DyAnNet: A Scene Dynamicity Guided Self-Trained Video Anomaly Detection Network

Kamalakar Vijay Thakare[1], Yash Raghuwanshi[1], Debi Prosad Dogra[1], Heeseung Choi[2,3], and Ig-Jae Kim[2,3]

[1]Indian Institute of Technology, Bhubaneswar, Odisha, 752050, India
[2]Artificial Intelligence and Robotics Institute, Korea Institute of Science and Technology, Seoul 02792, Republic of Korea
[3]Yonsei-KIST Convergence Research Institute, Yonsei University, Seoul 03722, Republic of Korea
{*tkv15, yr15, dpdogra*}*@iitbbs.ac.in,* {*hschoi, drjay*}*@kist.re.kr*

Abstract

Unsupervised approaches for video anomaly detection may not perform as good as supervised approaches. However, learning unknown types of anomalies using an unsupervised approach is more practical than a supervised approach as annotation is an extra burden. In this paper, we use isolation tree-based unsupervised clustering to partition the deep feature space of the video segments. The RGB- stream generates a pseudo anomaly score and the flow stream generates a pseudo dynamicity score of a video segment. These scores are then fused using a majority voting scheme to generate preliminary bags of positive and negative segments. However, these bags may not be accurate as the scores are generated only using the current segment which does not represent the global behavior of a typical anomalous event. We then use a refinement strategy based on a cross-branch feed-forward network designed using a popular I3D network to refine both scores. The bags are then refined through a segment re-mapping strategy. The intuition of adding the dynamicity score of a segment with the anomaly score is to enhance the quality of the evidence. The method has been evaluated on three popular video anomaly datasets, i.e., UCF-Crime, CCTV-Fights, and UBI-Fights. Experimental results reveal that the proposed framework achieves competitive accuracy as compared to the state-of-the-art video anomaly detection methods.

1. Introduction

Video Anomaly Detection (VAD) imposes a critical requirement in visual surveillance. Generally, video anomaly detection task covers a large spectrum including road traffic monitoring [33, 37], violence detection [21, 24, 31], human behaviour [14, 23, 25], crowd monitoring [3, 43], etc. Visual surveillance is primarily done by public and private agencies on a large scale. Hence researchers easily get humongous data analytic task while analyzing and annotating large video data. Moreover, recent existing video anomaly detection methods [9, 23, 25, 33, 38, 48, 49] heavily depend on full or weak supervision. However, generating annotations for such huge datasets is labor-intensive and time-consuming.

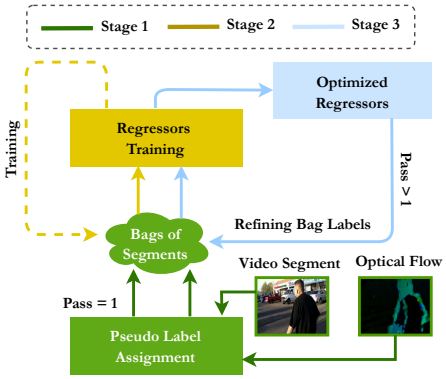

Figure 1. **Overview**. In the first stage, we obtain low-confidence pseudo labels. In the second stage, we incorporate *iterative learning* to train regressor networks using these labels. After successful training, we replace older labels with more confident labels in the third stage and retrain the regressors. After a few passes, an optimized version of regressors is used to predict the anomaly score.

In recent years, unsupervised approaches for video anomaly detection are being outnumbered by supervised or semi-supervised methods. Ravanbaksh *et al*. [36] have trained Generative Adversarial Nets (GANs) for video anomaly detection. Nguyen *et al*. [27] have concatenated appearance and motion encoders and decoders for accomplishing the job. Gong *et al*. [10] have proposed Memory-augmented Autoencoders (MemAEs) to detect video anomalies. The main advantage of using GANs or AEs is their capability to capture high-level video features. Recently, Doshi *et al*. [8] have proposed a continual learning frame-

work in which the model incrementally trains as the data arrives without forgetting the learnt (past) information. This type of framework can be feasible in visual surveillance as video data keep coming into the monitoring systems. However, all these approaches suffer a few limitations as follows: (1) in continual learning, a separate mechanism needs to be designed to avoid catastrophic forgetting [8], (2) GANs and AEs are highly vulnerable to unstable training, i.e., a subtle change in data imposes large changes in the labels, thus affecting the normal distribution, (3) most of the state-of-art VAD methods heavily depend on labeled normal/abnormal data, and (4) VAD approaches either utilize appearance-based features or deep features.

To address these limitations, we adopt an iterative learning [44] mechanism in which models are repeatedly tuned with more refined data during each pass. Moreover, we aim to combine the technical advantages of continual and AEs learning. Our proposed framework combines the power of DNNs with well-justified handcrafted motion features. These spatio-temporal features equipped with low-level motion features help to detect wide range of anomalies. The framework can also be retrained in an end-to-end fashion as input data arrives. The overview of the proposed framework is depicted in Fig. 1. It is divided into three stages: i) pseudo label assignment, ii) regressors training, and iii) refinement of labels using optimized regressors. For enabling the regressors to understand subtle anomalies, we have obtained motion features, namely dynamicity score using optical flow. In the first stage, we do not know the actual labels; hence we have obtained intermediate low confidence anomaly labels using OneClassSVM and iForest [19]. We also obtain the dynamicity labels using dynamicity scores. We have trained two regressor networks in the second stage by using the labels generated in the first stage. This is an iterative process to improve the confidence scores. In this way, both regressors are trained over refined labels and they learn discriminating features. The iterative learning approach also ensures that both the regressors learn new distinguish patterns without losing the past information. We have experimentally found that for first few iterations, both regressors gradually learn internal patterns and stabilizes after some iterations. Both regressors are trained independently in parallel. Precisely, in iterative learning, the model is retrained using refined data in each iteration. In this way, the proposed approach do not need any level of supervision. However, some form of supervision is mandatory for continual learning [8] or weakly-supervised methods [27, 38, 48]. These methods consider a video anomalous even if a small segment contains anomaly. In contrast, we identify anomalous segments using dynamicity and anomaly scores estimated using unsupervised ways, thus eliminating the requirement of supervision. To achieve this, we have made the following contributions:

- design an unsupervised end-to-end video anomaly detection framework that uses iterative learning to tune the model using refined labels in each iteration;
- propose a novel technique to assign intermediate labels in unsupervised scenarios by combining deep features with well-justified motion features and;
- conduct extensive experiments to understand the effectiveness of the proposed framework with respect to other state-of-the-art methods.

