CLAWS: Clustering Assisted Weakly Supervised Learning with Normalcy Suppression for Anomalous Event Detection

Muhammad Zaigham Zaheer[1,2], Arif Mahmood[3], Marcella Astrid[1,2], and Seung-Ik Lee[1,2]

[1] University of Science and Technology, Daejeon, South Korea.
[2] Electronics and Telecommunication Research Institute, Daejeon, South Korea.
{mzz, marcella.astrid}@ust.ac.kr, the_silee@etri.re.kr
[3] Information Technology University, Ferozpur Road, Lahore, Pakistan
arif.mahmood@itu.edu.pk

Abstract. Learning to detect real-world anomalous events through video-level labels is a challenging task due to the rare occurrence of anomalies as well as noise in the labels. In this work, we propose a weakly supervised anomaly detection method which has manifold contributions including 1) a random batch based training procedure to reduce inter-batch correlation, 2) a normalcy suppression mechanism to minimize anomaly scores of the normal regions of a video by taking into account the overall information available in one training batch, and 3) a clustering distance based loss to contribute towards mitigating the label noise and to produce better anomaly representations by encouraging our model to generate distinct normal and anomalous clusters. The proposed method obtains 83.03% and 89.67% frame-level AUC performance on the UCF-Crime and ShanghaiTech datasets respectively, demonstrating its superiority over the existing state-of-the-art algorithms.

Keywords: Weakly Supervised Learning, Anomaly Detection, Abnormal Event Detection, Noisy Labeled Training, Event Localization

1 Introduction

Anomalous event detection is an important computer vision domain because of its real-world applications in autonomous surveillance systems [29, 47, 14, 46, 24]. Attributed to the rare occurrences, anomalies are often seen as deviations from normal patterns, activities, or appearances. Hence, a widely popular approach for anomaly detection is to train a one-class classifier which can encode frequently occurring behaviors using only normal training examples [20, 62, 25, 56, 11, 40, 45, 37, 38, 46, 10]. Anomalies are then detected based on their distinction from the learned normalities. A drawback of such methods is the lack of representative training data capturing all variations of the normal behavior [5]. Therefore, a new occurrence of a normal event may deviate enough from the trained patterns to be flagged as anomaly, hence causing false alarms [10]. With

the recent popularity in weakly supervised learning algorithms [19, 21, 59, 30, 43, 54], another interesting approach is to train a binary classifier using weakly labeled training data containing both normal and anomalous videos [46, 63]. A video is labeled as normal if all of its frames are normal and labeled as anomalous if some frames are abnormal. It means that a video labeled as anomalous may also contain numerous normal scenes. In such weakly supervised algorithms, neither temporal nor spatial annotations are needed which considerably reduces the laborious efforts required to obtain the fine-grained manual annotations of the training dataset.

Weakly supervised anomaly detection problem has recently been formulated as Multiple Instance Learning (MIL) task [1, 46]. A video is considered as a bag of segments in which each segment contains several consecutive frames and the training is carried out using video-level annotations by computing a ranking loss between the two top-scoring segments, one from normal and the other from anomalous bag [46]. However, this method requires each video to have the same number of segments throughout the dataset which may not always be a practical approach. Since real-world scenarios may contain significantly varying length of videos, the events occurring in a small temporal range will not be represented well due to this rigid formulation. More recently, another approach towards weakly supervised anomaly detection has also been proposed as learning under noisy labels in which the noise refers to normal segments within anomalous videos [63]. Although it demonstrates superior performance, the method is prone to data correlation since it is trained using a whole video at each training iteration. Such correlation can get even stronger in the case of datasets recorded using stationary cameras. As reported in several existing works [4, 27, 28], data correlation can significantly deteriorate the learning performance of a deep network.

In contrast, we propose a batch based training architecture where a batch consists of several temporally consecutive segments of a video. Depending on the length of a video, several batches may be extracted from it. In each training iteration, we arbitrarily select a batch across the whole training dataset to make the batches independent and identically distributed, thus eradicating inter-batch correlation for a stable and enhanced training. We still, however, utilize the temporal consistency information within a batch for better classification. Extensive experiments demonstrated the efficacy of our proposed random batch selection mechanism as it yields significant performance improvements. Detailed discussion on this is provided in the results section.

Together with this batch based training scheme, a complementary attention-like mechanism may also contribute towards the improvement of an anomaly detection system. Various forms of attention have been introduced in many machine learning architectures with a common goal to highlight important regions within an input image or video [6, 42, 55, 12, 52]. Such mechanisms can be suitable with fully-supervised algorithms in which the attention layers is trained to highlight the important features corresponding to the class annotations of the training data. However, as our proposed approach is weakly supervised, we consider the problem as suppressing the features that correspond to normal events.

Therefore, we propose a normalcy suppression mechanism that operates over a full batch and learns to suppress the normal features towards smaller values. Our formulation exploits the fact that the labels in normal training videos are noise free. Thus, in the case of an input containing anomalous portions, suppression tries to reduce the impact of normal regions within that input while keeping only the high anomaly regions active. Whereas, in the case of an input containing only normal portions, suppression distributes across the whole batch hence complementing the backbone network towards generating lower anomaly scores.

Furthermore, we also propose a clustering distance based loss, the intuition of which is derived from the semi-supervised usage of clustering techniques [15, 7, 44]. To this end, we propose unsupervised clustering to collaborate with our network resulting in an improved overall performance. We first assume two clusters considering that anomaly detection is a binary problem and an anomalous labeled video may also contain normal segments. Then, with each training iteration, the network attempts to minimize (or maximize) the inter-cluster distance in the case of a normal (or an abnormal) video. The clustering is performed on an intermediate representation taken from our backbone network. Therefore, minimization of our formulated clustering loss encourages the network to produce discriminative intermediate representations thus enhancing its anomaly detection performance.

