervised training, thus experiment results highly agree with the ground truths. In addition to this, the proposed method can identify normal and abnormal video segments with FAR below 1%. The paper also presents a class-wise performance comparisons with recent anomaly detection methods. It can be observed that the proposed method significantly improves the performance by a margin of 10%–15%.

Even though the proposed method achieves state-of-the-art performance, the normal–abnormal feature encoding technique is still vulnerable to the subjective definition of an anomalous event. In future, it is possible to further strengthen the detection framework to ensure that it can handle complex anomalies. The noise filtering technique introduced in graph-based convolution method (Zhong et al., 2019) can be incorporated to improve the classification accuracy. The current model obtains raw anomaly score for each segment using improved MIL network. However, the obtained score can be further refined by employing additional bag-loss technique such the method introduced in Zhang et al. (2019). Also, specific research like violence detection (Mohammadi et al., 2016), human aggression detection (Kooij et al., 2016), attention-aware feature encoding (Zhu & Newsam, 2019) and crowd anomaly behavior (Behera et al., 2021) are some exciting future research directions of the current study. Moreover, it can be explored whether a classical optimization technique such as the one used in Dey et al. (2021) performs better than the fuzzy-based fusion strategy or not.

CRediT authorship contribution statement

Thakare Kamalakar Vijay: Conceptualization, Formal analysis, Investigation, Implementation of methodology, Testing and validation, Writing – original draft. **Nitin Sharma:** Conceptualization, Implementation, Testing, and validation. **Debi Prosad Dogra:** Conceptualization, Investigation, Supervision, Writing – review & editing. **Heeseung Choi:** Conceptualization, Investigation, Supervision, Writing – review & editing. **Ig-Jae Kim:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The work was funded under Korea Institute of Science and Technology (KIST) Institutional Program (Project No. 2E31082), NRF Project (2018M3E3A1057288) executed at IIT Bhubaneswar under the Project code CP106.

References

Behera, S., Vijay, T. K., Manish Kausik, H., & Dogra, D. P. (2021). PIDLNet: A physics-induced deep learning network for characterization of crowd videos. In *2021 17th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1–8). http://dx.doi.org/10.1109/AVSS52988.2021.9663817.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 4724–4733). http://dx.doi.org/10.1109/CVPR.2017.502.

Chan, T. H., Jia, K., Gao, S., Lu, J., Zeng, Z., & Ma, Y. (2015). PCANet: A simple deep learning baseline for image classification? *IEEE Transactions on Image Processing*, *24*(12), 5017–5032. http://dx.doi.org/10.1109/TIP.2015.2475625.

Dey, B. K., Bhuniya, S., & Sarkar, B. (2021). Involvement of controllable lead time and variable demand for a smart manufacturing system under a supply chain management. *Expert Systems with Applications*, *184*, Article 115464. http://dx.doi.org/10.1016/j.eswa.2021.115464.

Diba, A., Fayyaz, M., Sharma, V., Karami, A. H., Arzani, M. M., Yousefzadeh, R., & Van Gool, L. (2017). Temporal 3D ConvNets: New architecture and transfer learning for video classification. arXiv.

Gao, Y., Liu, H., Sun, X., Wang, C., & Liu, Y. (2016). Violence detection using oriented violent flows. *Image and Vision Computing*, *48–49*, 37–41. http://dx.doi.org/10.1016/j.imavis.2016.01.006.

Giorno, A. D., Bagnell, J. A., & Hebert, M. (2016). A discriminative framework for anomaly detection in large videos. ArXiv, arXiv:1609.08938.

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning temporal regularity in video sequences. *2016-Decem*, In *2016 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 733–742). http://dx.doi.org/10.1109/CVPR.2016.86.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv.

Ionescu, R., Smeureanu, S., Alexe, B., & Popescu, M. (2017). Unmasking the abnormal events in video. In *2017 IEEE international conference on computer vision (ICCV)* (pp. 2914–2922). http://dx.doi.org/10.1109/ICCV.2017.315.

Jiang, Y. G., Wu, Z., Wang, J., Xue, X., & Chang, S. F. (2018). Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(2), 352–364. http://dx.doi.org/10.1109/TPAMI.2017.2670560.

Kooij, J. F., Liem, M. C., Krijnders, J. D., Andringa, T. C., & Gavrila, D. M. (2016). Multi-modal human aggression detection. *Computer Vision and Image Understanding*, *144*, 106–120. http://dx.doi.org/10.1016/j.cviu.2015.06.009.

Li, N., & Chang, F. (2019). Video anomaly detection and localization via multivariate gaussian fully convolution adversarial autoencoder. *Neurocomputing*, *369*, 92–105. http://dx.doi.org/10.1016/j.neucom.2019.08.044.

