applied sciences

MDPI

Article

Anomalous Event Recognition in Videos Based on Joint Learning of Motion and Appearance with Multiple Ranking Measures

Shikha Dubey [1], Abhijeet Boragule [1], Jeonghwan Gwak [2,3,*] and Moongu Jeon [1]

[1] Gwangju Institute of Science and Technology (GIST), School of Electrical Engineering and Computer Science, Gwangju 61005, Korea; shikha.d@gm.gist.ac.kr (S.D.); abhijeet@gist.ac.kr (A.B.); mgjeon@gist.ac.kr (M.J.)
[2] Department of Software, Korea National University of Transportation, Chungju 27469, Korea
[3] Department of IT· Energy Convergence (BK21 FOUR), Korea National University of Transportation, Chungju 27469, Korea
* Correspondence: jgwak@ut.ac.kr; Tel.: +82-43-841-5852

Abstract: Given the scarcity of annotated datasets, learning the context-dependency of anomalous events as well as mitigating false alarms represent challenges in the task of anomalous activity detection. We propose a framework, Deep-network with Multiple Ranking Measures (DMRMs), which addresses context-dependency using a joint learning technique for motion and appearance features. In DMRMs, the spatial-time-dependent features are extracted from a video using a 3D residual network (ResNet), and deep motion features are extracted by integrating the motion flow maps' information with the 3D ResNet. Afterward, the extracted features are fused for joint learning. This data fusion is then passed through a deep neural network for deep multiple instance learning (DMIL) to learn the context-dependency in a weakly-supervised manner using the proposed multiple ranking measures (MRMs). These MRMs consider multiple measures of false alarms, and the network is trained with both normal and anomalous events, thus lowering the false alarm rate. Meanwhile, in the inference phase, the network predicts each frame's abnormality score along with the localization of moving objects using motion flow maps. A higher abnormality score indicates the presence of an anomalous event. Experimental results on two recent and challenging datasets demonstrate that our proposed framework improves the area under the curve (AUC) score by 6.5% compared to the state-of-the-art method on the UCF-Crime dataset and shows AUC of 68.5% on the ShanghaiTech dataset.

Keywords: anomalous event; deep multiple instance learning; deep motion flow; multiple ranking measures; data fusion

check for updates

Citation: Dubey, S.; Boragule, A.; Gwak, J.; Jeon, M. Anomalous Event Recognition in Videos Based on Joint Learning of Motion and Appearance with Multiple Ranking Measures. *Appl. Sci.* **2021**, *11*, 1344. https://doi.org/10.3390/app11031344

Academic Editor: Juan Rodríguez
Received: 12 December 2020
Accepted: 27 January 2021
Published: 2 February 2021

1. Introduction

Anomalous (abnormal) event detection for video surveillance is an influential area of computer vision research. Surveillance cameras monitored by humans have been installed to inhibit much of the crime happening around us. In the era of artificial intelligence, our fundamental objective is to efficiently and accurately automate most of the existing technologies, especially those that require extensive human effort and time. Anomalous event detection is one such form of automation that is especially worthwhile in light of the rising use of surveillance cameras due to camera technologies' advancement. Monitoring a large number of cameras currently requires the tedious application of human time and effort. Therefore, a system that could automatically detect anomalous events in live surveillance footage, such as robberies, road accidents, etc., is indisputably advantageous and cost-effective.

In the past several years, substantial efforts have been devoted to accomplishing the task of anomalous event detection [1–15]. Nevertheless, several challenges remain, which require consideration while proposing a technical method as follows. First, anomalous events occur rarely in real life, resulting in a scarcity of large datasets. Second, defining all possible types of anomalous events is unfeasible due to their subjectivity, complexity, and ambiguity. Third, anomalous events are context-dependent; for example, jumping while playing basketball is a normal event, though jumping into someone's house through a window is not. Fourth, the temporal annotation of all data is unfeasible. Finally, proper feature extraction is key to an effective anomalous event detection method.

Recent studies have exploited the latest advancements in deep learning techniques [1,3,4,10,12,13,16–18]. However, one shortcoming of these methods is the lack of relation between appearance and motion features, which affects these methods' performance and outcomes. One study on deep representations of appearance and motion [12] achieved the goal of learning the relationship between these features, but it was done by training three separate networks. Motivated by this work [12], we accomplished the goal of learning the relationship between appearance and motion features by training only a single network. Moreover, our framework has the advantage of preserving the individuality of the motion flow maps and spatial-time-dependent features until they are passed together for joint learning after a fusion technique. This method allows for the deep feature extraction of vital information related to motion and visual data individually, which helps our framework to maximize the utilization of both features, while the fusion technique allows for learning the relationship between the features.

In another recent study, Sultani et al. [14] has proposed a new approach to detect anomalous events. The authors used video-level annotations to train their model in a weakly supervised manner using multiple instance learning (MIL) [19,20]. Their work motivates our method. However, our method is quite different from their work in several aspects. First, our method exploits the motion flow maps along with spatial-time-dependent features. Second, we use a different feature extraction technique. Third, we propose a new objective function for DMIL using MRMs. Fourth, we include anomalous event recognition as well as localization of all moving objects. Our feature extraction technique uses a 3D ResNet [21] inspired by related studies [21,22] and the recent advancements in the action recognition task [21–23]. Additionally, the training of our proposed framework, DMRMs, is simpler than other previously introduced deep learning methods [3,4,9,11–13], since it utilizes a pre-trained network along with a simple joint learning technique.

In our proposed method, motion flow maps are integrated with a 3D deep neural network in order to capture deep motion flow information such as relative motion information of objects (spatial and time-dependent features of motion flow) and local variation of motion flow, which is further fused with spatial-time-dependent features for the joint learning of all features. In a recent study [10], the author has extracted the motion flow using the convolutional kernel method. The difference between our method and this study [10] is that we have utilized a 3D deep neural network to extract the relative motion and variational motion flow features from the dense motion flow maps, which helps our framework maximize the utilization of motion flow. In contrast, in the study [10], the author has extracted only motion flows using a convolutional kernel, and no relative motion information is extracted. Usually, motion flow maps are passed separately in the network, as in studies [9,11], showing a relationship gap between appearance and motion features. Our proposed framework, DMRMs, seeks to address this gap without neglecting the context-dependency that is critical to anomalous event detection due to the subjective definition of anomalous events. Therefore, our work addresses multiple challenges of the task and helps mitigate the false alarm rate.

The primary contributions of this study are summarized as follows:

1. A pre-trained 3D ResNet is used, not only to extract global, context-based features of moving objects but also to extract the deep local motion-related information from their motion flow maps. The network is trained to learn the context-dependency or the relationship between spatial-time-dependent features and deep motion features using joint learning.

2. A new objective function is proposed for DMIL using MRMs. Anomalous events are detected with their class (type) and with the localization of moving objects using motion flow maps.
3. The proposed framework has been tested on the two most recent and challenging datasets–UCF-Crime [14] and ShanghaiTech [5]. The results show that our method improves by 6.5% AUC on the UCF-Crime dataset compared with the state-of-the-method [14] and gives 68.5% AUC on the ShanghaiTech. The ablation study demonstrates the effectiveness of our proposed framework.

The remainder of this paper is structured as follows. Section 2 presents a brief analysis of existing anomalous event detection algorithms. Section 3 presents the details of our proposed anomalous event detection algorithm. Experimentation and ablation studies are mentioned in Section 4. At last, Section 5 concludes the paper and provides future research directions.

2. Related Works

In this section, we present a brief review of some recent anomalous event detection algorithms. Previously, some particular task-specific anomalous event detectors, including an abandoned object detector [24], traffic monitoring systems [25–27], and crowd violence detector [28], have been proposed. The generalization of these detectors is challenging. Several studies [5–9,29] have used traditional approaches to solve these challenges, such as the sparse coding technique [5,8], or the dictionary learning technique [6,7]. These techniques considered the anomalous event detection task as an outlier-detection task. They train these models only with normal event videos, and these models detect abnormality if they detect outliers from the learned pattern of the model, leading these models vulnerable to false alarms.

As previously mentioned, proper feature extraction is key to an effective anomalous event detection method. Previously, hand-crafted feature extraction methods [30,31] have been proposed in this task, which utilizes trajectory features at a low-level. Low-level trajectory features represent the regular pattern as a sequence of image coordinates. Since trajectory features are mainly conditioned on the object tracking concept, these features are not robust in case of crowded and complex videos with multiple shadows and occlusions. Thus, these methods [30,31] quickly fail in such cases. To overcome these shortcomings, low-level spatial-temporal feature extraction methods, such as histogram of oriented flows (HOF) [32] and histogram of oriented gradients (HOG) [33], have been proposed. Later, Zhang et al. [34] modeled normal patterns with spatial-temporal features by exploiting the Markov random field (MRF). In the study [35], the local optical flow pattern is modeled using a mixture of probabilistic PCA (MPPCA). However, in another study [5], local histograms of optical flow are modeled using an exponential distribution. Mahadevan et al. [7] proposed the Mixture of Dynamic Textures (MDT). However, to extract useful features from the videos, a 3D ResNet [21] is utilized in our study, as inspired by related studies [21,22].

A breakthrough of deep learning techniques in the field of computer vision tasks, such as in image processing [22], action recognition [15,18,21–23,36,37], object detection [38], object tracking [39,40], and re-identification [41,42] tasks, inspired the researchers to explore these techniques in the anomalous event detection task. Consequently, some recent studies [1–4,10,12,13,43–45] were proposed mostly using auto-encoder methods [1,3,12,13] and recurrent neural networks (RNNs) [3,4].

Author Xu et al. [12] demonstrated the efficacy of deep learning methods by designing a multi-layer auto-encoder. To model normal activities, a 3D convolutional auto-encoder (Conv-AE) is proposed by Hasan et al. [13]. Furthermore, convolutional neural networks (CNNs) extract useful spatial features, whereas (RNNs) are generally used for modeling sequential or time-dependent data using its variants of long short-term memory (LSTM). Consequently, in some studies [3,17], a convolutional LSTMs auto-encoder (ConvLSTM-AE) is proposed in which appearance and motion features are extracted simultaneously by leveraging both CNNs and RNNs, which further advances the performance of the Conv-AE based method [13]. In the study of Luo et al. [5], a stacked RNN framework is proposed by utilizing a temporally coherent sparse coding based technique. Furthermore, in another study [16], the authors proposed the combined study of detection and recounting of anomalous events. Moreover, in the task of anomalous event detection, an increase in several event patterns results in a hike of complexity and uncertainty of feature distributions. Therefore, it is challenging to construct a well-generalized model that can be strictly-discriminative for normal events and anomalous events simultaneously.