The rest of the paper is organized as follows. In the next section, we present the related work. In Sec. 3, we present the proposed framework. Experiments and results are presented in Sec. 4. The conclusions and future works are presented in Sec. 5.

2. Related Work

Existing work in the Video Anomaly Detection (VAD) domain largely draw motivation from activity recognition and scene understanding [38]. These methods utilize various types of video features, training procedures or both. In this section, we briefly discuss the main categories that are extensively followed in very recent VAD approaches.

2.1. Reconstruction-based Approaches

Several VAD approaches [1, 10, 22, 27, 29, 30, 39, 46] employ Autoencoders (AEs), Generative Adversarial Nets (GANs) and their variants under the assumption that the models that are explicitly trained on normal data may not be successful to reconstruct abnormal event as such samples are usually absent in the training set. Park *et al*. [29] have used AE to generate cuboids within normal frames using spatial and temporal transformation. Zaheer *et al*. [46] have generated good quality reconstructions using the current generator and used the previous state generator to obtain bad quality examples. This way, the new discriminator learns to detect even small distortions in abnormal input. Gong *et al*. [10] have introduced a memory module to AE and constructed MemAE. This is an improved version of existing AE. Szymanowicz *et al*. [39] have trained an AE to obtain saliency maps using five consecutive frames and per-pixel prediction error. Ravanbakhsh *et al*. [36] have imposed classic adversarial training using GANs to detect anomalous activity. However, the effectiveness of these approaches is highly dependent on the reconstruction capabilities of the model. Failing which, it may significantly degrade the model's performance.

2.2. Features-based Approaches

Primarily, features-based VAD approaches can be categorized by anomaly detection using either handcrafted or

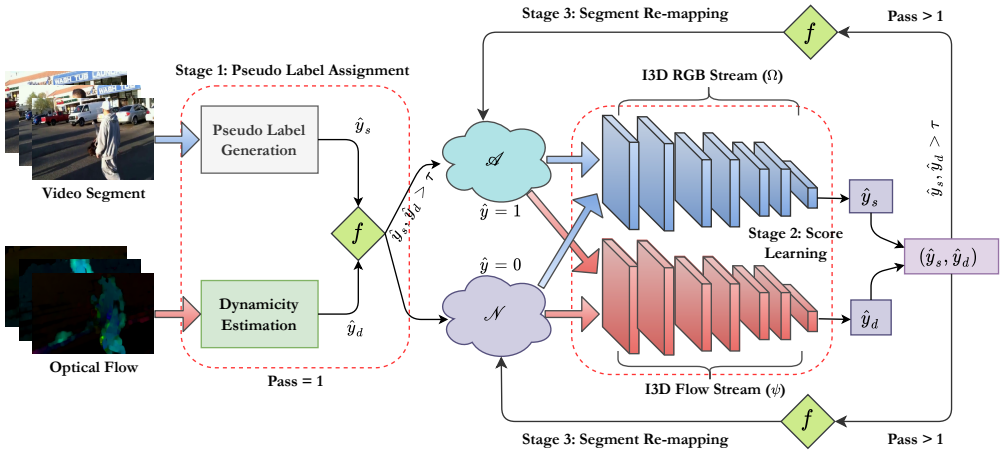

Figure 2. **DyAnNet**. Architecture of the proposed framework. The whole framework is divided into three stages: (1) Pseudo label assignment, (2) Score learning, and (3) segment re-mapping using refined labels. We have employed iterative learning mechanism to train the regressors Ω and ψ, and redefined the input bag $\mathcal{A}$ and $\mathcal{N}$ at the end of each pass. We construct a set of optimized regressors obtained through each pass and used it to predict anomaly and dynamicity score of each segment. (Thinner arrows represent the label passing and the thicker arrows are video features, i.e., the blue arrows are raw RGB frames, whereas the red arrows are representing optical flow of the segment.)

deep features. Early attempts have used handcrafted features such as object trajectories [26, 47], gradients of histograms (HOGs) [18], Hidden Markov Model (HMM) [16], and appearance-based features [11]. However, very recent deep learning approaches [9, 38, 48, 49] have achieved robust results for video anomaly detection. Feng *et al*. [9] have introduced Self-guided attention during the feature encoding process, Zhu *et al*. [49] have injected motion-aware features that increases the recognition capabilities of the classifier. Sultani *et al*. [38] addresses anomaly detection problem using weak supervision and following this, [48] has used Graph Convolutions Network (GCN). In addition to this, different training mechanisms have been employed such as continual learning [8], adversarial training [36], Self-trained [9, 28], and active learning [41] to obtained robust video anomaly detection results.

Even though the aforementioned techniques have achieved decent performance, they still suffer from a few avoidable limitations: (1) they heavily depend on manually labelled normal/abnormal data. However, generating annotations for huge data is time consuming and error prone, (2) due to the absence of universal definition the anomaly events, a few anomalous events that are normal in one context may be regarded abnormal in another context, e.g. marathon run vs. criminal run. These scenarios often lead to unstable training of the AEs and GANs. We have addressed these limitations using iterative learning combined with low and high-level features.