The main contributions of this work are as follows:

- Our CLAWS (CLustering Assisted Weakly Supervised) Net framework is trained in a weakly supervised fashion using only video-level labels to localize anomalous events.
- We propose a simple yet effective random batch selection scheme which enhances the performance of our system by removing inter-batch correlation.
- We also propose a normalcy suppression mechanism which, by exploiting full information of a batch, learns to suppress the features corresponding to the normal portions of an input.
- We formulate a clustering distance based loss which encourages the network to decrease the distance between the clusters created using a normal video and increase it for the clusters created using an anomalous video, resulting in an improved discrimination between normal and abnormal events.
- Our framework achieves frame level AUC performance of 83.03% on UCF-Crime [46] dataset and 89.67% on ShanghaiTech [24] dataset, outperforming the existing state-of-the-art approaches [63, 23, 10, 46].

2   Related Work

Anomaly Detection as One Class Classification Conventionally, anomaly detection has been tackled as learning normalcy in which test data is matched against learned representations of a normal class and the deviating instances are declared as anomalies. Some researchers proposed to train one-class classifiers using handpicked features [26, 3, 53, 61, 36] while others proposed to use

the features extracted from pre-trained deep convolution models [45, 38]. Image regeneration based architectures [9, 39, 57, 13, 32, 33, 58, 40, 41] make use of a generative network to learn normal data representations in an unsupervised fashion. The intuition is that a generator cannot reconstruct unknown classes, hence it may generate high reconstruction errors while reconstructing anomalies. Few researchers also proposed pseudo-supervised methods in which fake anomaly examples are created using the normal data to transform the one-class problem into a binary-class problem [13, 60]. Our architecture, however significantly deviates from the one-class training protocol used in these methods as we utilize both normal and weakly-labeled anomalous data during training.

Anomaly Detection as Weakly Supervised Learning This category utilizes the noisy annotations to carry out training on image datasets [18, 8, 50, 35, 2, 31, 17]. In such models, either the loss correction is applied [2] or noise models are specifically trained to separate out the noisy labeled data [18, 50]. Our work is different from these image based weakly supervised methods as we try to tackle video based anomalies in which temporally-ordered sequence of frames are required.

In essence, the most similar works to ours are [46] and [63], which also attempt to train anomaly detection models using video-level annotations. Sultani et al. [46] propose to formulate the weakly supervised problem as Multiple Instance Learning (MIL). A video is considered as a bag of segments. To train the network, a ranking loss is computed between the top scoring segments from a normal and an anomalous bag. Each training iteration involves several pairs of such bags. In Zhong et al. [63], the authors attempt to clean noisy labels in anomalous videos by using graph convolutional neural networks. A training iteration is carried out based on one complete video from the training dataset. In contrast to these methods, our approach attempts to train a batch based model which learns to maximize scores of the anomalous parts of an input, where a batch corresponds to a portion of a training video. Furthermore, we also propose a normalcy suppression module which, by exploiting the *noise free* labels in the normal labeled videos, learns to suppress the features corresponding to normal regions of a video. Lastly, a clustering distance based loss is also introduced which improves the capability of our model to produce better anomaly representations.

Normalcy Suppression The normalcy suppression module in our architecture can be seen as a variant of attention [51, 6, 42, 55, 12, 52]. However, attributed to the rare occurrence of anomalies, we tackle the problem in terms of suppressing features as opposed to highlighting [55, 12, 52]. Specifically, we define the problem by relying on the availability of *noise free* normal video annotations. Therefore, unlike the conventional attention which utilizes a weighted linear combination of features as in attention, our model learns to suppress the normalcy by obtaining an element-wise product of the suppression scores with the features.

3   Proposed Architecture

In this section, we present our CLAWS Net framework. Various components of the proposed architecture, as shown in Fig. 1, are discussed below:
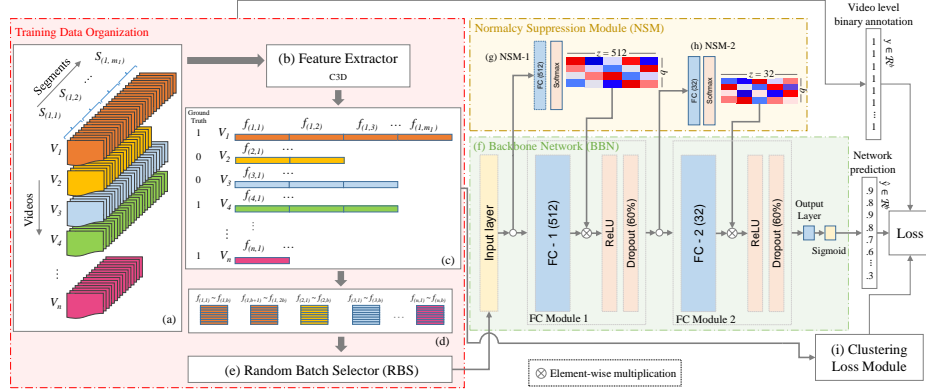
Fig. 1: CLAWS Net: the proposed framework for anomaly detection using video-level weak supervision. (a) Input videos are divided into segments of equal length. (b)&(c) A feature vector is extracted from each segment. (d) Feature vectors of each video are divided into batches of the same size. (e) Batches are randomly selected for training. (f) Backbone network. (g)&(h) Normalcy suppression modules. (i) Clustering loss module.