Additionally, all of these methods utilize the concept of reconstructing the normal video events and consider the large reconstruction errors associated with the anomalous event–a possible disadvantage in the anomalous event detection task, because this technique contains a possible risk of producing smaller reconstruction errors for the anomalous event [2,46] than for the expectation. This statement implies that a model designed using a reconstruction technique can yield a smaller reconstruction error for an anomalous event, resulting in limited discrimination between normal and anomalous events.

Although these models performed better than other traditional methods, [5–8], they failed to efficiently learn the correlation between motion features and appearance features. Moreover, these models [1–4,10,12,13] are complex and have a heavy network to train the anomalous event detection task. These networks are not suitable to train on small datasets efficiently, which also affected these methods' performance. In the study [11], the author has utilized the concept of dynamic skeleton features to detect the abnormality in the video using the message-passing encoder-decoder recurrent network (MPED-RNN). Their network can separate the abnormal sequence from the normal sequence after training it with the help of normal videos.

Moreover, Xu et al. [12] have exploited unsupervised learning. The relation between motion features and appearance features has been learned by training three auto-encoders using both early and late fusion techniques. In our proposed framework, DMRMs, we have made the network learn these kinds of dependency using a single network and intermediate fusion technique. In the state-of-the-art framework [14] on the UCF-Crime dataset, the authors have exploited both normal and anomalous videos for training the anomalous event detection network in a weakly-supervised manner using C3D [47] and MIL. Our proposed framework is motivated by [12,14]. Compared to them, we include context-dependency learning, an important point to consider to detect an anomalous event. Details are given in the following sections.

3. The Proposed Framework for Anomalous Event Recognition

The proposed framework for anomalous event recognition, Deep-network with Multiple Ranking Measures (DMRMs) shown in Figure 1, consists of two modules or phases: training and inference. In the training phase, our network is able to detect the presence of an abnormality acquired through the joint learning of deep motion and appearance features. These features undergo a data fusion technique [12,48] after being extracted from a 3D residual network (3D ResNet) [21]. This training takes place with the help of deep multiple instance learning (DMIL) [14,19,20] in a weakly-supervised manner, using the proposed multiple ranking measures (MRMs). In the inference phase, DMRMs predicts the abnormality score along with the location of moving objects and an event type (road accident, robbery, etc.). The following sections provide details about the training phase along with feature extraction in Sections 3.1 and 3.2. The joint learning model & MRMs are explained in Section 3.3, and finally, the inference phase is explained.

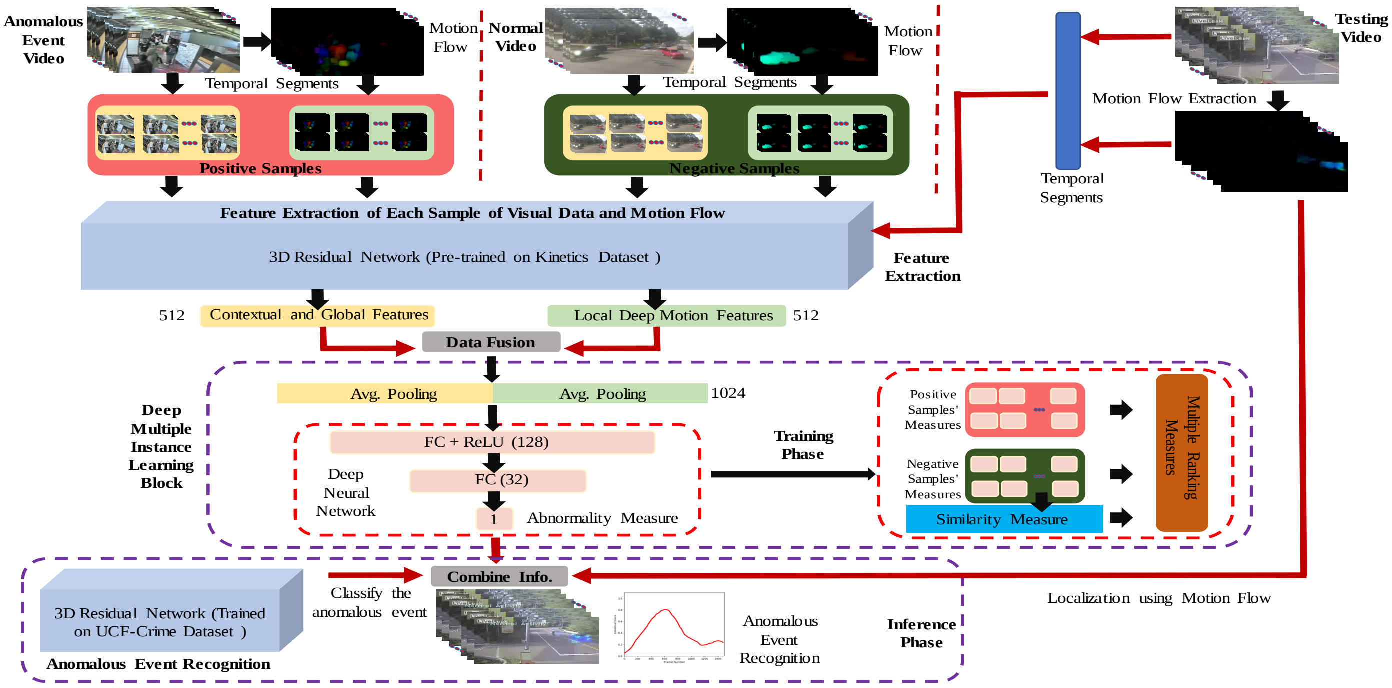

Anomalous Event Video
Motion Flow
Temporal Segments
Positive Samples
Normal Video
Motion Flow
Temporal Segments
Negative Samples
Testing Video
Motion Flow Extraction
Temporal Segments
Feature Extraction of Each Sample of Visual Data and Motion Flow
3D Residual Network (Pre-trained on Kinetics Dataset)
Feature Extraction
512
Contextual and Global Features
Local Deep Motion Features
512
Data Fusion
Deep Multiple Instance Learning Block
Avg. Pooling
Avg. Pooling
1024
FC + ReLU (128)
Deep Neural Network
FC (32)
1
Abnormality Measure
Training Phase
Positive Samples' Measures
Negative Samples' Measures
Similarity Measure
Multiple Ranking Measures
3D Residual Network (Trained on UCF-Crime Dataset)
Anomalous Event Recognition
Classify the anomalous event
Combine Info.
Anomalous Event Recognition
Inference Phase
Localization using Motion Flow

Figure 1. The proposed framework for the anomalous event recognition framework, DMRMs. Training phase: All videos, along with their respective motion flow maps, are classified into two groups of samples: positive (anomalous) or negative (normal). Contextual/global features, as well as local deep motion features, are extracted using their respective techniques. These two types of features are then fused in order to be passed through the DMIL block for joint learning. The proposed MRMs aid this learning in a weakly supervised manner. Inference phase: The testing video is passed through the network, which predicts the abnormality score of each frame along with the position of its moving objects and recognizes the type of anomalous events (road accident, robbery, etc.).

Training Phase

3.1. Extraction of Motion Flow Maps

The proposed framework, DMRMs, begins by extracting each frame's motion flow map from both normal and anomalous videos. This work uses a simpler technique than other previously published studies [9,12,13] have been used, which is Gunner Farneback's dense optical flow algorithm [49]. This algorithm gives us motion-related information, such as the direction and velocity of moving (foreground) objects. It tracks every pixel's movement information from two consecutive frames. To extract deep features from motion flow maps, a dense set of motion features is required. Therefore, in this work, we opted not to use Lucas–Kanade's optical flow algorithm [50], which extracts motion flow with a sparse set of features. Here, the term "dense" indicates that optical or motion flow is measured for each pixel of the frame. The example of extracted motion flow maps is shown in Figure 2. These features give us the local motion information of moving objects, which can be further combined with the global information in order to teach the network about the correlation among moving objects, as explained in the following sections.

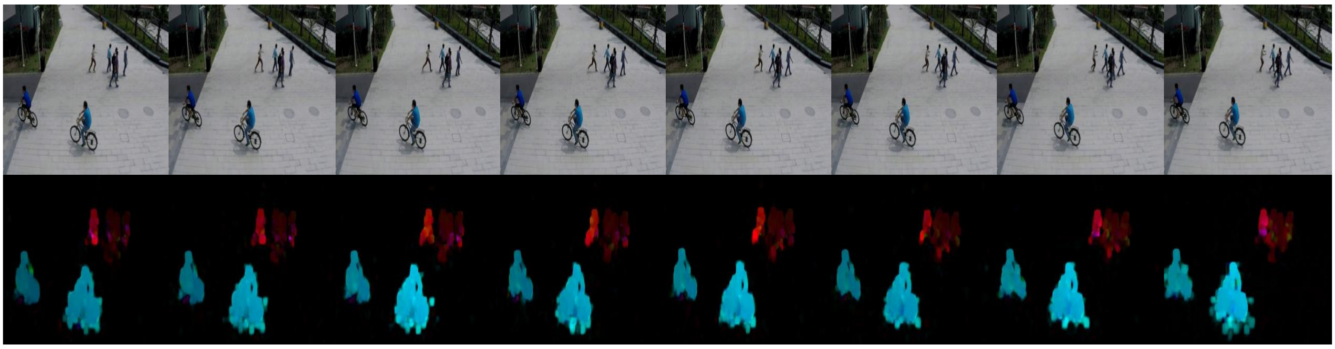


Figure 2. An example of motion flow maps extraction from the ShanghaiTech dataset [5]: the upper row represents eight consecutive frames, and the lower row shows their corresponding motion flow maps.

3.2. Extraction of Deep Motion and Appearance Features

Prior to extracting the deep motion and appearance features in Figure 1, we divide all training videos (both normal and anomalous, along with their motion flow maps) into a matching number of non-overlapping temporal segments. These video segments (N segments) and their respective motion flow segments (N segments) are classified as either positive (anomalous) or negative (normal) samples. In this work, we exploited the information from video-level annotations as follows. We group all segments from anomalous event videos as the positive samples group, G_a, and all segments from normal event videos are grouped as the negative samples group, G_n.