3. Proposed Method

We first provides a detailed description of the proposed video anomaly detection framework. Our framework encompasses with the following three stages: (1) pseudo label assignment, (2) anomaly score learning, and (3) segment re-mapping.

3.1. Overall Architecture

A high-level architecture of the proposed framework is depicted in Fig. 2. The problem formulation is follows: Assume an input video (V) is divided into a n number of segments such that $V = \{S_1, S_2, \ldots, S_n\}$. The goal is to design a function as given in Eq. 1 that generates an anomaly score y_s and a dynamicity score y_d to predict the label $y \in \{0, 1\}$ for each video segment.

$$\Theta : V \rightarrow y \in \{0, 1\} \tag{1}$$

A positive segment contains anomalous activity and ideally has a higher anomaly and dynamicity score than the normal segments such that $\Theta(S_i) > \Theta(S_j)$, where S_i is an anomalous and S_j is a normal segment. Note that, no labeled data is available during this training. To tackle this scenario, we have employed iterative learning [44] and bag formation [38]. First, we have assigned pseudo anomaly scores $\hat{y}_s$ and pseudo dynamicity scores $\hat{y}_d$ to the video segments S_i. These intermediate labels help to form two separate bags $\mathcal{A}$ and $\mathcal{N}$. Here, $\mathcal{A} \subset V$ is bag of positive segments, where $S \in \mathcal{A}$ if $y = 1$ for S, which generally has higher $\hat{y}_s$ and $\hat{y}_d$ values. Similarly, $\mathcal{N} \subset V$ is the bag of negative segments, where $S \in \mathcal{N}$ if $y = 0$ for normal segment S and we expect a lower value for both $\hat{y}_s$ and $\hat{y}_d$. Note that, $\mathcal{A} \cap \mathcal{N} = \phi$. In the second stage, two separate regressors, e.g. Ω and ψ have been trained using these pseudo labels. In the third stage, we have used these trained regressors to refine the contents of the bags. A training pass redefines the membership of each

segment of a bag. In the next pass, Ω and ψ are tuned using $\mathcal{A}$ and $\mathcal{N}$. In the subsequent sections, we provide detailed descriptions of the stages.

3.2. Pseudo Label Assignment

The training procedure begins with unlabeled data. Hence we don't know $\mathcal{A}$ and $\mathcal{N}$ in the first place. To handle this problem, we initialize $\mathcal{A}$ and $\mathcal{N}$ via generating pseudo anomaly score $\hat{y}_s$ and pseudo dynamicity score $\hat{y}_d$. To obtain $\hat{y}_s$, we have employed OneClassSVM and iForest [19] in combination. Note that, both algorithms run on feature vectors of the video segment. We have extracted segment features using I3D pretrained on Kinetic dataset [5]. OneClassSVM is similar to the SVM algorithm. But, it uses hypersphere to cover all data instances. This algorithm tries to construct the smallest possible hypersphere using the support vectors. All the data instances that lie outside the hypersphere are likely to be anomalies. Let $F = f(S)$ be the feature extraction function of segment S. The anomaly score can be defined using Eq. 2,

$$d(F) = \max_{F \in V} \delta(c, F) \tag{2}$$

where F is the feature point, c is the center of the smallest hypersphere constructed by the SVM, and δ is the distance function. iForest isolates data instances by randomly selecting any feature and a split value. A tree structure can depict this recursive partitioning; hence the number of partitions is equal to the path length of the data instance up to the root node. The inverse of the path from the root to leaf is the anomaly score of F. It is estimated using Eq. 3,

$$d(F) = 2^{[\frac{-E(l(F))}{g(|F|)}]} \tag{3}$$

where $l(F)$ is the path length of F, $E(.)$ denotes the average path length of F on n isolation trees, and $g(.)$ is the expected path length for a given sub-sample. We normalize the anomaly scores of each feature point within [0,1] interval and take the average score over n isolation trees to obtain the pseudo anomaly score $\hat{y_s}$ of a video segment.

In addition to the anomaly score, we have also obtained the dynamicity score of each segment. The dynamicity of the segment refers to the rate of change in displacement of the pixel over time which is obtained using motion information. It is expected that for a rapidly changing video scene, the dynamicity score is expected to be higher. Let P_k represents the coordinate of the k^{th} pixel of the immediate preceding frame and M_k be the estimated position of the pixel obtained using optical flow in the next frame. The displacement of the pixel (S_k) can be calculated using Eq. 4,

$$S_k = SAD(P_k, M_k) \tag{4}$$

where SAD is the sum of absolute difference. We have used absolute displacement to consider movement in any direction to estimate the dynamicity score. Now, the frame-level dynamicity score D_i of the i^{th} frame is estimated using Eq. 5,

$$D_i = \frac{1}{m \times n} \sum_{k=1}^{m \times n} S_k \tag{5}$$

where m and n represent height and width of the frame. We then obtain the dynamicity scores of all the frames within a segment. It is represented by $[D_i, D_{i+1}, \ldots, D_{p-1}]$ assuming there are p number of frames in a segment. We average all frame-level dynamicity scores to obtain a pseudo dynamicity score $\hat{y}_d$ of the segment. The score is then normalized within [0,1]. We now assign an intermediate label $\hat{y}$ to a segment using the heuristic presented in Eq. 6.

$$\hat{y} = f(\hat{y}_s, \hat{y}_d) = \begin{cases} 1 & \text{if } \hat{y}_s, \hat{y}_d > \tau \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Eq. 6 ensures that the segment with a higher anomaly and dynamicity scores than a predefined threshold τ should be placed in $\mathcal{A}$ with the intermediate label $\hat{y} = 1$.