3.1 Training Data Organization

The proposed training data organization is shown in Fig. 1(a)-(e). Given n training videos in a dataset, each training video V_i is divided into m_i non-overlapping segments $S_{(i,j)}$ of size p frames, where $i \in [1, n]$ is the video index and $j \in [1, m_i]$ is the segment index. For each segment $S_{(i,j)}$, a feature vector $f_{(i,j)} \in \mathcal{R}^d$ is computed as $f_{(i,j)} = \mathcal{E}(S_{(i,j)})$ using a feature extractor $\mathcal{E}(\cdot)$ such as Convolution 3D (C3D) [49]. Consecutive feature vectors are arranged in batches B_k, each consisting of b feature vectors such that $B_k = \{f_{(i,j)}, f_{(i,j+1)}, \cdots, f_{(i,j+b-1)}\} \in \mathcal{R}^{b \times d}$, where $k \in [1, K]$ is the batch index of K number of batches in the training data which is used by the Random Batch Selector (RBS) in Fig. 1(e) to retrieve batches randomly for training. All feature vectors maintain their temporal order within a batch, as shown in Fig. 1(d). For each video we have binary labels as {normal, abnormal} represented as {0, 1}. As the training is performed in a weakly supervised fashion, each batch also inherits the labels of its features from the parent video.

In the existing weakly supervised anomaly detection systems, a training iteration (i.e. one weight update) is carried out on one or several complete videos [46, 63]. In contrast to this practice, we propose several batches to be extracted from videos and then input to the backbone network in an arbitrary order using RBS. The configuration serves two purposes: 1) It minimizes correlation between consecutive batches while keeping the temporal order within batches which is necessary to carry out the weakly supervised anomaly detection training. 2) It allows our network to have more learning instances as our training iteration is carried out using a small portion of a video (batch) instead of a complete video. We also observed that breaking the temporal order between consecutive batches
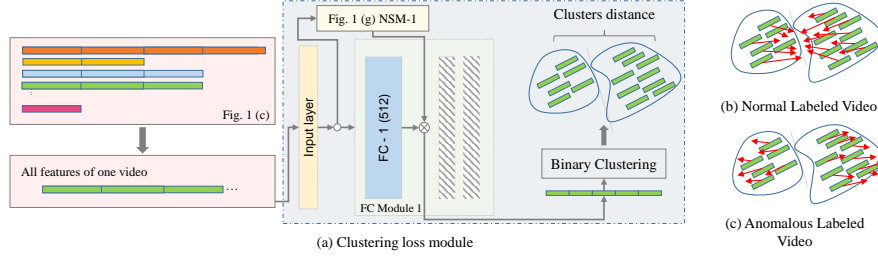
Fig. 2: The proposed clustering loss module (a) which encourages our network to bring the clusters closer in the case of normal labeled videos (b) or push farther in the case of anomalous labeled videos (c). The FC Module 1 and the NSM-1 are same as the ones in Fig. 1 and are only used for inference in this module.

results in a significant increase in the backbone performance, which is discussed in the results section.

3.2   Backbone Network

The backbone network (BBN) consists of two fully connected (FC) modules each containing an FC layer followed by a ReLU activation function and a dropout layer. The input layer receives each batch from RBS. The output layer has a sigmoid activation function which produces anomaly score prediction $y \in \mathcal{R}^b$ of range $[0, 1]$. Training of the BBN is carried out using video-level labels. Hence, a batch from a normal video will have labels $y = \mathbf{0} \in \mathcal{R}^b$ whereas a batch from an anomalous video will have labels $y = \mathbf{1} \in \mathcal{R}^b$, where $\mathbf{0}$ is an all-zeros vector, $\mathbf{1}$ is an all-ones vector, and b is batch size.

3.3   Normalcy Suppression

Our proposed architecture has multiple normalcy suppression modules (NSM) as shown in Figs. 1 (g)-(h). Each NSM contains an FC layer, kept consistent in dimensions with the FC layer in its respective FC Module, followed by a softmax layer. An NSM computes probability values across temporal dimension of the input batch therefore, serves as a global information collector over the whole batch. Based on the FC layer dimension z and the number of features b in each input batch (Fig. 1 (d)), an NSM outputs a probability matrix $\mathcal{P}$ of size $b \times z$, such that the sum of each column in this matrix is 1. An element-wise multiplication between $\mathcal{P}$ and the output of FC layer is performed in the corresponding FC Module.

    Our approach exploits the fact that all normal labeled videos have noise free labels at segment level as the anomalies do not appear in these videos. During training, in case an input batch is taken from a normal labeled video, all features in this batch are labeled as normal and the BBN is expected to produce low anomaly scores on all input feature vectors. Therefore, the NSM learns to complement the BBN towards minimizing the overall training loss by distributing its probabilities across the whole input batch and not highlighting any portions of

an input. This phenomenon is particularly the desideratum behind choosing such element-wise configuration which provides more freedom to the NSM towards minimizing its values at each feature dimension of the whole batch. In case an input batch is taken from an abnormal video and consequently all features of the batch are labeled as abnormal, it may be expected that the BBN will produce high anomaly scores on all input features. However, with normal batches as part of the training data, the BBN is also concurrently trained to produce low anomaly scores on the normal segments. Therefore, to some extent, the BBN achieves the capability to distinguish between normal and anomalous segments of an anomalous video. NSM further complements this capability of the BBN by suppressing the features of normal segments. Given an anomalous batch as input, to minimize the overall training loss, NSM thus learns to suppress the portions of the input batch that do not contribute strongly towards the anomaly scoring in the BBN, consequently highlighting anomalous portions.