After dividing them into groups, each video sample, along with its motion flow sample, is individually sent through the pre-trained 3D ResNet [21]. It extracts spatial-time-dependent features (from the video sample) and deep motion flow features (from the motion flow samples), such as relative motion information of objects and local variation of motion flow. Though 3D ResNet captures spatial-time-dependent features from visual data, we still separately extract deep motion flow using 3D ResNet. The reason for extracting deep motion flow is to extract and utilize the video's maximum motion information, like relative motion and local flow information, which can not be extracted only using 3D ResNet for visual data. Consequently, it further helps our framework, DMRMs, in learning context-dependency more accurately (shown in Section 4.4.2). Previous work [12] has required training three separate networks in order to extract all of these features (appearance, motion, and their correlation features). However, in our framework, DMRMs, we use only a single pre-trained network to extract them all. We use 3D ResNet-34, which was pre-trained on the Kinetics dataset [51], a vast and challenging action-recognition dataset. As demonstrated by authors S. Tiago et al. [52], pre-trained convolutional neural networks (CNNs) are considered effective in extracting features for the task of anomalous event detection. Additionally, K. Hara et al. [21,22] demonstrates that a 3D ResNet outperforms any other 3D CNNs in the action recognition task. Motivated by these studies [21,22,52], our work utilizes a pre-trained 3D ResNet for its efficacy as a spatial-time-dependent feature extractor for an anomalous event detection task. Its effectiveness is shown in Section 4.4.3.

Moreover, we trained the 3D ResNet-34 on the UCF-Crime dataset [14] to recognize the type of anomalous events such as road accidents, robbery, etc. The results of activity recognition appear in Section 4.4.4. Simultaneously, Section 4.4.4 also shows the network's effectiveness in this task on the UCF-Crime dataset.

3.3. Proposed Joint Learning Model and Multiple Ranking Measures (MRMs)

The spatial-time-dependent features extracted (explained in Section 3.2) from the video segments or samples provide the network with contextual and global features of those segments or samples. In contrast, deep motion flow features such as relative motion information of objects and local variation of motion flow, provide the network with local motion features of moving objects. Hence, we propose joint learning of these features to learn the correlations among all these features and detect the extent of abnormality, as shown in Figure 1. Details are given below:

3.3.1. Data Fusion

Both deep motion flow and spatial-time-dependent features are fused before passing them on to further processing. The extracted features from 3D ResNet-34 [21] of both motion flow maps as $\{m_1, m_2, ..., m_D\}$ (represents features of single instance) and video samples as $\{f_1, f_2, ..., f_D\}$ (represents features of single instance) are concatenated together at the last fully-connected layer of 3D ResNet-34 using the feature-level fusion technique [53], where $m_i \in \mathbb{R}, f_i \in \mathbb{R}$ and D equals to 512. Resulting in a large feature vector as $\{m_1, m_2, ..., m_D, f_1, f_2, ..., f_D\}$. The feature-level fusion helps the model learn a joint representation of each of the extracted features from different or similar models. Then, we pass the result to the DMIL block for joint learning, as shown in Figure 1.

3.3.2. Deep Multiple Instance Learning Block

The joint learning of both concatenated features is conducted in the DMIL block with the help of the proposed MRMs in a weakly-supervised manner, as shown in Figure 1. Anomalous events are infrequent in the real-world, and annotating all videos at the temporal-level is laborious and time-consuming; therefore, utilizing video-level annotations is an essential advantage of using DMIL. Moreover, training a deep neural network with any available information is advantageous. Therefore, a DMIL approach [19,20] is used similar to Sultani et al. [14] work in the absence of temporal annotations.

As introduced in Section 3.2, the group with positive (anomalous) samples is represented as G_a, while the group with negative (normal) samples is G_n. Both of them are then updated with their concatenated extracted features. Each group has N number of samples. Exploiting video-level annotations, and with the help of the proposed MRMs, the DMIL block trains the network to predict the abnormality score of each sample, given that at least one of the positively annotated samples contains an abnormality in it. Since normal video samples have no abnormalities, this result also indicates the accurate annotations of negative samples.

Therefore, in the absence of accurate annotations of positive samples, an optimization function, shown in Equation (1), can be used with the highest scored instance from each group or video. Note that we use an optimization function that is different from the optimization function of the standard supervised binary classification (where instance-level annotations are available) [20].

$$\min_{\mathbf{w}} \frac{1}{v} \sum_{l=1}^{v} \overbrace{\max\left(0, 1 - Y_{G_l}\left(\max_{s \in G_l}(\mathbf{w}.\phi(x_s)) - b\right)\right)}^{\widehat{h}} + \frac{1}{2} \parallel \mathbf{w} \parallel^2 \tag{1}$$

where v is the total number of training videos, Y_{G_l} represents the video-level or group-level annotation (groups G_n and G_a), $\phi(x_s)$ stands for feature representation of a sample (an instance), $\widehat{h}$ denotes the hinge loss function, b stands for a bias term, and $\mathbf{w}$ denotes model weights. The following section describes the proposed objective function for DMIL.

3.3.3. Proposed Multiple Ranking Measures

Similar to Sultani et al. [14] work, we also considered the anomalous event detection problem as a regression problem, where we want the sample with an anomalous event to have a greater abnormality score than the normal event instance. To achieve this, one of the possible solutions could be the ranking loss to train our network. Ranking loss with instance-level or sample-level annotations can be mentioned as follows:

$$R(S_a) > R(S_n) \tag{2}$$

where anomalous videos' samples and normal videos' samples are represented as S_a and S_n, respectively. Functions $R(S_a)$ and $R(S_n)$ stand for the corresponding predicted abnormality scores. The range for the abnormality score function $R(x)$ is between 0 and 1.

Because of the unavailability of instance-level or sample-level annotations, Equation (2) cannot be used. Consequently, we propose a new objective function using MRMs to train our model using DMIL. Moreover, the idea of proposing a new objective function is motivated by another critical goal of reducing the false alarm rates, which may be produced by our proposed anomalous event detection framework, DMRMs.

The following are two cases of false alarms:

(Case 1) False positive: The detector predicts a normal event as an anomalous event.

(Case 2) False negative: The detector predicts an anomalous event as a normal event.

To alleviate the aforementioned false alarms and develop a more suitable ranking loss function than the one described in Equation (2), as well as to train our model using DMIL, the following ranking measure conditions are proposed in Equations (3) and (4).

$$\max_{i \in S_a} R\left(S_a^i\right) > \max_{i \in S_n} R\left(S_n^i\right) \tag{3}$$

$$\max_{i \in S_a} R\left(S_a^i\right) > \min_{i \in S_a} R\left(S_a^i\right) \tag{4}$$

Equations (3) and (4) are proposed to avoid the false alarm in case 1. In Equations (3) and (4), the highest-ranked instance or sample of the positive group (S_a)–most likely a true positive–is compared with the highest-ranked instance or sample of the negative group (S_n) and the lowest-ranked sample of the positive group (S_a), respectively–most likely false positives.

We still need to alleviate the second type of false alarms of case 2. To this end, we proposed additional ranking measure conditions in Equations (6) and (7). The abnormality scores of each instance of the positive group are arranged in descending order of scores as described below:

$$[J_1, J_2, J_3, ..., J_N] = order_{i \in S_a}^{desc} R\left(S_a^i\right) \tag{5}$$

where each group has a total N number of instances.

$$J_2 > \max_{i \in S_n} R\left(S_n^i\right) \tag{6}$$

$$J_3 > \max_{i \in S_n} R\left(S_n^i\right) \tag{7}$$

$$...$$

Equations (6) and (7) are proposed to avoid both cases (case 1 and case 2) of false alarms. In Equations (6) and (7), the second-highest and third-highest ranked samples of the positive group (S_a), respectively, are compared with the highest-ranked sample of the negative group (S_n)–most likely a false positive.

Since the training videos are large in size and lengthy, there is a possibility of the presence of anomalous events in multiple instances (as shown in Section 4.4.2). Therefore, we propose the aforementioned ranking conditions, Equations (6) and (7), to reduce the incidence of false alarm cases, 1 and 2. The proposed conditions also maximize the abnormality scores of instances or samples of the positive group (S_a) and minimize those of the negative group (S_n). However, a comparison between a less highly ranked positive instance (as in Equation (7) and so on) and the highest-ranked instance of negative groups starts affecting the framework's accuracy (as shown in Section 4.4.3). Therefore, based on the conducted experiments, we have excluded all the ranking conditions with less highly ranked instances, including Equation (7) ranking condition, and so on. Moreover, segment-level or instance-level annotations are not required in these ranking conditions.

In addition to these ranking conditions, there is one more ranking measure condition, Similarity Measure (SM), which needs to be considered, as follows:

$$SM(S_n) = \sum_{i}^{n-1}\left(R\left(S_n^i\right) - R\left(S_n^{i+1}\right)\right)^2 \tag{8}$$

Equation (8) refers to the fact that consecutive instances of a negative group should have similar features. Therefore, the model should predict approximately the same abnormality score (close to zero) for all such instances. This condition forces the network to learn the similarity of features among the instances of a negative group, leading the network to understand the similarity among the instances of positive groups.