3.3. Learning of Anomaly and Dynamicity Scores

Ideally, when a score learner feeds with an anomalous segment, it should generate a high anomaly score as compared to a normal segment. However, in the present scenario, the labels are inaccurate due to the absence of ground truths. Moreover, the label of each segment has been decided using anomaly and dynamicity scores. Hence we have carefully designed a function Θ as given in Eq. 7 using two different score learner functions, namely Ω and ψ,

$$\Theta(S) = f(\Omega(Z_R), \psi(Z_F)), \tag{7}$$

where Z_R represents the RGB frame, Z_F denotes the optical flow of the segment S, $\hat{y}_s = (\Omega(Z_R))$, $\hat{y}_d = (\psi(Z_F))$, either $S \in \mathcal{A}$ or $S \in \mathcal{N}$ and $f(.)$ is the label mapping function defined in Eq. 6. Typical 3D CNNs can be incorporated here to implement Ω and ψ. We have employed RGB and flow modalities of I3D [5] network followed by a 3-layer FCN to implement score learners Ω and ψ, respectively. We train the anomaly score learner $\Omega(Z_R, \mathbf{W_\Omega})$ and dynamicity score learner $\psi(Z_R, \mathbf{W}_\psi)$ using Mean-squared error (MSE) loss, where $\mathbf{W}_\Omega$ and $\mathbf{W}_\psi$ are trainable weights of Ω and ψ, respectively.

3.4. Segment Re-mapping via Iterative Learning

The training procedure begins with the pseudo labels assigned to the segments in the first stage. However, labels are not as correct as the ground truth. In this stage, we aim to fine-tune Ω and ψ with more accurate labels to achieve stable performance. To achieve this, we have incorporated an iterative learning mechanism. Let P_i be the i^{th} pass in which $\mathcal{A}$ and $\mathcal{N}$ have been initialized based on pseudo

anomaly and dynamicity scores. We then train Ω and ψ via MSE loss using these pseudo labels. We have obtained sub-optimized version Ω_{P_i} and ψ_{P_i} of both regressors. Lastly, we re-estimate both scores using Eq. 8 via these optimized versions of the regressors,

$$\hat{y}_s^{P_{i+1}} = \Omega_{P_i}(Z_R) \text{ and } \hat{y}_d^{P_{i+1}} = \psi_{P_i}(Z_F) \quad (8)$$

where $\hat{y}_s^{P_{i+1}}$ and $\hat{y}_d^{P_{i+1}}$ are new scores obtained through sub-optimized regressors. Now, we use these new scores to refine $\mathcal{A}$ and $\mathcal{N}$ using Eq. 6 and retrained Ω_{P_i} and ψ_{P_i} in the next pass P_{i+1} using a new input batch. In particular, for each pass in iterative learning, we utilize a completely new set of $\mathcal{A}$ and $\mathcal{N}$ and retrain the regressors. We only utilize new scores instead of combining them with older scores because such a mixing without any supervision usually generates erroneous scores. We have empirically found that the proposed approach performs better on popular video anomaly datasets. Finally, each pass generates an optimized version of Ω and ψ and hence the proposed iterative learning approach results in a set of optimized regressor models.

3.5. Training and Inference

We have employed iterative learning to achieve stable performance of the regressors. During the first pass, we have obtained pseudo anomaly and dynamicity scores to initialize $\mathcal{A}$ and $\mathcal{N}$. However, the actual training takes place in the second stage, where two regressor models Ω and ψ are trained using the pseudo labels. Note that, Ω and ψ are I3D [5] networks followed by a 3-layer FCN with a single neuron at the end to produce respective scores. Hence we have incorporated MSE loss for network the training as the formulation is recognized as regression rather than a binary classification. In each pass, both regressor networks have been trained using a fixed number of training iterations depending on the number of samples available in the training set. Finally, the sub-optimized version of Ω and ψ are used to rearrange the content of $\mathcal{A}$ and $\mathcal{N}$ for the next pass.

Each pass in iterative learning outputs an optimized version of $\mathcal{A}$ and $\mathcal{N}$. In the inference stage, we use a set of sub-optimized models to generate optimized anomaly and dynamicity scores. The final score generation can be summarized using Eq. 9,

$$y_s = \sum_{i=1}^{k} \Omega_i(Z_R) \text{ and } y_d = \sum_{i=1}^{k} \psi_i(Z_F) \quad (9)$$

where k is the number of passes. Ω_i and ψ_i represent the optimized models obtained after the i^{th} pass. y_s and y_d are anomaly and dynamicity scores obtained using Ω_i and ψ_i, respectively. The output neuron from both the regressors use softmax, hence anomaly and dynamicity scores always fall between [0,1] for an input video segment.

4. Experiments

In this section, we present implementation details, datasets, evaluation metrics, comparisons of the proposed method with recent state-of-the-art VAD methods, qualitative results, ablation experiments, and the effect of training and testing iterations on performance.

4.1. Implementation Details

Following [38, 49, 48], we divide each video into 32 non-overlapping temporal segments. We then extract features from *mixed-5C* layer of the I3D [5] network resulting in 1024D feature components, and feed them to PCA to reduce dimensionality to 100 components. These components have been used to train the OneClassSVM and iForest [19] classifiers to generate pseudo anomaly scores. We have used default parameters of OneClassSVM and iForest [19] given in the scikit-learn during experiments.

We have used SelFlow [20] and Farneback algorithm for optical flow estimation to calculate the dynamicity score of the segment. We have implemented the regressors Ω and ψ using I3D [5] as a backbone network pre-trained on the Kinetic dataset as recommended in I3D original work. We have replaced FCN layers of I3D with a 3-layer FCN. The first layer contains 512 units, followed by 32 units, and 1 unit at the end to generate the scores. We have also experimented with deeper networks. However, we have not observed significant performance deviation. We have trained the regressors with initial learning rate of 0.005 and AdaGrad optimizer. Following [12, 38, 48, 49], we set $\tau = 0.50$ for comparisons. We have experimented with even lower values of (τ). Such analysis can be found in supplementary document. The experiments reveal that both regressors get substantially improved only in the first few passes while achieving a stable performance. We have discussed results by varying the number of training iterations and passes in the subsequent sections.