3.4   Clustering Loss Module

A clustering distance based loss is formulated with an intuition to encourage our network to successfully group deep video features into normal and anomalous clusters. As mentioned previously, each feature vector $f_{(i,j)}$ inherits its label of being normal or abnormal from the parent video. A normal label means all segments are normal. However an abnormal label does not mean all segments are anomalous rather, some segments are anomalous while the remaining are normal. We propose to cluster a sub-representation of all the feature vectors $f_{(i,j)}$ of each training video into two clusters. In case of a normal video, we try to bring the centers of the two cluster as close to each other as possible assuming that both clusters correspond to normal segments (Fig. 2(b)). In case of a video with abnormal label, we try to push the centers of the two clusters away from each other assuming that one cluster should contain normal segments while the other should contain abnormal (Fig. 2(c)). As the clustering is performed on an intermediate representation inferred from the BBN, this loss results in an improved capability of our network to represent anomalies and consequently an enhanced anomaly detection performance of the proposed model. Specifically, the distance d_i between centers of the two clusters for a video V_i containing m_i number of segments is given as:

$$d_i = \frac{1}{m_i}\|\mathbf{c_1} - \mathbf{c_2}\|_2, \tag{1}$$

where $\mathbf{c_1}$ and $\mathbf{c_2}$ are the centers of the two clusters. As the training videos may have varying length, longer videos will have more batches than shorter videos therefore, m_i is used to normalize distance values across the training dataset.

At the beginning of each training epoch, for all feature vectors $f_{(i,j)}$ of a video V_i, intermediate representations are computed using the BBN as shown in Fig. 2(a). These resulting vectors are then grouped into two clusters using k-means clustering [22] and d_i is computed. This d_i is then used with the corresponding batches at each training iteration.

3.5   Training Losses of the Proposed Algorithm

Training of our model is carried out to minimize regression, clustering distance, temporal smoothness, and sparsity losses, as explained below:

Regression Loss: The proposed CLAWS Net mainly performs regression to minimize mean square error using the video labels directly assumed towards each feature of the input batch: $\mathcal{L}_{pred} = \frac{1}{b} \sum_{l=1}^{b} (y_l - \hat{y}_l)^2$, where y_l and $\hat{y}_l$ denote l-th ground truth and l-th predicted values in a batch respectively, and b is the batch size.

Clustering Distance Loss: Given clustering distance (d_i), the clustering distance loss $\mathcal{L}_c$ of a video V_i is given as:

$$\mathcal{L}_c = \begin{cases} min(\alpha, d_i), & \text{if } V_i \text{ is Normal} \\ \frac{1}{d_i}, & \text{if } V_i \text{ is Abnormal,} \end{cases} \tag{2}$$

where α is an upper bound on the clustering distance loss which helps to make the training more stable in the presence of much varied video data.

Temporal Smoothness Loss: It is applied based on the fact that most events are temporally consistent. Our proposed architecture maintains temporal order among the feature vectors in each input batch therefore, similar to [46], we apply temporal smoothness loss $(\mathcal{L}_{ts})$ as: $\mathcal{L}_{ts} = \sum_{l=1}^{b-1} (\hat{y}_{l+1} - \hat{y}_l)^2$, where $\hat{y}_l$ is the l-th prediction in a batch of size b.

Sparsity Loss: The sparsity loss, proposed previously in [46], exploits the fact that anomalous events occur rarely as compared to the normal events. Hence, cumulative anomaly score of a complete video should be comparatively small. We compute this loss on each batch during training as: $\mathcal{L}_s = \sum_{l=1}^{b} \hat{y}_l$, where $\hat{y}_l$ is the l-th prediction in a batch of size b.

Overall Loss Function: Finally, complete loss of the proposed network is computed as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{pred} + (1 - \lambda_1)\mathcal{L}_c + \lambda_2(\mathcal{L}_s + \mathcal{L}_{ts}), \tag{3}$$

where λ_1 and λ_2 assign relative importance to different loss parameters.

4   Experiments

4.1   Datasets

Experiments have been conducted on two different video anomaly detection datasets including UCF-Crime [46] and ShanghaiTech [24].

UCF-Crime is a large-scale complex dataset which spans over 128 hours of videos (resolution 240×320 pixels), captured through CCTV surveillance cameras and contains 13 different classes of real world anomalies [46]. Its training split contains 800 normal and 810 anomalous videos, while the test split has 150 normal and 140 anomalous videos. Videos labeled as normal do not contain any abnormal scenes whereas, videos labeled as abnormal contain anomalous as well
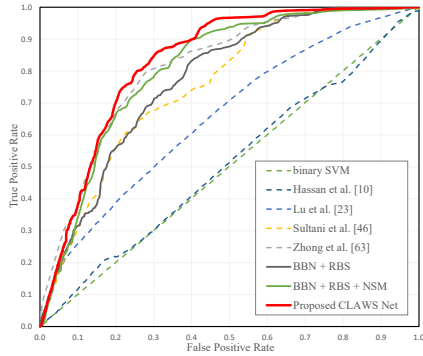
Fig. 3: ROC curves comparison with state-of-the-art using C3D features on UCF-Crime.
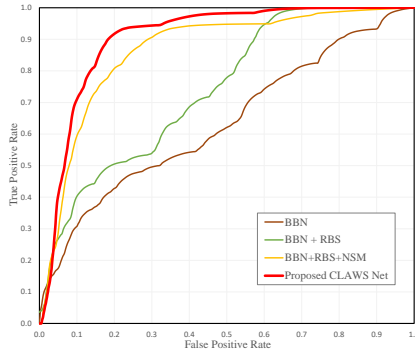
Fig. 4: ROC curves comparison of the variants of our proposed algorithm on ShanghaiTech.

as normal scenes. In the training split, video-level binary labels are provided which can only be used by weakly supervised algorithms. In the test split the labels are provided at the frame level to facilitate the evaluation of anomaly localization.