To define MRMs, all predicted abnormality scores of negative group instances should be kept far apart from all the positive group instances. Consequently, MRMs (MRM) is a combination of all the aforementioned ranking condition measures. Therefore, the MRMs is defined as the following hinge loss formulation and satisfy all the ranking measure conditions as mentioned earlier from Equations (3)–(8) except Equation (7):

$$MRM(S_a, S_n) = rm_1(S_a, S_n) + rm_2(S_a, S_a) + rm_3(S_a, S_n) + \gamma_1 SM(S_n) \tag{9}$$

where γ_1 is the model's hyper-parameters, and ranking measures $rm_1(S_a, S_n), rm_2(S_a, S_a)$, $rm_3(S_a, S_n)$, and $rm_4(S_a, S_n)$ are described in Equations (10)–(13), respectively, as follows:

$$rm_1(S_a, S_n) = \max\left(0, 1 - \max_{i\in S_a} R\left(S_a^i\right) + \max_{i\in S_n} R\left(S_n^i\right)\right) \tag{10}$$

$$rm_2(S_a, S_a) = \max\left(0, 1 - \max_{i\in S_a} R\left(S_a^i\right) + \min_{i\in S_a} R\left(S_a^i\right)\right) \tag{11}$$

$$rm_3(S_a, S_n) = \max\left(0, 1 - J_2 + \max_{i\in S_n} R\left(S_n^i\right)\right) \tag{12}$$

$$rm_4(S_a, S_n) = \max\left(0, 1 - J_3 + \max_{i\in S_n} R\left(S_n^i\right)\right) \tag{13}$$

The ranking loss mentioned by Sultani et al. [14] does not consider both cases of false alarms. Considering both cases result in a reduced false alarm rate for our network in comparison to [14]. Yet similarly to [14], the two following constraints (ε_1 and ε_2) are utilized in order to preserve the sparsity ε_1 in Equation (14) and temporal smoothness ε_2 in Equation (15) of the abnormality score:

$$\varepsilon_1 = \gamma_2 \sum_{i}^{n} R\left(S_n^i\right) \tag{14}$$

$$\varepsilon_2 = \gamma_3 \sum_{i}^{n-1}\left(R\left(S_a^i\right) - R\left(S_a^{i+1}\right)\right)^2 \tag{15}$$

where γ_2, and γ_3 are the model's hyper-parameters. Therefore the new proposed objective function using MRMs for the DMIL can be defined as follows:

$$\Im(S_a, S_n) = MRM(S_a, S_n) + \varepsilon_1 + \varepsilon_2 \tag{16}$$

Finally, our proposed objective function, along with a regularization term, is mentioned as follows:

$$m(\mathbf{W}) = \Im(S_a, S_n) + \gamma_4 \parallel \mathbf{W} \parallel^2 \tag{17}$$

where **W** represents model weights and γ_4 is the model's hyper-parameters.

Consequently, the network learns to predict the abnormality scores for all instances or samples of both groups using the aforementioned objective function for DMIL and using a large number of positive and negative instances from the training dataset. With our proposed framework, DMRMs, the network learns the context-dependency of the normal and anomalous videos (shown in Section 4.4.2). The framework also addresses the problem of the unavailability of a large annotated dataset by exploiting the video-level annotations.

Inference Phase

During the inference phase, the video is given to the proposed framework (DMRMs), and its motion flow maps are then extracted. These maps, along with the video, are divided into temporal segments. Subsequently, these segments are passed through the feature extraction process. Data fusion occurs, and each segment's abnormality score is predicted using the trained deep neural network. Finally, the type of anomalous event is predicted using the trained 3D ResNet on the UCF-Crime dataset [14], along with the position of moving objects in the video.

4. Experimentation

4.1. Datasets

There are numerous standard datasets [5,8,14,24,54,55] available for the anomalous event detection task. Despite that, we chose to perform our experimentation using the proposed framework, DMRMs, on two very recent and challenging datasets for the anomalous event detection task: UCF-Crime [14] and ShanghaiTech [5] datasets. These datasets have videos that are long in duration, large in size, and consist of real-world scenarios of anomalous events. Moreover, to train our network efficiently and accurately, we needed a large number of anomalous and normal videos.

The UCF-Crime dataset consists of a total of 1900 long, untrimmed real-world videos, including 950 normal videos and 950 anomalous event videos. It has 13 classes of real-world anomalous events such as robberies, car accidents, fighting, and more. We also trained our activity recognition network, 3D ResNet-34 [21], using this dataset.

The ShanghaiTech dataset consists of a total of 437 videos (with 13 different scenes and $316,154$ frames)–that is, 330 training videos (normal videos) and 107 test videos (anomalous videos like cycling and skating on the pedestrian-only path). This dataset contains videos with different camera angles and lighting conditions.

Training and Testing Datasets: In DMRMs, the network was trained individually on both datasets. First, the UCF-Crime dataset was divided into a training dataset and a testing dataset. Whereas the training group contains 800 normal and 810 anomalous event videos, the testing group contains 150 normal and 140 anomalous event videos. It is worth mentioning that the dataset was divided in such a way that both the training and testing datasets each contained all 13 classes of anomalous events. Second, our network was also trained on the ShanghaiTech dataset, which was already divided into training and testing datasets. However, we needed both normal and anomalous videos to train our network; therefore, we used a portion of the given testing dataset (randomly chosen 40 testing videos) as an anomalous video group for the training phase. Moreover, we used only 30% of the given training dataset (normal event videos). Moreover, both the training and testing datasets contain all 13 scenes. Using less part of the training dataset was an advantageous part of our proposed framework, DMRMs.

4.2. Implementation Details

As mentioned in Section 3.2, all videos from the training and testing datasets and their associated motion flow maps were divided into a specific number (N) of non-overlapping temporal segments, though the AUC of the proposed detector, DMRMs, is unaffected by utilizing multi-scaled overlapped temporal segments. In our experiments, we chose to divide them into 16 segments (N = 16). However, the ablation study (Section 4.4) demonstrated the effect of choosing a different number of segments N. The input size for the 3D ResNet-34 was 112 × 112 pixels; the results were similar for size 160 × 160 (AUC = 68.54 on the ShanghaiTech Testing dataset). For training on the UCF-Crime dataset, 30 frames per second (fps) was chosen. Then, 60 randomly chosen videos (30 normal and 30 anomalous) were passed through the proposed framework to train it for the anomalous event detection task. However, for training on the ShanghaiTech dataset, 24 fps was chosen, and only 8 randomly chosen videos (four normal and 4 anomalous) were passed through the framework, DMRMs.

Afterward, a 3D ResNet-34 [21] extracted 512 features from every 16 frames of the video segments/samples as well as 512 features from their respective motion flow maps, as shown in Figure 1. Here, $3-$channels of motion flow maps are constructed by repeating $1-$channel information to all $3-$channels. These extracted features were fused in the data fusion block. Next, these integrated features were passed through a three-layer fully connected (FC) deep neural network, as shown in Figure 1 for the joint learning of these features. We have also used a 60% dropout among all 3 FC layers, similar to previous work [14].

The proposed objective function described in Equation (17) using MRMs was used to train DMIL along with an Adagrad optimizer with an initial learning rate of 0.001 and 0.0001 for UCF-Crime and ShanghaiTech datasets, respectively. In the case of the ShanghaiTech dataset–all of the hyper-parameters $\gamma_1, \gamma_2, \gamma_3$, and γ_4 described in the proposed objective function in Equation (17) have values of 8×10^{-5}, 8×10^{-5}, 8×10^{-5}, and 0.01, respectively. Similarly in the case of UCF-Crime dataset, all of the hyper-parameters, $\gamma_1, \gamma_2, \gamma_3$, and γ_4 have values of $8 \times 10^{-5}, 8 \times 10^{-4}, 8 \times 10^{-4}$, and 0.01, respectively. These values achieved the best performance from DMRMs, as shown in Table 1. The proposed network was trained for $15,000$ epochs for the UCF-Crime dataset and 5000 epochs for the ShanghaiTech dataset.

For the anomalous event recognition task, we trained a 3D ResNet-34 on the UCF-Crime dataset using a 4-fold cross-validation strategy and using 50 videos from each category (normal and anomalous), dividing them into the ratio of $75:25$ (see results in Section 4.4.4).

4.3. Evaluation Metrics

In this study, the effectiveness of our proposed framework, DMRMs, was evaluated using a receiver operating characteristics curve (ROC-Curve), which is a trade-off plot between False Positive Rate (FPR) and True Positive Rate (TPR), and the area under the curve (AUC) metric. This metric is a quantitative measure for the ROC-Curve, which has been used in similar previous studies [8,13,14] for evaluating the anomalous event detection task. Moreover, FPR and TPR can be calculated as follows:

$$TPR\ or\ Recall = \frac{\sum TP}{\sum (TP+FN)}$$

$$FPR = \frac{\sum FP}{\sum (TP+FN)} \tag{18}$$

where *TP*, *FN*, and *FP* stand for True Positive, False Negative, and False Positive, respectively. Additionally, the performance of the anomalous event recognition task is evaluated using an accuracy metric.

4.4. Experiment Results and Ablation Study

4.4.1. Comparison among State-of-the-Art Methods

The UCF-Crime and ShanghaiTech datasets are the newest datasets; therefore, only recent algorithms [5,8,9,11,13,14,44,45,56] have reported their results on these datasets. In the above-mentioned algorithms, the training phase is different from our framework, DMRMs. Unlike ours, most of them have used only normal videos to train their deep neural networks. However, we have used the same UCF-Crime and ShanghaiTech testing datasets for fair comparison among all algorithms.

Table 1. Effect of different values of Hyper-parameters on the AUC of UCF-Crime and ShanghaiTech datasets.

Hyper-Parameters	γ_1	γ_2	γ_3	γ_4	**lr**	**UCF-Crime**	**ShanghaiTech**
1	8×10^{-5}	8×10^{-5}	8×10^{-5}	0.01	0.001	81.21	65.47
2	8×10^{-5}	8×10^{-5}	8×10^{-5}	0.01	0.01	77.23	-
3	8×10^{-5}	8×10^{-5}	8×10^{-5}	0.01	0.0001	78.53	68.50
4	8×10^{-5}	8×10^{-5}	8×10^{-5}	0.01	0.00001	-	65.83
5	1×10^{-5}	1×10^{-5}	1×10^{-5}	0.01	0.001	78.44	-
6	1×10^{-5}	1×10^{-5}	1×10^{-5}	0.01	0.0001	-	67.32
7	8×10^{-5}	8×10^{-5}	8×10^{-5}	0.1	0.001	77.80	-
8	8×10^{-5}	8×10^{-5}	8×10^{-5}	0.1	0.0001	-	65.70
9	8×10^{-5}	8×10^{-5}	8×10^{-5}	0.001	0.001	79.51	-
10	8×10^{-5}	8×10^{-5}	8×10^{-5}	0.001	0.0001	-	65.83
11	8×10^{-4}	8×10^{-4}	8×10^{-4}	0.01	0.001	81.40	-
12	8×10^{-4}	8×10^{-4}	8×10^{-4}	0.01	0.0001	-	67.72
13	8×10^{-6}	8×10^{-6}	8×10^{-6}	0.01	0.001	78.56	-
14	8×10^{-6}	8×10^{-6}	8×10^{-6}	0.01	0.0001	-	67.86
15	8×10^{-5}	8×10^{-5}	8×10^{-4}	0.01	0.001	81.62	-
16	8×10^{-5}	8×10^{-5}	8×10^{-4}	0.01	0.0001	-	67.75
17	8×10^{-5}	8×10^{-4}	8×10^{-5}	0.01	0.001	78.12	-
18	8×10^{-5}	8×10^{-4}	8×10^{-5}	0.01	0.0001	-	67.68
19	8×10^{-4}	8×10^{-5}	8×10^{-5}	0.01	0.001	78.52	-
20	8×10^{-4}	8×10^{-5}	8×10^{-5}	0.01	0.0001	-	67.84
21	8×10^{-5}	8×10^{-4}	8×10^{-4}	0.01	0.0001	-	67.71
22	8×10^{-4}	8×10^{-5}	8×10^{-4}	0.01	0.0001	-	67.82
23	8×10^{-4}	8×10^{-4}	8×10^{-5}	0.01	0.0001	-	67.85
24	8×10^{-4}	8×10^{-4}	8×10^{-5}	0.01	0.001	78.98	-
25	8×10^{-4}	8×10^{-5}	8×10^{-4}	0.01	0.001	78.62	-
26	8×10^{-5}	8×10^{-3}	8×10^{-3}	0.01	0.001	80.64	-
27	8×10^{-5}	8×10^{-6}	8×10^{-6}	0.01	0.001	81.48	-
28 (ours)	8×10^{-5}	8×10^{-4}	8×10^{-4}	0.01	0.001	81.91	-
	8×10^{-5}	8×10^{-5}	8×10^{-5}	0.01	0.0001	-	68.50