4.2. Datasets

We have used three real-world video anomaly datasets for experiments, namely UCF-Crime [38], CCTV-Fights [31], and UBI-Fights [6].

UCF-Crime [38]: It is a video anomaly dataset containing 13 real-world anomalies recorded using CCTV cameras. It contains 1900 real-world videos of normal and criminal activities such as robbery, vandalism, burglary, shooting, abuse, etc.

CCTV-Fights [31]: This dataset offers 1000 fighting videos recorded in real-world scenarios. The total duration of these videos is 17.68 hours and collected using search keywords like kicking, punching, physical violence, mugging, etc.

UBI-Fights [6]: It holds 1000 real-world videos, where 784 are normal and 216 are real-life fighting scenarios. It

contains videos recorded in indoor and outdoor environment with no administrative control or supervision, high occlusion, and varying illumination conditions.

4.3. Performance Evaluation Metrics

All test video frames from the above-aforementioned datasets are marked as either normal or abnormal. Hence following the previous works [9, 27, 28, 29, 31, 38, 39, 45, 48, 49] on anomaly detection, we compute frame-level receiver operating characteristics (ROC) curve and area under the curve (AUC) as evaluation metrics.

4.4. Comparisons with State-of-the-art

We compare our method with recent state-of-the-art video anomaly detection methods [7, 8, 9, 11, 12, 13, 15, 17, 27, 28, 29, 34, 35, 38, 49] on three aforementioned datasets. Tab. 1 shows the performance of all methods. It can be observed that the proposed unsupervised method outperforms other weakly-supervised methods [9, 38, 49] by a substantial margin across all three datasets. Zhu [49], Pang *et al*. [28], and Leroux *et al*. [17] have achieved decent performance on all datasets by introducing attention-based deep features, ordinal regression, and multi-branch deep autoencoders, respectively. However, incorporating multiple deep networks and adding attention-based features into the network are insufficient to detect multiple anomalous events. It can be observed that the multi-branch framework introduced by Leroux *et al*. [17] performs well on CCTV-Fights [31] and UBI-Fights [6] as these datasets focus on fighting events only. However, it performs moderately on UCF-Crime [38] as the dataset addresses multiple anomalous activities. Doshi *et al*. [8] have employed continual learning in which the model learns new patterns as the input data arrives without forgetting the learnt information. However, such type of learning requires constant flow of incoming data. Moreover, such continual learning approach can efficiently utilize the temporal information of single fixed location [8, 34]. However, the chosen VAD datasets [6, 31, 38] for the experiments are multi-scene and provide complex temporal richness. To tackle this problem, Doshi *et al*. [8] have constructed NOLA video anomaly dataset using fixed location camera. However, to the best of our knowledge, this dataset is yet to be published. Perez *et al*. [31] have introduced CCTV-Fights dataset and computed the performance of C3D [40], I3D [5], and other popular backbone architectures. However, the popular 3D-CNN-based backbone architecture such as C3D [40] and I3D [5] have already been incorporated in the proposed framework as well as with other methods [9, 38]. Hence we have not explicitly included the method used in [31] for comparisons. However, we have studied the effectiveness of the backbone architectures in the proposed framework. Tab. 2 shows AUC (in %) for four popular architectures, namely Pseudo-ResNet 3D [32], Temporal Segments Network [42], C3D [40], and Inception V3 [5].

Table 1. Frame-level AUC scores (in %) of the state-of-the-art methods on three video anomaly datasets, D1: CCTV-Fights [31], D2: UBI-Fights [6], and D3: UCF-Crime [38]. The top two results are shown in red and blue.

Year	Method	D1	D2	D3	Superv.
2016	Hasan *et al*. [11]	52.43	64.87	50.6	Semi.
2017	Hinami *et al*. [12]	56.70	67.12	57.10	Semi.
2018	Ravanbaksh *et al*. [35]	60.37	69.45	61.61	Unsuper.
2018	Sultani *et al*. [38]	72.55	78.70	75.41	Weak.
2019	Ionescu *et al*. [13]	73.86	78.49	76.20	Unsuper.
2019	Nguyen *et al*. [27]	76.43	77.18	75.65	Semi.
2019	Zhu *et al*. [49]	75.20	81.02	79.0	Weak.
2020	Degardin *et al*. [6]	77.14	84.60	76.90	Weak.
2020	Ramachandra *et al*. [34]	73.81	82.45	75.46	Semi.
2020	Pang *et al*. [28]	76.78	84.65	78.50	Unsuper.
2021	Feng *et al*. [9]	**81.43**	**85.19**	**82.30**	Weak.
2021	Kopuklu *et al*. [15]	74.90	79.63	75.12	Weak.
2022	Doshi *et al*. [8]	75.86	80.71	79.46	Semi.
2022	Park *et al*. [29]	73.28	77.23	75.40	Unsuper.
2022	Leroux *et al*. [17]	76.20	78.06	76.78	Unsuper.
	Ours (Farneback Flow)	79.31	84.12	81.40	Unsuper.
	Ours (SelFlow [20])	**81.01**	**86.31**	**84.50**	Unsuper.

Following the limitations imposed by Zhong *et al*. [48], Pang *et al*. [28] have formulated anomaly detection as unsupervised ordinal regression and performed image-level anomaly detection. However, focusing only on spatial-features and ignoring temporal aspect of the video is not advisable in the context of anomalous event. Our framework utilizes both spatial and temporal information and hence outperforms the method proposed by Pang *et al*. [28] by a notable margin.

Table 2. Performance of the proposed method in terms of AUC (%) with different backbone architectures used for implementation of Ω and ψ regressors.