ShanghaiTech is a medium-scale dataset for abnormal event detection captured in a university campus with staged anomalous events. It contains 437 videos ($317, 398$ frames of resolution 480×856 pixels) captured at 13 different locations with varying lighting conditions and camera angles. The original protocol of this dataset is to train one-class classifiers which means that the training dataset contains only normal videos. In order to make it suitable for evaluating weakly supervised binary classification architectures, Zhong et al. [63] reorganized the dataset into a mix of normal and anomalous videos in both testing and training splits. Their training split contains 175 normal and 63 anomalous videos and the test split contains 155 normal and 44 anomalous videos. For fair comparison, we follow this protocol for the training and evaluation of our system.

4.2   Evaluation Metric

Following previous works [23, 63, 46, 10], we use Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve, calculated with respect to the frame-level ground truth annotations of the test videos, as our evaluation metric. A larger AUC implies better discrimination performance at various thresholds. Since the number of normal frames in the test data is much larger than the number of anomalous frames, Equal Error Rate (EER) may not be a suitable measure [46].

4.3   Experimental Settings

Implementation of our model is carried out in PyTorch [34]. We train for a total of 100k iterations using RMSProp [48] optimizer with an initial learning rate of

Table 1: Frame-level AUC performance comparison on UCF-Crime (BBN: Backbone network, RBS: Random batch Selector, NSM: Normalcy suppression module).

Method	Features Type	AUC(%)
SVM Baseline	C3D	50.00
Hasan et al. [10]	C3D	50.60
Lu et al. [23]	C3D	65.51
Sultani et al. [46]	C3D	75.41
Zhong et al. [63]	C3D	81.08
Zhong et al. [63]	$TSN^{OpticalFlow}$	78.08
Zhong et al. [63]	TSN^{RGB}	82.12
BBN	C3D	69.50
BBN + RBS	C3D	75.95
BBN + RBS + NSM	C3D	80.94
Proposed CLAWS Net	**C3D**	**83.03**

0.0001 and a 10 times decrease after 80k iterations. In all our experiments, λ_1, λ_2 and α are set to 0.90, 8.0×10^{-5} and 1.0, respectively. FC layers of the FC Module 1 and FC Module 2 are set to have 512 and 32 channels. C3D [49], a 3D-convolution model for videos pre-trained on Sports-1M dataset [16], is employed as the feature extractor in Fig. 1 (b). Default feature extraction settings of C3D, as proposed in [49, 63, 46], are used. Mean normalization is applied on the extracted features. Batch size b is set to 64 feature vectors, segment $S_{(i,j)}$ size is set to 16 frames and feature $f_{(i,j)}$ size d is 2048.

4.4   Experiments on UCF-Crime Dataset

We train our proposed model on UCF-Crime dataset using only video-level labels. Fig. 3 and Table 1 visualize a comparison of our method with current state-of-the-art approaches. In both types of comparisons, our CLAWS Net shows superior performance than the compared algorithms. Interestingly, our proposed RBS enhances the performance of the BBN to 75.95% which is superior than the performance reported by Hasan et al. [10], Lu et al. [23] and Sultani et al. [46]. It is because of the significant decrease in the inter-batch correlation which improves the learning of BBN. Noticeably, compared to Zhong et al. [63], our approach is fairly simple as we do not train a deep action classifier. However, using the similar C3D features, our CLAWS Net depicts 1.95% improved performance.

4.5   Experiments on ShanghaiTech

The CLAWS Net framework is evaluated on ShanghaiTech following the test/train split proposed by Zhong et al. [63]. Being a recent split, other existing methods have not reported performance on this dataset. Using the same protocol, our framework outperforms Zhong et al. [63] by a significant margin of 13.23% when both algorithms use similar C3D features (see Table 2). Zhong et al [63] reported relatively better performance using TSN features, however our proposed model outperforms their TSN based performance as well by 5.23%. In our

Table 2: Frame-level AUC performance comparison on ShanghaiTech (BBN: Backbone network, RBS: Random batch Selector, NSM: Normalcy suppression module).

Method	Features Type	AUC %
Zhong et al. [63]	C3D	76.44
Zhong et al. [63]	$TSN^{OpticalFlow}$	84.13
Zhong et al. [63]	TSN^{RGB}	84.44
BBN	C3D	67.50
BBN + RBS	C3D	79.64
BBN + RBS + NSM	C3D	87.76
Proposed CLAWS Net	**C3D**	**89.67**

Table 3: Ablation studies of our proposed approach on UCF-Crime (BBN: Backbone network, RBS: Random batch Selector, NSM: Normalcy suppression module, $\mathcal{L}_s$: Sparsity loss, $\mathcal{L}_{tc}$: Temporal consistency loss, $\mathcal{L}_c$: Clustering distance loss).

BBN	RBS	NSM-1	NSM-2	$\mathcal{L}_s + \mathcal{L}_{tc}$	$\mathcal{L}_c$	AUC (%)
Bottom-up Approach						
✓	-	-	-	-	-	69.50
✓	✓	-	-	-	-	75.95
✓	✓	✓	-	-	-	78.60
✓	✓	✓	✓	-	-	80.94
✓	✓	✓	✓	✓	-	81.53
Top-down Approach						
✓	-	✓	✓	✓	✓	80.23
✓	✓	-	✓	✓	✓	77.39
✓	✓	✓	-	✓	✓	79.78
✓	✓	-	-	✓	✓	76.81
✓	✓	✓	✓	-	✓	82.41
✓	✓	✓	✓	✓	✓	83.03

work, the reported results are evaluated using C3D features which is in consistence with most of the existing methods. However, using features with better representation may result in a further improved performance. An ROC curve based performance comparison of the three variants of our proposed approach is provided in Figure 4. We observe a performance boost of 12.14% by adding RBS to the BBN. Furthermore, 8.12% boost is observed by using NSM along with BBN+RBS. Finally the complete system, which incorporates all losses as well, further enhances the performance by 1.91%. These experiments highlight the importance of each component in CLAWS Net.