This study compares our proposed anomalous event detection framework, DMRMs, with other state-of-the-art methods, Hasan et al. [13], Sohrab et al. [44], Lu et al. [8], GODS [56], Sultani et al. [14], and an SVM binary classifier using ROC-Curve and AUC metric on the UCF-Crime dataset. As shown in Figure 3, our proposed framework, DMRMs, outperforms all state-of-the-art methods, including the recent algorithm by Sultani et al. [14], which was the primary motivation for the present work, and gives the best FPR-TPR trade-off curve among them. Similarly, from Table 2, the highest AUC score reflects the effectiveness of our proposed framework. Moreover, AUC mentioned by authors Sohrab et al. [44] in Table 2 was reported in the paper GODS [56].

Similarly, the performance of our proposed framework, DMRMs, achieves competitive results on the ShanghaiTech dataset using the AUC metric evaluation method when compared with the state-of-the-art methods, as shown in Table 2. Here, we have evaluated the AUC of each previously proposed algorithms [5,9,11,13,45] using the same experimental setup on the testing dataset as ours. The AUC metric is calculated on the overall ShanghaiTech testing dataset (107 testing videos) in the Hasan et al. [13]* algorithm.

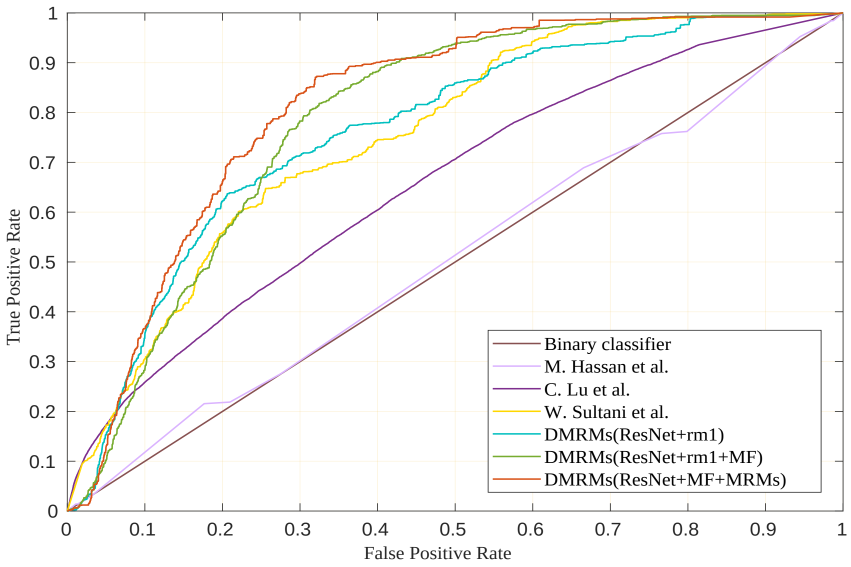


Figure 3. ROC-Curves comparison on UCF-Crime dataset on state-of-the-art methods, M. Hasan et al. [13], C. Lu et al. [8], Sultani et al. [14]. Our proposed framework's variations curves are shown in associated ROC-curves. The ROC-Curve of our proposed framework, DMRMs (3D ResNet+MF+MRMs), including the variation of it, shows the best FPR-TPRtrade-off ROC-Curve, where MF stands for Motion Flow maps.

Table 2. Frame-level AUC-Comparison (in %) with other state-of-the-art methods on UCF-Crime [14] and ShanghaiTech [5] datasets. Bold figures stand for the best performance in all.

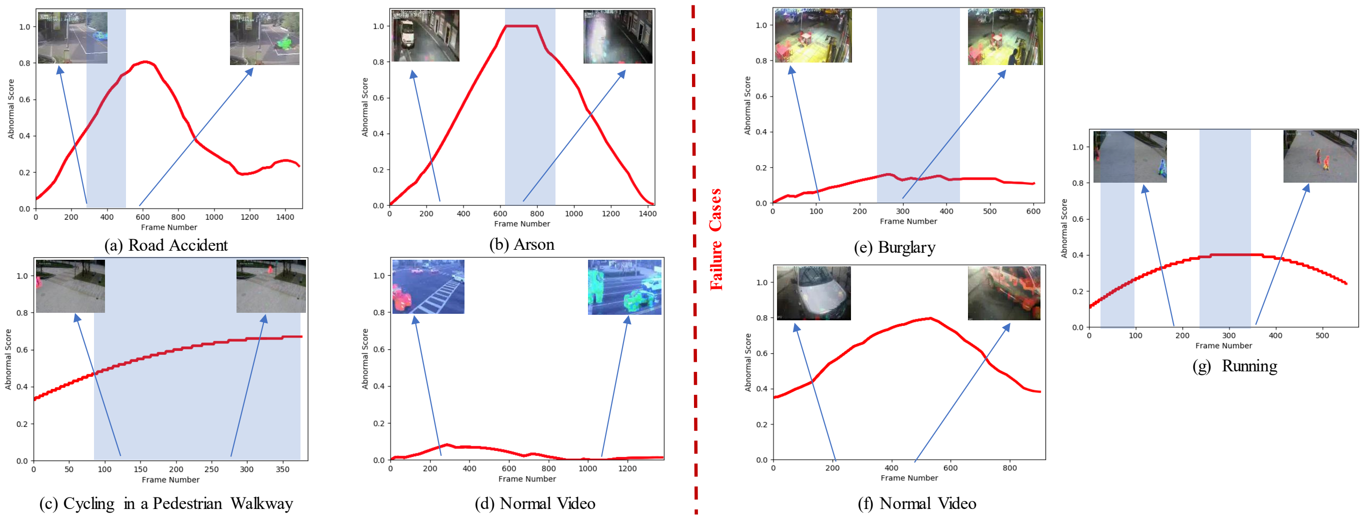
Algorithm	UCF-Crime	ShanghaiTech
SVM Binary Classifier	50.00	-
M. Hasan et al. [13]	50.60	60.90 *
F. Sohrab et al. [44]	58.50	-
C. Lu et al. [8]	65.51	-
Luo et al. [5]	-	60.73
GODS [56]	70.46	-
Autoreg-ConvAE (LLK) [45]	-	70.30
Liu et al. [9]	-	73.04
Autoreg-ConvAE (NS) [45]	-	73.85
MPED-RNN [11]	-	**76.24**
W. Sultani et al. [14]	75.41	-
Ours DMRMs	**81.91**	68.50

Table 2 shows that our framework, DMRMs, performs quite well on the UCF-Crime dataset compared to the ShanghaiTech dataset. There are mainly two reasons that our framework, DMRMs, does not function well on the ShanghaiTech dataset. First, our model requires a large amount of dataset to train the proposed architecture as it utilizes 3D ResNet-34 to extract deep features from the video. As 3D ResNet is not explored adequately in the anomalous activity detection task, consequently, according to the action recognition task's studies [22,23], similar to our task, 3D ResNet overfits on the smaller dataset. Moreover, compared to the ShanghaiTech dataset UCF-Crime dataset is vast. Second, the ShanghaiTech dataset has much noisier annotations compare to the UCF-Crime dataset, and this dataset includes videos with a lot of illumination and camera angle variations. Recently, the authors [18] have utilized the graphical noise cleaner approach resulting in better performance than our method. However, our study does not include noise cleaner. Therefore, there is an enormous variation in our algorithm's performance between these two datasets, and we can conclude that the proposed framework does not perform well for smaller and noisier datasets like the ShanghaiTech dataset.

4.4.2. Qualitative Results and Demonstration of Context-Dependency Learning of DMRMs

The qualitative results of our proposed framework are shown in Figures 4 and 5, after testing our framework on videos from the testing datasets of UCF-Crime and ShanghaiTech.

Figure 4 demonstrates how well our network learned the anomalous event detection task because of our framework's context-dependency learning. However, it failed in some cases like (e), (f) and (g), where (e) demonstrates the failure case in which DMRMs was unsuccessful due to extreme variation in lighting conditions, though success is achieved in (d) in the presence of less lighting variation. Figure 4f demonstrates the case where DMRMs failed to differentiate between normal "car washing" and "vehicle theft" cases, while (g) shows continuous anomalous event detection along with a low abnormality score because of random movements. On the other hand, (a), (b), (c), and (d) demonstrate the accurate detection of different anomalous events with different backgrounds and lighting conditions along with the location of moving objects. These show effectiveness of the proposed framework, DMRMs: (a), (b), (c) and (d) depict the accurate detection of a "road accident," "arson," "cycling in a pedestrian walkway," and a "normal event," respectively.