Backbone	CCTV-Fight [31]	UBI-Fight [6]	UCF-Crime [38]
P3D [32]	**78.42**	**84.20**	**84.78**
TSN [42]	77.10	83.08	81.22
C3D [40]	76.56	81.91	79.96
I3D [5]	**81.01**	**86.31**	**84.50**

Based on the performance results discussed so far, it is important to note the proposed framework i) addresses the feature selection problem faced by [9, 29, 49] using low-level motion features and spatio-temporal features, ii) employs an iterative training rather than depending on the weak labels [38, 48, 49]. Thus, our method has achieved a reasonable gain in terms of AUC (%) score.

4.5. Qualitative Analysis

We present a few qualitative results obtained using the proposed method on a few test videos taken from the CCTV-Fight [31], UBI-Fights [6], and UCF-Crime [38] datasets. Such results are presented in Figs. 3, 4, and 5, respectively. Note that, the trained regressors Ω and ψ generate corresponding segment-level anomaly and dynamicity scores.

Hence we have interpolated these scores using Cubic Interpolation to achieve smooth curves. It can be seen that the method successfully detects anomalous segments and generates higher anomaly and dynamicity scores as per the ground truths.

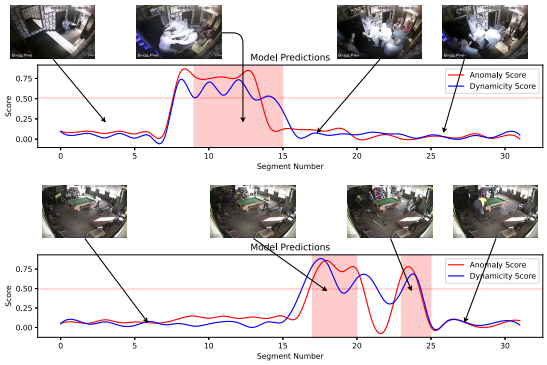

Figure 3. **Results Visualization**: Qualitative results on test videos taken from the CCTV-Fight [31] dataset. Each image represents a frame in a temporal segment. The shaded portions are ground truths and the horizontal line represents the threshold.

From Figs. 3 and 5, it can be seen that both regressors accurately detect anomalous patterns and abrupt change in the scene a few frames earlier. This indicates a quick response to sensitive contents. Moreover, the proposed framework is able to detect multiple occurrences of anomalous events in a video. From Figs. 3 and 4, it can be seen that in the absence of any anomalous activity, both regressors generate very low scores yielding lower false alarms toward the later part of the videos.

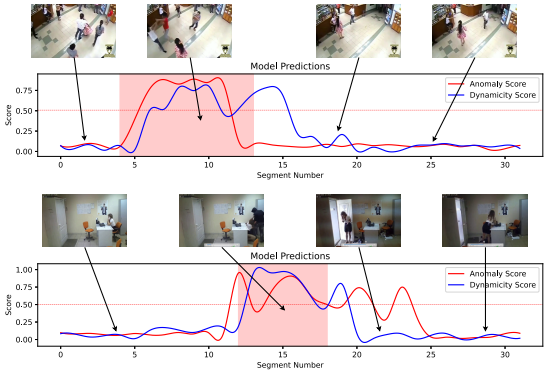

Figure 4. **Results Visualization**: Qualitative results on test videos taken from the UBI-Fight [6] dataset.

In Fig. 5, the first illustration depicting an explosion event from the UCF-Crime [38] dataset is very interesting. The explosion usually fills the whole field of view of the camera with a thick smoke that moves slowly or rapidly depending on the intensity of the explosion event. In this example, after successfully detecting the first explosion, due to faster

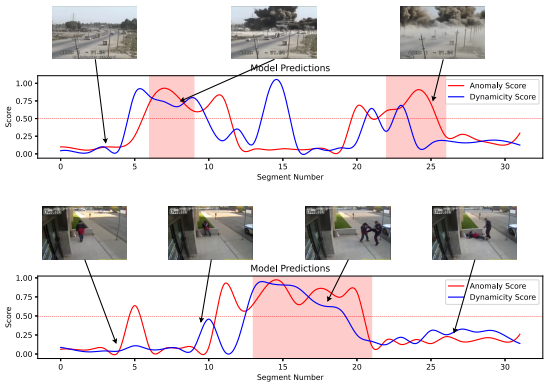

Figure 5. **Results Visualization**: Qualitative results on the test videos from the UCF-Crime [38] dataset.

moving smoke, the ψ regressor has generated a high dynamicity score. However, detecting smoke is not necessarily an anomalous event. Hence Ω has predicted a very low anomaly score for the same segment avoiding false positives. However, during the second mild-level explosion, both regressors agree to generate a relatively higher score. More qualitative results on anomaly detection have been provided in the supplementary material.

4.6. Number of Passes and Training Iterations

To understand the iterative training mechanism, we present the AUC (in %) results of the proposed framework at each pass during the training on CCTV-Fight [31], UBI-Fights [6], and UCF-Crime [38] in Figs. 6, 7, and 8. For CCTV-Fights [31], our method achieves stable performance at the 9th and 10th pass. However, for UBI-Fights [6] and UCF-Crime [38], the framework's performance has improved significantly in the first few passes across all datasets. It achieves a stable performance after the 7th or 8th pass. Note that, during each pass, the sub-optimized version of Ω and ψ is retrained with the refined pseudo labels. Hence it is necessary to restricts this training to avoid over-fitting. We have observed that the number of training iterations can be decided by the input batch size and the total number of samples in the training set. For example, CCTV-Fight [31], UBI-Fights [6], and UCF-Crime [38] datasets contain a few thousands of video samples in the training set. Hence 30 training iterations/pass with a batch size 32 is sufficient to train the model. However, we have experimentally found that the number of iterations/pass do not matter much as long as large number of iterations are done within a pass. This ensures that the models are retrained iteratively. Hence we can achieve same performance with fewer number of training iterations and a large number of passes and vise-versa. Since all datasets offer a few thousands of samples in the training set, we have found that 10 passes and 30 training iterations are sufficient to train both the regressors. All experiments

in this paper have thus been conducted under this uniform setting.