4.6   Ablation

We performed two types of ablation studies by using bottom-up as well as top-down approaches on the UCF-Crime dataset as shown in Table 3. In the former approach, we started with the evaluation of only Backbone Network (BBN) and kept on adding different modules while observing the performance boost. In the

(a) Output of NSM-1

(b) Output of NSM-2

(c) Cumulative response of NSM-1
on each feature segment
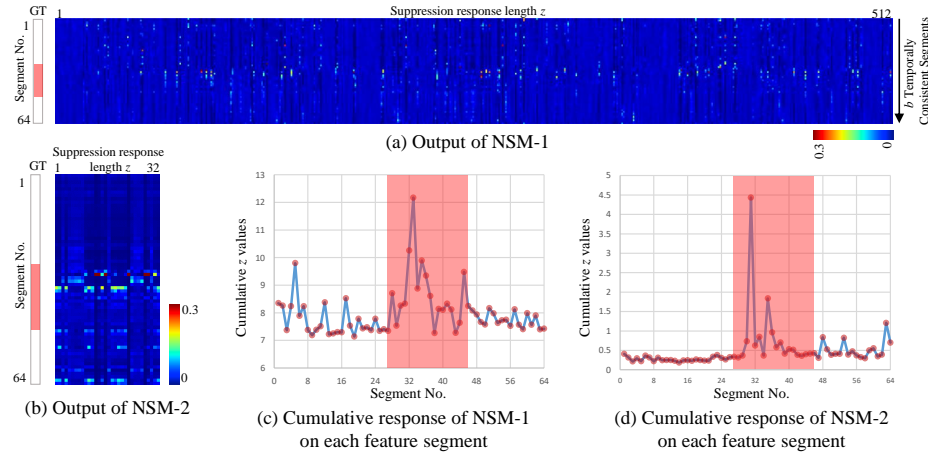
(d) Cumulative response of NSM-2
on each feature segment

Fig. 5: Softmax output of our proposed NSM-1 (a) and NSM-2 (b) on an input batch from an anomalous test video. It can be seen that both modules learn to successfully suppress the normal regions of an input. Cumulative suppression values of these modules on each individual feature of segments is visualized in (c) & (d) which also provides an overall insight on the response of each NSM towards an anomalous input. Red colored windows show anomaly ground truth.

later approach we started with the whole CLAWS Net and removed different modules to observe the consequent performance degradation.

The BBN achieves 69.5% AUC while the addition of RBS enhances the performance to 75.95% which validates its importance in our approach. Addition of the NSM-1 results in an improved performance of 78.60% whereas the NSM-2 further elevates the performance to 80.94%. The overall boost in the AUC by NSM-1 is larger than by NSM-2. It can mainly be attributed to the size of the FC layer in NSM-1, which is 16 times larger than the FC layer in NSM-2. Figs. 5 (a) & (b) show the responses of NSM-1 & 2 for an anomalous test video batch. The response corresponding to the anomalous events is much higher compared to the response on normal events. These values are multiplied with the FC layer output and thus suppress the values corresponding to the normal events. Figs. 5 (c) & (d) show the cumulative response of NSM-1 and NSM-2 modules computed by summation along the response length z. The cumulative response in the anomalous region is significantly higher than the response in the normal regions. It clearly demonstrates that the NSM modules successfully learn to suppress the normal regions, consequently highlighting the anomalous regions, at two intermediate levels of the network. Table 3 also shows that the addition of temporal consistency and sparsity losses brings the performance to 81.53%, and finally the addition of the clustering distance loss yields an overall system performance of 83.03%. This study depicts the importance of each component in our proposed architecture.
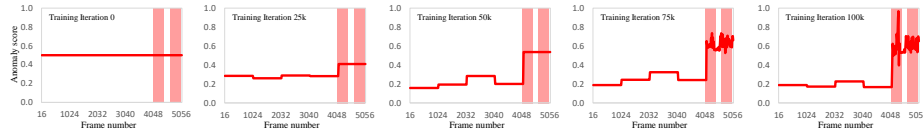
Fig. 6: Evolution of frame-level anomaly scores output by our network over several training iterations. Although weakly supervised, our framework learns to produce significantly higher scores in the anomalous portions whereas low scores in the normal portions of a video. Red colored windows show anomaly ground truth.

In the top-down approach (Table 3) deletion of RBS from our CLAWS Net causes a drop of 2.8% AUC. Compared to this, the addition of RBS to BBN in the bottom-up approach resulted in an improvement of 6.45%. Thus, some of the performance of RBS is compensated by the other components in the complete system, however it cannot be fully replaced. Furthermore, deletion of NSM-1 and NSM-2 from the overall architecture causes a drop of 5.64% and 3.25% respectively. Consistent to the results in the bottom-up approach, it demonstrates a relatively higher importance of NSM-1 than NSM-2, mainly due to the larger size of its FC layer. Another contributing factor is that the NSM-1 directly affects the clustering loss module. As shown in Fig. 2, the clustering is performed after multiplying the NSM-1 response with the output of FC-1. If both NSM-1 and NSM-2 are removed, the remaining system achieves 76.81% AUC which is 6.22% less than the complete CLAWS Net performance. However, the addition of NSM-1 and NSM-2 to BBN+RBS in the bottom-up approach caused an improvement of 4.99%. This demonstrates more importance of the NSM modules than what is observed in the bottom-up approach due to their direct effect on the clustering loss module as well as indirect effect on other losses.