(a) Road Accident

(b) Arson

(e) Burglary

(g) Running

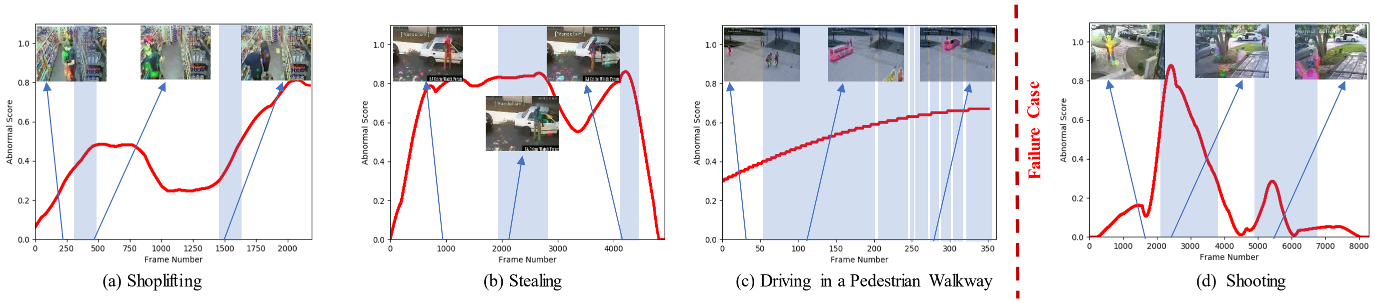
(c) Cycling in a Pedestrian Walkway

(d) Normal Video

(f) Normal Video

Figure 4. Qualitative analysis of the proposed framework, DMRMs, on UCF-Crime and ShanghaiTech testing datasets. The red curve shows the abnormality score of the corresponding frame (ranging between 0 and 1). The light blue box shows the ground truth of the temporal region of anomalous events, while the arrow indicates the corresponding visual frame of the recognized anomalous event along with the localization of moving objects. (**a**,**b**,**d**–**f**) are testing results from UCF-Crime dataset and (**c**,**g**) are from ShanghaiTech dataset, where (**e**–**g**) are failure cases of our proposed framework, DMRMs. Meanwhile, all successful results of anomalous event recognition are shown in (**a**,**b**,**d**,**e**).

Figure 5 demonstrates how well our network learned the anomalous event detection task when there are multiple anomalous events in a single video of proposed DMRMs (effect of proposed MRMs). However, it failed in some cases like (d), which demonstrates the failure case in which a "shooting" is detected only a single time, although it is present two times in the same video, a mistake occurring due to occlusion by trees, which hid some extent of information. Besides, (a), (b), and (c) demonstrate the accurate detection of multiple instances of anomalous events in a single video with different backgrounds and lighting conditions (illumination variations) along with the location of moving objects. These illustrate the effectiveness of DMRMs (especially proposed MRMs) in the presence of multiple instances of anomalous events: (a), (b) and (c) depict the detection of "shoplifting," "stealing," and "driving a car in a pedestrian walkway," respectively. These multiple anomalous events were detected with a significant abnormality score in a single video.

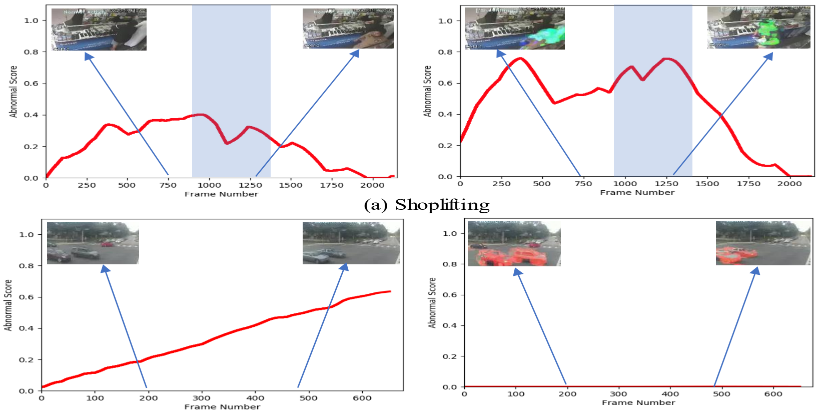


(a) Shoplifting (b) Stealing (c) Driving in a Pedestrian Walkway (d) Shooting

Figure 5. Qualitative analysis of the proposed framework, DMRMs, for multiple instances of anomalous events in a single video. The red curve shows the abnormality score of the corresponding frame (ranging between 0 and 1). The light blue box shows the ground truth of the temporal region of anomalous events, while the arrow indicates the corresponding visual frame of the recognized anomalous event along with the localization of moving objects. (**a**,**b**,**d**) are testing results from UCF-Crime testing dataset and (**c**) is from ShanghaiTech testing dataset, where (**d**) is a failure case of our proposed framework, DMRMs. Meanwhile, all successful results in recognizing multiple-instance anomalous events are shown in (**a**–**c**).

Demonstration of context-dependency learning appears in Figures 6 and 7. Figure 6 demonstrates the qualitative comparison between the framework without joint learning and our framework (DMRMs) with joint learning. The upper-left image in (a) reflects the result of an anomalous event without any joint learning of motion and appearance features that failed to recognize “shoplifting”—an anomalous event. In contrast, the upper-right image demonstrates the result of our proposed DMRMs after the joint learning technique, showing the successful recognition of “shoplifting”–an event that is very similar to the “normal event of shopping”—through context-dependency learning. Similarly, in (b), the left-hand image gives a false alarm for an anomalous event given that many cars are passing by. In contrast, in the right-hand image, our framework learned the context of this video as well as the difference between “car accident” cases and those of a “normal road with many cars passing by.”

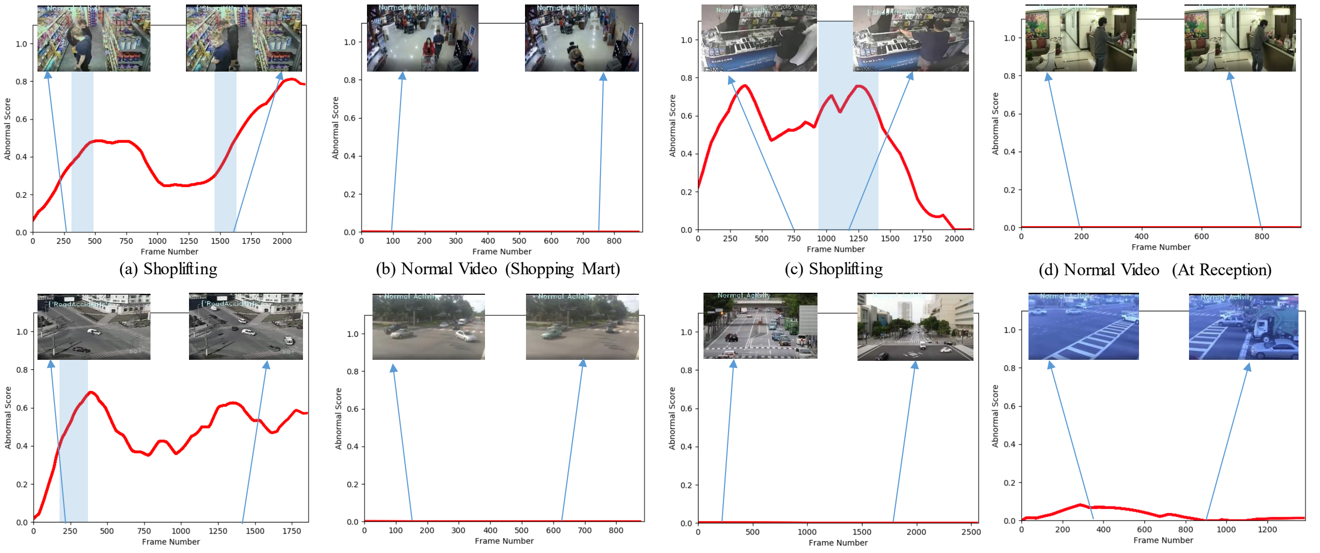
The qualitative analysis of context-dependency learning of DMRMs is demonstrated in Figure 7. For better visualization, the localization of moving objects is avoided in Figure 7. (a)–(h) reflect the result of an anomalous event detection task and context-dependency learning of DMRMs. Events (a), (c), and (e) have background similarities with the events (b), (d), and (f)–(h), respectively. (a) and (b) both are events from shopping marts—(a) is an anomalous event of “shoplifting,” while (b) is a “normal event of shopping.” Similarly, events (c) and (d) take place near the reception area—(c) is an anomalous event of “shoplifting at reception,” while (d) is a “normal event of some query at reception.” Furthermore, events (e) and (f) are cases of driving on crossroads—(e) is an anomalous event of “car accident at crossroads,” while (f) is a “normal event of driving at crossroads.” Additionally, similar to the event (e), events (g) and (h) are also cases of driving on crossroads with different scenarios (much rush on bigger crossroads) and with substantial illumination differences, respectively. Despite all these background similarities, our proposed framework, DMRMs, successfully recognized all the events from (a)–(h), demonstrating the context-dependency learning of DMRMs.



(a) Shoplifting

(b) Normal Video

Figure 6. Qualitative comparison between the framework without joint learning and our proposed framework, DMRMs, with joint learning (or context-dependency learning) on UCF-Crime testing dataset. The red curve shows the abnormality score of the corresponding frame (ranging between 0 and 1). The light blue box shows the ground truth of the temporal region of anomalous events, while the arrow indicates the corresponding visual frame of the recognized anomalous event along with the localization of moving objects. The left-side images in (**a**,**b**) show the cases of the framework without joint learning, whereas the right-side images show the effect of joint learning in the same cases.



(a) Shoplifting (b) Normal Video (Shopping Mart) (c) Shoplifting (d) Normal Video (At Reception)

(e) Road Accident (f) Normal Video (Driving on Crossroads) (g) Normal Video (Driving on Crossroads) (h) Normal Video (Driving on Crossroads)

Figure 7. Qualitative analysis of context-dependency learning of our proposed framework, DMRMs, on UCF-Crime testing dataset. The red curve shows the abnormality score of the corresponding frame (ranging between 0 and 1). The light blue box shows the ground truth of the temporal region of anomalous events, while the arrow indicates the corresponding visual frame of the recognized anomalous event. (**a**,**c**,**e**) depict cases of anomalous events, whereas (**b**,**d**,**f**–**h**) demonstrate normal events cases. Additionally, events (**a**,**c**,**e**) have background similarities with the events (**b**,**d**,**f**,**h**), respectively. Our proposed framework successfully recognizes all the events from (**a**–**h**) though having background similarities, demonstrating the context-dependency learning of DMRMs.