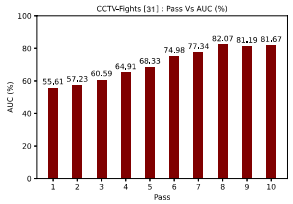

Figure 6. **Pass Vs. AUC**: The AUC (in %) performance of the proposed method for the CCTV-Fights [31] dataset videos against each pass, where x-axis is the number of passes and y-axis presents AUC.

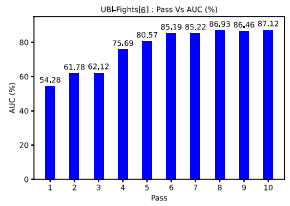

Figure 7. **Pass Vs. AUC**: The AUC (in %) performance of the proposed method for the UBI-Fights [6] dataset videos against each pass.

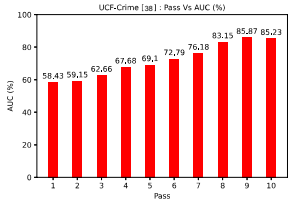

Figure 8. **Pass Vs. AUC**: The AUC (in %) performance of the proposed method for the UCF-Crime [38] dataset videos against each pass.

4.7. Ablation Study

The proposed method has three main modules: i) pseudo label assignment, ii) backbone architecture, and iii) dynamicity score to efficiently detect anomalies. In the first stage, we have employed OneClassSVM and iForest [19] to obtained pseudo anomaly score for each temporal segment. We have replaced these two unsupervised anomaly detection algorithms with Robust Covariance [2] and Local Outlier Factor (LOF) [4]. However, a significant degradation in the AUC performance has been observed (3% - 5%). In the second stage, we have employed two-stream I3D [5] followed by a 3-layer FCN to generate the scores. To check the efficiency of this backbone, we have re-conducted the experiments with same setting. The overall AUC performance with respect to the backbone is represented in Tab. 2. We have also represented the effect of considering low-level motion features to decide the abnormality of the scene. From Tab. 3 and qualitative results, it can be safely concluded that, inclusion of motion features helps to achieve good detection performance as well as lower False Alarms Rate (FAR). We have explored various unsupervised algorithms to generate pseudo anomaly scores. Tab. 4 presents the AUC performance of these experiments. It reveals that when OCSVM is combined with iForest, we get best performance.

Table 3. AUC (in %) and False Alarms Rate (FAR) of the proposed method with and without the dynamicity score. An improved AUC and corresponding FAR are shown in red and blue colors, respectively.

Dynamicity	**CCTV-Fight** [31]	**UBI-Fights** [6]	**UCF-Crime** [38]
No	75.21 (5.8)	81.64 (4.7)	79.76 (1.8)
Yes	**81.01 (1.7)**	**86.31 (1.4)**	**84.50 (0.5)**

Table 4. Performance of the proposed method in terms of AUC (%) with different unsupervised algorithms combined with iForest [19] to generate pseudo-anomaly scores.

Algorithm	**CCTV-Fights** [31]	**UBI-Fights** [6]	**UCF-Crime** [38]
MCD	77.24	84.07	81.11
PCA	**79.94**	**85.13**	**83.58**
LOF	77.60	84.86	82.02
OCSVM	**81.01**	**86.31**	**84.50**

5. Conclusion and Future Work

It has been discussed in this paper that large-scale video anomaly detection using iterative learning is a viable approach to avoid annotation dependency. We have shown that by employing iterative training, the model can learn discriminating features. Moreover, we have shown that by employing pseudo-label generation, one can avoid any type of supervision and still achieve very good performance. Two key insights are: i) low-level features are equally important for anomaly detection, and ii) iterative training helps to reduce FAR and it is possible to detect anomalous event a few frames earlier. We can explore more advance technique to utilize both low-level and deep features in future. However, it is not wise to assume that any AI assisted visual surveillance framework can be a complete replacement of manual surveillance. Essentially, the amount of training data and quality of the underlying model play important role in decision making.

Acknowledgement

This work was supported in part by the Korea Institute of Science and Technology (KIST) Institutional Program under Project 2E31082 and in part by the National Research Foundation (NRF) Project (Grant No. 2018M3E3A1057288) executed at IIT Bhubaneswar with project code CP220.

References

[1] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pages 481–490, 2019.

[2] Fatemah A. Alqallaf, Kjell P. Konis, R. Douglas Martin, and Ruben H. Zamar. Scalable robust covariance and correlation estimates for data mining. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 14–23, 2002.

[3] Shreetam Behera, Thakare Kamalakar Vijay, H Manish Kausik, and Debi Prosad Dogra. 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)*, pages 1–8, 2021.

[4] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. *SIGMOD Rec.*, 29:93–104, 2000.

[5] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4724–4733, 2017.

[6] Bruno Degardin and Hugo Proença. Human activity analysis: Iterative weak/self-supervised learning frameworks for detecting abnormal events. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–7. IEEE, 2020.

[7] Bruno Degardin and Hugo Proença. Iterative weak/self-supervised classification framework for abnormal events detection. *Pattern Recognition Letters*, 145:50–57, 2021.

[8] Keval Doshi and Yasin Yilmaz. Rethinking video anomaly detection - a continual learning approach. In *Wint. Conf. on Appli. of Comp. Vis. (WACV)*, pages 3036–3045, 2022.

[9] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pages 14009–14018, June 2021.

[10] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton Van Den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proc. Int. Conf. Comput. Vis. (ICCV)*, pages 1705–1714, 2019.

[11] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pages 733–742, 2016.