Li, W., Mahadevan, V., & Vasconcelos, N. (2014). Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(1), 18–32. http://dx.doi.org/10.1109/TPAMI.2013.111.

Liu, P., Lyu, M., King, I., & Xu, J. (2019). Selflow: Self-supervised learning of optical flow. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Lu, Z., Fu, Z., Xiang, T., Han, P., Wang, L., & Gao, X. (2017). Learning from weak and noisy labels for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(3), 486–500. http://dx.doi.org/10.1109/TPAMI.2016.2552172.

Lu, C., Shi, J., & Jia, J. (2013). Abnormal event detection at 150 FPS in MATLAB. In *2013 IEEE international conference on computer vision (ICCV)* (pp. 2720–2727). http://dx.doi.org/10.1109/ICCV.2013.338.

Luo, W., Liu, W., & Gao, S. (2017a). Remembering history with convolutional LSTM for anomaly detection. In *Proceedings - IEEE international conference on multimedia and expo* (pp. 439–444). http://dx.doi.org/10.1109/ICME.2017.8019325.

Luo, W., Liu, W., & Gao, S. (2017b). A revisit of sparse coding based anomaly detection in stacked rnn framework. In *2017 IEEE international conference on computer vision (ICCV)* (pp. 341–349). http://dx.doi.org/10.1109/ICCV.2017.45.

Mohammadi, S., Perina, A., Kiani, H., & Murino, V. (2016). Angry crowds: Detecting violent events in videos. In *LNCS*: *Vol. 9911*, *Lecture notes in computer science (artificial intelligence and lecture notes in bioinformatics)* (pp. 3–18). http://dx.doi.org/10.1007/978-3-319-46478-7_1.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175. http://dx.doi.org/10.1023/A:1011139631724.

Pang, G., Yan, C., Shen, C., van den Hengel, A., & Bai, X. (2020). Self-trained deep ordinal regression for end-to-end video anomaly detection. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 12170–12179). http://dx.doi.org/10.1109/CVPR42600.2020.01219.

Perera, P., & Patel, V. M. (2018). Learning deep features for one-class classification. *IEEE Transactions on Image Processing*, *28*(11), 5450–5463.

Sabokrou, M., Fayyaz, M., Fathy, M., Moayed, Z., & Klette, R. (2018). Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Computer Vision and Image Understanding*, *172*, 88–97. http://dx.doi.org/10.1016/j.cviu.2018.02.006.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, *1*(January), 568–576.

Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 6479–6488). http://dx.doi.org/10.1109/CVPR.2018.00678.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. *2015*, In *2015 IEEE international conference on computer vision (ICCV)* (pp. 4489–4497). http://dx.doi.org/10.1109/ICCV.2015.510.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *LNCS*: *Vol. 9912*, *Lecture notes in computer science (lecture notes in artificial intelligence)* (pp. 20–36).

Xu, K., Jiang, X., & Sun, T. (2018). Anomaly detection based on stacked sparse coding with intraframe classification strategy. *IEEE Transactions on Multimedia*, *20*(5), 1062–1074. http://dx.doi.org/10.1109/TMM.2018.2818942.

Yao, J., Wang, J., Tsang, I., Zhang, Y., Sun, J., Zhang, C., & Zhang, R. (2017). Deep learning from noisy image labels with quality embedding. *IEEE Transactions on Image Processing*, *28*(04), 1909–1922.

Zaheer, M. Z., Mahmood, A., Shin, H., & Lee, S. I. (2020). A self-reasoning framework for anomaly detection using video-level labels. *IEEE Signal Processing Letters*, *27*, 1705–1709. http://dx.doi.org/10.1109/LSP.2020.3025688.

Zhang, J., Qing, L., & Miao, J. (2019). Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. *Sept*, In *IEEE international conference on image processing* (pp. 4030–4034). http://dx.doi.org/10.1109/ICIP.2019.8803657.

Zhong, J. X., Li, N., Kong, W., Liu, S., Li, T. H., & Li, G. (2019). Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 1237–1246). http://dx.doi.org/10.1109/CVPR.2019.00133.

Zhou, J. T., Du, J., Zhu, H., Peng, X., Liu, Y., & Goh, R. S. M. (2019). AnomalyNet: An anomaly detection network for video surveillance. *IEEE Transactions on Information Forensics and Security*, *14*(10), 2537–2550. http://dx.doi.org/10.1109/TIFS.2019.2900907.

Zhu, Y., & Newsam, S. (2019). Motion-aware feature for improved video anomaly detection. In *British machine vision conference*.