4.4.3. Ablation Study

First, we demonstrate the results of dividing the videos into different numbers of segments, 8, 16, 32, 48, and 64—an essential step of our framework (Figure 1)–by studying their effect of the frame-level AUC metric on the UCF-Crime and ShanghaiTech datasets (given in Table 3). We concluded that the effect of the number of segments was more significant on the ShanghaiTech dataset than on the UCF-Crime dataset, and that dividing the videos into 16 segments provided an optimal performance on both datasets. The pre-trained 3D ResNet-34 [21] extracts features from each video clip of 16-frames. Therefore, in this work, a lower limit of the number of segments is considered 16, although the effect of 8 segments on AUC of ShanghaiTech testing dataset is shown in Table 3. Moreover, dividing the video into 16 segments speeds up our framework (as shown in Table 7), DMRMs. Therefore, we chose $N = 16$ in our study for both datasets.

Second, we demonstrate in Tables 4 and 5 the effectiveness of using our proposed MRMs along with 3D ResNet-34 and motion flow maps, a framework that outperforms the state-of-the-art methods on the UCF-Crime and ShanghaiTech datasets and mitigates the FAR as well. In Table 4, we show the effect of each ranking measure of our framework, DMRMs, on the AUC of both datasets. From Table 4, we can conclude that introducing higher-order ranking measures degrade the performance of the anomalous event detector, while a combination of ranking measures $rm1$, $rm2$, and $rm3$ increases the AUC of our detector. Table 5 shows the lowest FAR value on the UCF-Crime testing dataset, and these FAR values were calculated on the normal videos of the dataset at 50% of the threshold. The smallest FAR value demonstrates that our framework, DMRMs, mitigates false alarms.

Table 3. Comparison of choosing a different number of segments on UCF-Crime and ShanghaiTech datasets.

Segments	**8**	**16**	**32**	**48**	**64**
UCF-Crime (**AUC**)	-	81.91	81.17	80.71	80.98
ShanghaiTech (**AUC**)	68.20	68.50	65.60	66.20	66.00

Table 4. Effectiveness of proposed ranking measures rm_1, rm_2, rm_3, rm_4, MRMs and constraints (sparsity and temporal smoothness) as described in Equations (9)–(15), using 3D ResNet-34 [21] and using motion flow maps (MFM) is shown with frame-level AUC-Comparison (in %) on UCF-Crime and ShanghaiTech datasets. It shows that collectively using 3D ResNet, motion flow and MRMs in our proposed framework contributes to outperforming state-of-the-art methods. Bold figures stand for the best performance in all.

Ranking Measures	**UCF-Crime**	**ShanghaiTech**
3D ResNet-34 + rm_1	76.20	61.20
3D ResNet-34 + rm_1 + MF	78.68	65.30
3D ResNet-34 + rm_1 + MF + rm_2	78.60	64.90
3D ResNet-34 + rm_1 + MF + rm_3	80.08	66.10
3D ResNet-34 + rm_1 + MF + rm_4	80.01	65.60
3D ResNet-34 + rm_1 + MF + rm_2 +rm_3	80.92	67.70
3D ResNet-34 + rm_1 + MF + rm_2 +rm_4	79.80	65.40
3D ResNet-34 + rm_1 + MF + rm_2 + rm_3 + rm_4	80.10	67.00
3D ResNet-34 + MF + MRM	81.12	67.91
DMRMs 3D ResNet-34 + MF + MRM + ϵ_1 + ϵ_2	**81.91**	**68.50**

Table 5. Comparison of false alarm rate (FAR) on UCF-Crime [14] test dataset in normal event videos. Bold figure stands for the best performance in all.

Algorithm	**M. Hasan et al. [13]**	**C. Lu et al. [8]**	**W. Sultani et al. [14]**	**Ours DMRMs**
FAR	27.20	3.10	1.90	**0.85**

4.4.4. Anomalous Event Recognition and Localization of Moving Objects

Figure 8 demonstrates the anomalous event recognition task along with the localization of moving objects. Tables 4 and 6 show the effectiveness of the 3D ResNet [21] used in our study. Table 6 specifically shows its near-similar accuracy on the activity recognition task compared to the work of [57], which uses a heavier network than 3D ResNet. Therefore, we have utilized 3D ResNet for extracting useful features from videos, and its effectiveness is shown in Table 4. Table 6 also demonstrates that the UCF-Crime dataset is a challenging dataset not only in the anomalous event detection task but also in the activity recognition task. Figure 8 shows the success and failure of both cases of the anomalous event recognition task.

4.4.5. Analysis of Computational Cost of The Proposed DMRMs

All the experiments are performed on the Ubuntu 16.04 LTS system with an Intel Core i7 processor along with the NVIDIA GeForce GTX1080 Ti GPU. In Table 7, we have demonstrated the computational cost of the proposed anomalous event detector. We achieve 26 frames per second (FPS) and 18 FPS on UCF-Crime and ShanghaiTech datasets, respectively. As the size of each frame is 856 × 480 and 320 × 240 in ShanghaiTech and UCF-Crime datasets, respectively, therefore, the computational cost of extracting motion flow maps is higher in ShanghaiTech dataset than that of the UCF-Crime dataset, which affects the overall FPS of the anomalous event detector's computational cost.

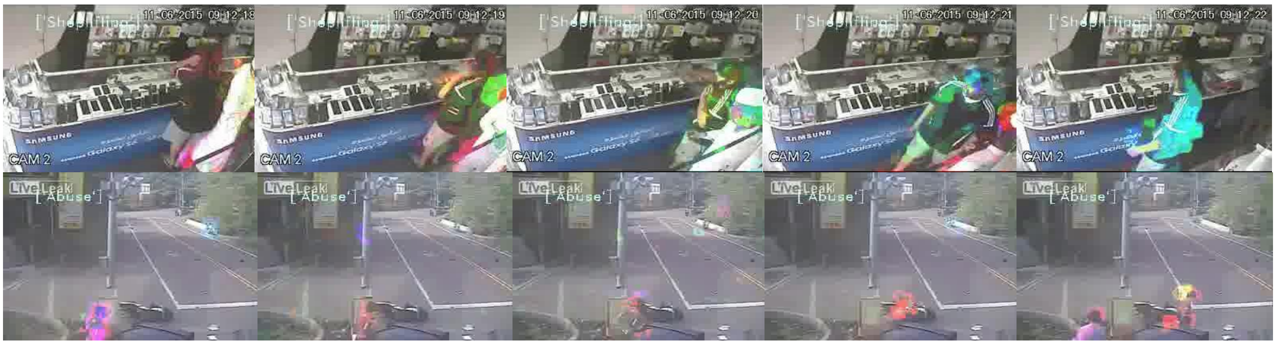


Figure 8. Demonstration of the anomalous event recognition task along with localization of moving objects of our proposed framework, DMRMs: the first and second rows represent five frames from different anomalous events with a gap of 30 frames, where our framework, DMRMs, recognizes the anomalous event of "shoplifting" accurately in the first row. In contrast, it failed in the second case. The second row shows the failure case of our recognition task, where DMRMs recognized it as an "abuse" event though it is a "car accident" case.

Table 6. Evaluation of Anomalous Activity Recognition on UCF-Crime dataset.

Algorithm	**C3D [47]**	**TCNN [57]**	**3D ResNet-34 [21]**
Accuracy	23.0	28.4	27.2

Table 7. Analysis of Computational Cost during Inference Phase (in seconds per frame).

Computational Process	**UCF-Crime**	**ShanghaiTech**
Extraction of Motion Flow Maps	0.016	0.03
16 Temporal Segments and Pre-process.	0.0012	0.0016
32 Temporal Segments and Pre-process.	-	0.0019
Feature Extraction from Visual Data	0.011	0.012
Feature Extraction from Motion Maps	0.011	0.012
Detection	0.0004	0.0005
Total	0.0396	0.0561
FPS	≈ 26	≈ 18

5. Conclusions

We introduced new ranking measures for learning the DMIL in the absence of temporal annotations. Moreover, with the help of joint learning of deep motion and appearance features, our framework, DMRMs, competitively learned the context-dependency compared to other deep learning algorithms [1,3,12,13]. We achieved promising performance on challenging datasets. The experimental results demonstrated noticeable improvement in context-learning, mitigating the false alarm rate, and overall accuracy in the anomalous event detection task. However, there are several drawbacks to our method, which should be handled in the future. First, a large-scale and well-annotated dataset for both normal events and abnormal events is necessary to train the proposed framework since it performs in a semi-supervised manner. A large-scale dataset requirement is an inherent problem in almost all existing visual recognition methods using deep learning. Second, our framework is not robust to the noises (occlusion, camera jitters, and illumination variations). Therefore, our future work will focus on training our network in an unsupervised manner using a smaller amount of training data; consequently, this will address the problem of the requirement of a large-scale and well-annotated dataset. Moreover, our future work will also focus on overcoming the shortcomings of noise due to illumination variations, camera jitters, and occlusion. Finally, we note that in the ShanghiTech dataset, some algorithms outperform our proposed model on the given experimental setting. However, performance improvement is still possible by adopting different models (e.g., 3D ResNet-50) and their optimization by retraining (with practically many training samples), fine-grained transfer learning, and knowledge distillation, and this is our ongoing work.

Author Contributions: Conceptualization, S.D. and M.J.; methodology, S.D., J.G. and M.J.; software, S.D.; validation, S.D. and A.B.; formal analysis, S.D., A.B. and J.G.; investigation, J.G. and M.J.; resources, M.J.; data curation, S.D.; writing—original draft preparation, S.D., and M.J.; writing—review and editing, J.G. and M.J.; visualization, S.D.; supervision, M.G.; project administration, M.J.; funding acquisition, J.G and M.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2014-3-00077, Development of Global Multitarget Tracking and Event Prediction Techniques Based on Real-time Large-Scale Video Analysis), Ministry of Culture, Sports and Tourism (MCST) Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development (R2020070004) Program 2021, and the Brain Research Program through the NRF funded by the Ministry of Science, ICT and Future Planning (Grant No. NRF-2019M3C7A1020406).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available in publicly accessible repositories. Publicly available datasets were analyzed in this study. These datasets can be found in the locations: (1) UCF-Crime Dataset: https://www.crcv.ucf.edu/projects/real-world/ (2) ShanghaiTech Dataset: https://svip-lab.github.io/dataset/campus_dataset.html

Acknowledgments: We are grateful to Paulina Martinez for English language editing. Additionally, the authors would like to thank the editors and all the reviewers for their valuable comments on this manuscript.