[12] Ryota Hinami, Tao Mei, and Shin'Ichi Satoh. Joint Detection and Recounting of Abnormal Events by Learning Deep Generic Knowledge. In *Proc. Int. Conf. Comput. Vis. (ICCV)*, pages 3639–3647, 2017.

[13] Radu Tudor Ionescu, Sorina Smeureanu, Marius Popescu, and Bogdan Alexe. Detecting abnormal events in video using narrowed normality clusters. In *Wint. Conf. on Appli. of Comp. Vis. (WACV)*, pages 1951–1960, 2019.

[14] Okan Kopuklu, Jiapeng Zheng, Hang Xu, and Gerhard Rigoll. Driver anomaly detection: A dataset and contrastive learning approach. In *Wint. Conf. on Appli. of Comp. Vis. (WACV)*, pages 91–100, January 2021.

[15] Okan Kopuklu, Jiapeng Zheng, Hang Xu, and Gerhard Rigoll. Driver anomaly detection: A dataset and contrastive learning approach. In *Wint. Conf. on Appli. of Comp. Vis. (WACV)*, pages 91–100, January 2021.

[16] Louis Kratz and Ko Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1446–1453, 2009.

[17] Sam Leroux, Bo Li, and Pieter Simoens. Multi-branch neural networks for video anomaly detection in adverse lighting and weather conditions. In *Wint. Conf. on Appli. of Comp. Vis. (WACV)*, pages 2358–2366, January 2022.

[18] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):18–32, 2014.

[19] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Trans. Knowl. Discov. Data*, 6(1), 2012.

[20] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4566–4575, 2019.

[21] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 FPS in MATLAB. In *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2720–2727, 2013.

[22] Yiwei Lu, K Mahesh Kumar, Seyed shahabeddin Nabavi, and Yang Wang. Future frame prediction using convolutional vrnn for anomaly detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2019.

[23] Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. Graph embedded pose clustering for anomaly detection. In *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, June 2020.

[24] Sadegh Mohammadi, Alessandro Perina, Hamed Kiani, and Vittorio Murino. Angry crowds: Detecting violent events in videos. In *Eur. Conf. on Comp. Vis. (ECCV)*, volume 9911, pages 3–18, 2016.

[25] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, June 2019.

[26] Brendan Tran Morris and Mohan Manubhai Trivedi. Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(11):2287–2301, 2011.

[27] Trong Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proc. Int. Conf. Comput. Vis. (ICCV)*, pages 1273–1283, 2019.

[28] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pages 12170–12179, 2020.

[29] Chaewon Park, MyeongAh Cho, Minhyeok Lee, and Sangyoun Lee. Fastano: Fast anomaly detection via spatio-temporal patch transformation. In *Wint. Conf. on Appli. of Comp. Vis. (WACV)*, pages 2249–2259, January 2022.

[30] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pages 14360–14369, 2020.

[31] Mauricio Perez, Alex C. Kot, and Anderson Rocha. Detection of real-world fights in surveillance videos. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2662–2666, 2019.

[32] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *Proc. Int. Conf. Comput. Vis. (ICCV)*, pages 5534–5542, 2017.

[33] Bharathkumar Ramachandra and Michael J. Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Wint. Conf. on Appli. of Comp. Vis. (WACV)*, pages 2569–2578, Feb. 2020.

[34] Bharathkumar Ramachandra, Michael J. Jones, and Ranga Raju Vatsavai. Learning a distance function with a siamese network to localize anomalies in videos. In *Wint. Conf. on Appli. of Comp. Vis. (WACV)*, pages 2587–2596, 2020.

[35] Mahdyar Ravanbakhsh, Moin Nabi, Hossein Mousavi, Enver Sangineto, and Nicu Sebe. Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection. In *Wint. Conf. on Appli. of Comp. Vis. (WACV)*, pages 1689–1698, 2018.

[36] Mahdyar Ravanbakhsh, Enver Sangineto, Moin Nabi, and Nicu Sebe. Training adversarial discriminators for cross-channel abnormal event detection in crowds. In *Wint. Conf. on Appli. of Comp. Vis. (WACV)*, pages 1896–1904, 2019.

[37] Ankit Parag Shah, Jean-Bapstite Lamare, Tuan Nguyen-Anh, and Alexander Hauptmann. Cadp: A novel dataset for cctv traffic camera based accident analysis. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–9. IEEE, 2018.

[38] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-World Anomaly Detection in Surveillance Videos. In *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pages 6479–6488, 2018.

[39] Stanislaw Szymanowicz, James Charles, and Roberto Cipolla. Discrete neural representations for explainable anomaly detection. In *Wint. Conf. on Appli. of Comp. Vis. (WACV)*, pages 1506–1514, 2022.

[40] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proc. Int. Conf. Comput. Vis. (ICCV)*, pages 4489–4497, 2015.

[41] Vercruyssen Vincent, Meert Wannes, and Davis Jesse. Transfer learning for anomaly detection through localized and unsupervised instance selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6054–6061, 2020.

[42] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 41:2740–2755, 2019.

[43] Tian Wang, Meina Qiao, Zhiwei Lin, Ce Li, Hichem Snoussi, Zhe Liu, and Chang Choi. Generative neural networks for anomaly detection in crowded scenes. *IEEE Transactions on Information Forensics and Security*, 14(5):1390–1399, 2019.

[44] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8688–8696, 2018.

[45] Peng Wu, jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Eur. Conf. on Comp. Vis. (ECCV)*, 2020.

[46] Muhammad Zaigham Zaheer, Jin-Ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pages 14171–14181, 2020.

[47] Tianzhu Zhang, Hanqing Lu, and Stan Z. Li. Learning semantic scene models by object classification and trajectory clustering. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1940–1947, 2009.

[48] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H. Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1237–1246, 2019.

[49] Yi Zhu and Shawn D. Newsam. Motion-aware feature for improved video anomaly detection. In *Proc. British Machine Vis. Conf. (BMVC)*, pages 270–282, 2019.