4.7   Qualitative Analysis

The evolution of the proposed anomaly detection system over several training iterations is shown in Fig. 6. As the number of training iterations increases, the difference between the response of our system on normal and anomalous regions also increases. Fig. 7 provides a comparison of the anomaly scores produced by different configurations of our proposed architecture on two normal and four anomalous test videos from the UCF crime dataset. In these cases, the BBN was not able to accurately discriminate between the normal and anomalous regions. The addition of RBS showed significant improvements except in the *shooting002* case where the improvement is small. It can be observed that with the addition of normalcy suppression modules, the difference between anomaly scores on normal and abnormal regions became higher which is a desirable property in the anomaly detection systems. The proposed suppression not only pushed the anomaly scores of the normal regions towards 0 but also created a smoothing effect. The response of our complete system, CLAWS Net, is more stable as well as discriminative than all the other variants in most of the cases. It should also be noted that Fig.
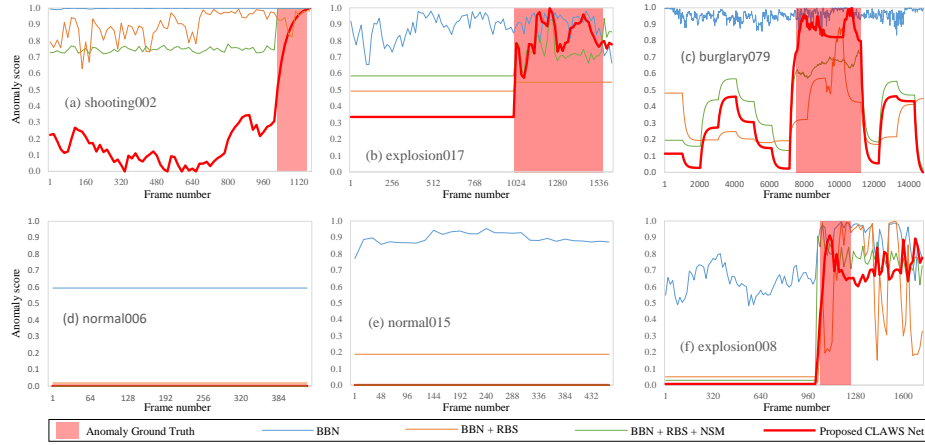
Fig. 7: Qualitative results of our method on test videos of the UCF-Crime dataset. (a), (b) & (c) show anomaly cases while (d) & (e) depict normal cases. (f) illustrate a relatively unsuccessful case of our system in which, the network keeps showing higher anomaly scores after the explosion. Red colored windows show anomaly ground truth.

7 (f) shows a relatively unsuccessful case in which the system continues to show high anomaly score even after the anomalous event is over, which is not unlikely due to the aftermath of an explosion. The annotations in the dataset are marked only for the duration of explosion, while the scenes in such an event may remain abnormal for a significantly longer time.

5   Conclusions

In this study, we present a weakly supervised anomalous event detection system trainable using only video-level labels. Unlike the existing systems which utilize complete video based training iterations, we adopt a batch based training. A batch may have several temporally ordered segments of a video and one video may be divided into several batches depending on its length. Selection of these batches in random order helps breaking inter-batch correlation and demonstrates a significant boost in performance. A normalcy suppression mechanism is also proposed which collaborates with the backbone network in detecting anomalies by learning to suppress the features corresponding to the normal portions of an input video. Moreover, a clustering distance based loss is formulated, which improves the capability of our network to better represent the anomalous and normal events. Validation of the proposed architecture on two large datasets (UCF-Crime & ShanghaiTech) demonstrates SOTA results.

6   Acknowledgements

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Advances in neural information processing systems. pp. 577–584 (2003)
2. Azadi, S., Feng, J., Jegelka, S., Darrell, T.: Auxiliary image regularization for deep cnns with noisy labels. arXiv preprint arXiv:1511.07069 (2015)
3. Basharat, A., Gritai, A., Shah, M.: Learning object motion patterns for anomaly detection and improved object detection. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
4. Bengio, Y.: Practical recommendations for gradient-based training of deep architectures. In: Neural networks: Tricks of the trade, pp. 437–478. Springer (2012)
5. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM computing surveys (CSUR) **41**(3), 1–58 (2009)
6. Chen, X., Xu, C., Yang, X., Tao, D.: Attention-gan for object transfiguration in wild images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 164–180 (2018)
7. Fogel, S., Averbuch-Elor, H., Cohen-Or, D., Goldberger, J.: Clustering-driven deep embedding with pairwise constraints. IEEE computer graphics and applications **39**(4), 16–27 (2019)
8. Goldberger, J., Ben-Reuven, E.: Training deep neural-networks using a noise adaptation layer (2016)
9. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
10. 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. pp. 733–742 (2016)
11. 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. pp. 3619–3627 (2017)
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
13. Ionescu, R.T., Khan, F.S., Georgescu, M.I., Shao, L.: Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7842–7851 (2019)
14. Kamijo, S., Matsushita, Y., Ikeuchi, K., Sakauchi, M.: Traffic monitoring and accident detection at intersections. IEEE transactions on Intelligent transportation systems **1**(2), 108–118 (2000)
15. Kamnitsas, K., Castro, D.C., Folgoc, L.L., Walker, I., Tanno, R., Rueckert, D., Glocker, B., Criminisi, A., Nori, A.: Semi-supervised learning via compact latent space clustering. arXiv preprint arXiv:1806.02679 (2018)
16. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)
17. Larsen, J., Nonboe, L., Hintz-Madsen, M., Hansen, L.K.: Design of robust neural network classifiers. In: Proceedings of the 1998 IEEE International Conference on

Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181). vol. 2, pp. 1205–1208. IEEE (1998)

18. Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., Li, L.J.: Learning from noisy labels with distillation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1910–1918 (2017)