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

References

1. Wang, L.; Zhou, F.; Li, Z.; Zuo, W.; Tan, H. Abnormal Event Detection in Videos Using Hybrid Spatio-Temporal Autoencoder. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 2276–2280. [CrossRef]
2. Perera, P.; Nallapati, R.; Xiang, B. OCGAN: One-Class Novelty Detection Using GANs with Constrained Latent Representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019; pp. 2893–2901.
3. Chong, Y.S.; Tay, Y.H. Abnormal Event Detection in Videos Using Spatiotemporal Autoencoder. In *Advances in Neural Networks—ISNN 2017*; Springer: Cham, Japan, 2017; pp. 189–196. [CrossRef]
4. Medel, J.; Savakis, A. Anomaly Detection in Video Using Predictive Convolutional Long Short-Term Memory Networks. *arXiv* **2016**, arXiv:1612.00390.

5. Luo, W.; Liu, W.; Gao, S. A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 341–349. [CrossRef]
6. Zhao, B.; Fei-Fei, L.; Xing, E.P. Online Detection of Unusual Events in Videos via Dynamic Sparse Coding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011.
7. Mahadevan, V.; LI, W.X.; Bhalodia, V.; Vasconcelos, N. Anomaly Detection in Crowded Scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 1975–1981.
8. Lu, C.; Shi, J.; Jia, J. Abnormal event detection at 150 FPS in MATLAB. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 2720–2727. [CrossRef]
9. Liu, W.; Luo, W.; Lian, D.; Gao, S. Future Frame Prediction for Anomaly Detection—A New Baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6536–6545.
10. Xu, K.; Sun, T.; Jiang, X. Video Anomaly Detection and Localization Based on an Adaptive Intra-Frame Classification Network. *IEEE Trans. Multimed.* **2020**, *22*, 394–406. [CrossRef]
11. Morais, R.; Le, V.; Tran, T.; Saha, B.; Mansour, M.; Venkatesh, S. Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 11988–11996.
12. Xu, D.; Yan, Y.; Ricci, E.; Sebe, N. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Comput. Vis. Image Underst.* **2017**, *156*, 117–127. [CrossRef]
13. Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A.K.; Davis, L.S. Learning Temporal Regularity in Video Sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 733–742.
14. Sultani, W.; Chen, C.; Shah, M. Real-World Anomaly Detection in Surveillance Videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6479–6488. [CrossRef]
15. Dubey, S.; Boragule, A.; Jeon, M. 3D ResNet with Ranking Loss Function for Abnormal Activity Detection in Videos. In Proceedings of the International Conference on Control, Automation and Information Sciences (ICCAIS), Chengdu, China, 23–26 October 2019.
16. Hinami, R.; Mei, T.; Satoh, S. Joint Detection and Recounting of Abnormal Events by Learning Deep Generic Knowledge. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3639–3647.
17. Luo, W.; Liu, W.; Gao, S. Remembering history with convolutional LSTM for anomaly detection. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 439–444.
18. Zhong, J.X.; Li, N.; Kong, W.; Liu, S.; Li, T.H.; Li, G. Train a plug-and-play action classifier for anomaly detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
19. Babenko, B. *Multiple Instance Learning: Algorithms and Applications*; Tech. Rep.; University of California: San Diego, CA, USA, 2008.
20. Andrews, S.; Tsochantaridis, I.; Hofmann, T. Support Vector Machines for Multiple-Instance Learning. In *Advances in Neural Information Processing Systems 15 (NIPS)*; MIT Press: Cambridge, MA, USA, 2003; pp. 577–584.
21. Hara, K.; Kataoka, H.; Satoh, Y. Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition. In Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCV), Venice, Italy, 22–29 October 2017; pp. 3154–3160. [CrossRef]
22. Hara, K.; Kataoka, H.; Satoh, Y. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6546–6555. [CrossRef]
23. Wang, L.; Xu, Y.; Cheng, J.; Xia, H.; Yin, J.; Wu, J. Human Action Recognition by Learning Spatio-Temporal Features With Deep Neural Networks. *IEEE Access* **2018**, *6*, 17913–17922. [CrossRef]
24. Bird, N.; Atev, S.; Caramelli, N.; Martin, R.; Masoud, O.; Papanikolopoulos, N. Real time, online detection of abandoned objects in public areas. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Orlando, FL, USA, 15–19 May 2006; pp. 3775–3780.
25. Kamijo, S.; Matsushita, Y.; Ikeuchi, K.; Sakauchi, M. Traffic monitoring and accident detection at intersections. In Proceedings of the 199 IEEE/IEEJ/JSAI International Conference on Intelligent Transportation Systems, Tokyo, Japan, 5–8 October 1999; pp. 703–708.
26. Wei, J.; Zhao, J.; Zhao, Y.; Zhao, Z. Unsupervised Anomaly Detection for Traffic Surveillance Based on Background Modeling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 129–1297. [CrossRef]
27. Bajestani, M.F.; Abadi, S.S.H.R.; Fard, S.M.D.; Khodadadeh, R. AAD: Adaptive Anomaly Detection through traffic surveillance videos. *arXiv* **2018**, arXiv:1808.10044.
28. Mohammadi, S.; Perina, A.; Kiani, H.; Murino, V. Angry Crowds: Detecting Violent Events in Videos. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016.
29. Cong, Y.; Yuan, J.; Liu, J. Sparse reconstruction cost for abnormal event detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 3449–3456.

30. Wu, S.; Moore, B.E.; Shah, M. Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 2054–2060.
31. Tung, F.; Zelek, J.S.; Clausi, D.A. Goal-based trajectory analysis for unusual behaviour detection in intelligent surveillance. *Image Vis. Comput.* **2011**, *29*, 230–240. [CrossRef]
32. Dalal, N.; Triggs, B.; Schmid, C. Human Detection Using Oriented Histograms of Flow and Appearance. In Proceedings of the European Conference on Computer Vision (ECCV), Graz, Austria, 7–13 May 2006; pp. 428–441.
33. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893.
34. Zhang, D.; Gatica-perez, D.; Bengio, S.; Mccowan, I. Semi-supervised adapted hmms for unusual event detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; pp. 611–618.
35. Kim, J.; Grauman, K. Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 2921–2928.
36. Yu, J.; Kim, D.Y.; Yoon, Y.; Jeon, M. Action Matching Network: Open-set Action Recognition using Spatio-Temporal Representation Matching. *Vis. Comput.* **2019**, *36*, 1457–1471. [CrossRef]
37. Yoon, Y.; Yu, J.; Jeon, M. Spatio-Temporal Representation Matching-Based Open-Set Action Recognition by Joint Learning of Motion and Appearance. *IEEE Access* **2019**, *7*, 165997–166010. [CrossRef]
38. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
39. Sun, S.; Akhtar, N.; Song, H.; Mian, A.; Shah, M. Deep Affinity Network for Multiple Object Tracking. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2018**, *43*, 104–119. [CrossRef] [PubMed]
40. Ciaparrone, G.; Sańchez, F.L.; Tabik, S.; Troiano, L.; Tagliaferri, R.; Herrera, F. Deep learning in video multi-object tracking: A survey. *Neurocomputing* **2020**, *381*, 61–88. [CrossRef]
41. Wu, D.; Zheng, S.J.; Zhang, X.P.; Yuan, C.A.; Cheng, F.; Zhao, Y.; Lin, Y.J.; Zhao, Z.Q.; Jiang, Y.L.; Huang, D.S. Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing* **2019**, *337*, 354–371. [CrossRef]
42. Chen, Y.; Zhu, X.; Gong, S. Person Re-identification by Deep Learning Multi-scale Representations. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 2590–2600.
43. Yu, J.; Yow, K.C.; Jeon, M. Joint Representation Learning of Appearance and Motion for Abnormal Event Detection. *Mach. Vis. Appl.* **2018**, *29*, 1157–1170. [CrossRef]
44. Sohrab, F.; Raitoharju, J.; Gabbouj, M.; Iosifidis, A. Subspace Support Vector Data Description. In Proceedings of the IEEE International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018.
45. Abati, D.; Porrello, A.; Calderara, S.; Cucchiara, R. Latent Space Autoregression for Novelty Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
46. Luo, W.; Liu, W.; Lian, D.; Tang, J.; Duan, L.; Peng, X.; Gao, S. Video Anomaly Detection With Sparse Coding Inspired Deep Neural Networks. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2019**. [CrossRef] [PubMed]
47. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
48. Lan, Z.Z.; Bao, L.; Yu, S.I.; Liu, W.; Hauptmann, A.G. Double Fusion for Multimedia Event Detection. In *Advances in Multimedia Modeling*; Springer: Berlin/Heidelberg, Germany, 2012.
49. Farnebäck, G. Two-Frame Motion Estimation Based on Polynomial Expansion. In *Image Analysis*; Springer: Berlin/Heidelberg, Germany, 2003.
50. Lucas, B.D.; Kanade, T. An Iterative Image Registration Technique with an Application to Stereo Vision (DARPA). In Proceedings of the 1981 DARPA Image Understanding Workshop, Washington, DC, USA, 23 April 1981; pp. 121–130.
51. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The Kinetics Human Action Video Dataset. *arXiv* **2017**, arXiv:705.06950.
52. Nazaré, T.S.; de Mello, R.F.; Ponti, M.A. Are pre-trained CNNs good feature extractors for anomaly detection in surveillance videos? *arXiv* **2018**, arXiv:1811.08495.
53. Ding, C.; Tao, D. Robust Face Recognition via Multimodal Deep Face Representation. *IEEE Trans. Multimed.* **2015**, *17*, 2049–2058. [CrossRef]
54. LI, W.X.; Mahadevan, V.; Vasconcelos, N. Anomaly Detection and Localization in Crowded Scenes. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2013**, *36*, 18–32. [CrossRef]
55. Adam, A.; Rivlin, E.; Shimshoni, I.; Reinitz, D. Robust Real-Time Unusual Event Detection using Multiple Fixed-Location Monitors. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2008**, *30*, 555–560. [CrossRef] [PubMed]
56. Wang, J.; Cherian, A. GODS: Generalized One-Class Discriminative Subspaces for Anomaly Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019.
57. Hou, R.; Chen, C.; Shah, M. Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5823–5832.