19. Liu, D., Jiang, T., Wang, Y.: Completeness modeling and context separation for weakly supervised temporal action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1298–1307 (2019)

20. 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. pp. 6536–6545 (2018)

21. Liu, Z., Wang, L., Zhang, Q., Gao, Z., Niu, Z., Zheng, N., Hua, G.: Weakly supervised temporal action localization through contrast based evaluation networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3899–3908 (2019)

22. Lloyd, S.: Least squares quantization in pcm. IEEE transactions on information theory **28**(2), 129–137 (1982)

23. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in matlab. In: Proceedings of the IEEE international conference on computer vision. pp. 2720–2727 (2013)

24. 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. pp. 341–349 (2017)

25. 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. pp. 341–349 (2017)

26. Medioni, G., Cohen, I., Brémond, F., Hongeng, S., Nevatia, R.: Event detection and analysis from video streams. IEEE Transactions on pattern analysis and machine intelligence **23**(8), 873–889 (2001)

27. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing atari with deep reinforcement learning. In: NIPS Deep Learning Workshop (2013)

28. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)

29. Mohammadi, S., Perina, A., Kiani, H., Murino, V.: Angry crowds: Detecting violent events in videos. In: European Conference on Computer Vision. pp. 3–18. Springer (2016)

30. Narayan, S., Cholakkal, H., Khan, F.S., Shao, L.: 3c-net: Category count and center loss for weakly-supervised action localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8679–8687 (2019)

31. Natarajan, N., Dhillon, I.S., Ravikumar, P.K., Tewari, A.: Learning with noisy labels. In: Advances in neural information processing systems. pp. 1196–1204 (2013)

32. Nguyen, T.N., Meunier, J.: Anomaly detection in video sequence with appearance-motion correspondence. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)

33. Nguyen, T.N., Meunier, J.: Hybrid deep network for anomaly detection. arXiv preprint arXiv:1908.06347 (2019)

34. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)

35. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1944–1952 (2017)
36. Piciarelli, C., Micheloni, C., Foresti, G.L.: Trajectory-based anomalous event detection. IEEE Transactions on Circuits and Systems for video Technology **18**(11), 1544–1554 (2008)
37. Ravanbakhsh, M., Nabi, M., Mousavi, H., Sangineto, E., Sebe, N.: Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1689–1698. IEEE (2018)
38. Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C., Sebe, N.: Abnormal event detection in videos using generative adversarial nets. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 1577–1581. IEEE (2017)
39. Ren, H., Liu, W., Olsen, S.I., Escalera, S., Moeslund, T.B.: Unsupervised behavior-specific dictionary learning for abnormal event detection. In: BMVC. pp. 28–1 (2015)
40. Sabokrou, M., Fayyaz, M., Fathy, M., Klette, R.: Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Transactions on Image Processing **26**(4), 1992–2004 (2017)
41. Sabokrou, M., Khalooei, M., Fathy, M., Adeli, E.: Adversarially learned one-class classifier for novelty detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3379–3388 (2018)
42. Shen, Y., Ni, B., Li, Z., Zhuang, N.: Egocentric activity prediction via event modulated attention. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 197–212 (2018)
43. Shou, Z., Gao, H., Zhang, L., Miyazawa, K., Chang, S.F.: Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 154–171 (2018)
44. Shukla, A., Cheema, G.S., Anand, S.: Semi-supervised clustering with neural networks. arXiv preprint arXiv:1806.01547 (2018)
45. Smeureanu, S., Ionescu, R.T., Popescu, M., Alexe, B.: Deep appearance features for abnormal behavior detection in video. In: International Conference on Image Analysis and Processing. pp. 779–789. Springer (2017)
46. 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. pp. 6479–6488 (2018)
47. Sultani, W., Choi, J.Y.: Abnormal traffic detection using intelligent driver model. In: 2010 20th International Conference on Pattern Recognition. pp. 324–327. IEEE (2010)
48. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning **4**(2), 26–31 (2012)
49. 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. pp. 4489–4497 (2015)
50. Vahdat, A.: Toward robustness against label noise in training deep discriminative neural networks. In: Advances in Neural Information Processing Systems. pp. 5596–5605 (2017)

51. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
52. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3156–3164 (2017)
53. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1386–1393 (2014)
54. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 4325–4334 (2017)
55. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19 (2018)
56. Xia, Y., Cao, X., Wen, F., Hua, G., Sun, J.: Learning discriminative reconstructions for unsupervised outlier removal. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1511–1519 (2015)
57. Xu, D., Ricci, E., Yan, Y., Song, J., Sebe, N.: Learning deep representations of appearance and motion for anomalous event detection. arXiv preprint arXiv:1510.01553 (2015)
58. Xu, D., Yan, Y., Ricci, E., Sebe, N.: Detecting anomalous events in videos by learning deep representations of appearance and motion. Computer Vision and Image Understanding **156**, 117–127 (2017)
59. Yu, T., Ren, Z., Li, Y., Yan, E., Xu, N., Yuan, J.: Temporal structure mining for weakly supervised action detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5522–5531 (2019)
60. Zaheer, M.Z., Lee, J.h., Astrid, M., Lee, S.I.: Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14183–14193 (2020)
61. Zhang, T., Lu, H., Li, S.Z.: Learning semantic scene models by object classification and trajectory clustering. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1940–1947. IEEE (2009)
62. Zhang, Y., Lu, H., Zhang, L., Ruan, X., Sakai, S.: Video anomaly detection based on locality sensitive hashing filters. Pattern Recognition **59**, 302–311 (2016)
63. Zhong, J.X., Li, N., Kong, W., Liu, S., Li, T.H., Li, G.: Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1237–1246 (2019)