DAM: Dissimilarity Attention Module for Weakly-supervised Video Anomaly Detection

Snehashis Majhi, Srijan Das, François Brémond

DAM : Dissimilarity Attention Module for Weakly-supervised Video Anomaly Detection

Snehashis Majhi[1], Srijan Das[2], François Brémond[1]
[1] INRIA Sophia Antipolis, 2004 Route des Lucioles, 06902, Valbonne, France
[2] Stony Brook University, USA

Abstract

Video anomaly detection under weak supervision is complicated due to the difficulties in identifying the anomaly and normal instances during training, hence, resulting in non-optimal margin of separation. In this paper, we propose a framework consisting of Dissimilarity Attention Module (DAM) to discriminate the anomaly instances from normal ones both at feature level and score level. In order to decide instances to be normal or anomaly, DAM takes local spatio-temporal (i.e. clips within a video) dissimilarities into account rather than the global temporal context of a video. This allows the framework to detect anomalies in real-time (i.e. online) scenarios without the need of extra window buffer time. Further more, we adopt two-variants of DAM for learning the dissimilarities between successive video clips. The proposed framework along with DAM is validated on two large scale anomaly detection datasets i.e. UCF-Crime and ShanghaiTech, outperforming the online state-of-the-art approaches by 1.5% and 3.4% respectively. The source code and models will be available at `https://github.com/snehashismajhi/DAM-Anomaly-Detection`

1. Introduction

Anomaly detection in videos has drawn a significant attention in the vision domain due to its huge applications in intelligent surveillance systems enabling crime prevention and investigation. Detection of anomalies in real-world scenarios is challenging due to the unavailability of large annotated data, sparsity in anomaly occurrence, and ambiguous definition of anomalies. To define, anomaly detection for a given stream of video, we aim at learning the start and end of an anomaly event occurring in a video. For this, previous studies [3, 1, 6] have been learning the distribution of only normal activities defined for a specific environment through uni-class unsupervised learning and treat anomaly as an outlier w.r.t. the learned normal distribution. Since, it is difficult to generalize representations for all possible normal activities, these methods are highly biased to generate false positives in real-world scenarios.

To combat this, recent anomaly detection approaches [11, 17, 15, 7, 4, 16, 13, 14] adopt a weakly-supervised binary classification paradigm where both normal and anomaly videos are taken into account during training. In this setting, for a long untrimmed video sequence, only coarse video-level labels (*i.e. normal and anomaly*) are required for training instead of frame-level annotations. However, in these methods a major challenge lies in identifying the anomaly instances from normal ones to take part in maximizing the margin of separation between the classes. Recent approaches [11, 17, 15, 4, 7] have been addressing this challenge by using Multiple Instance Learning (MIL) with a ranking loss. These approaches inputs feature representation of non-overlapping temporal segments of each video to the MIL model which initially assigns scores to each segments for optimizing the separation between anomaly and normal temporal segments based on their scores by a ranking loss. Broadly, these approaches discriminate the anomaly instances either at the score level [11, 17, 15] by selecting the maximum scoring instances as anomaly or at the feature level [7] by combining several modalities. Either of these approaches are limited due to the inappropriate selection of anomaly instances at score level and difficulties in obtaining discriminative feature representation. This is mainly due to the previous algorithms focusing on modeling the global temporal information ignoring the local contextual information localized in the temporal segments that pertain to have anomalies.

To this end, we propose a framework that not only highlights the salient anomaly instances at the feature level but also at the score level through a Dissimilarity Attention Module (DAM) by taking the local contextual information into account. In contrast to earlier methods [7, 17] where both RGB and Motion modalities are used, we only use RGB modality in our framework to achieve real-time performance as well as to obtain discriminative representation for anomaly instances. Since anomaly videos

differ significantly in their temporal context from that of normal ones, we aim at modeling this attribute for obtaining discriminative representations. However, discriminating anomaly instance by global temporal context modeling using TCNs, LSTMs, are limited by the fact that they can not detect anomalies in real-time detection systems, since they require buffer as the input to the systems. Thus, the key building bock of our framework DAM exploits the temporal context of a video by looking at the neighboring instances rather than the whole video to assign independent attention weights both at the feature and score levels. The proposed framework is validated on two publicly available large scale datasets, namely *UCF-crime* and *ShanghaiTech* achieving state-of-the-art anomaly detection performances.

In summary, the contributions of the paper include:

- A weakly-supervised anomaly detection framework that learns a discriminative representation for anomaly instances both at feature and score levels.
- A dissimilarity attention module that incorporates local contextual information in the anomaly detection for real-time applications.
- An exhaustive experimental analysis to corroborate the robustness of proposed method on two competitive anomaly detection datasets.

2. Related Work

Weakly-supervised anomaly detection has been studied extensively in the past few years [11, 16, 13, 15, 12, 17, 7, 14, 4]. Major previous work can be divided into two categories based on the learning paradigm: (1) Multiple Instance Learning (MIL) based approach, (2) Cleaning Noisy Labels based approach. **Multiple Instance Learning (MIL) Based Approach** was introduced by Sultani *et al.* [11] to overcome the drawbacks of traditional unsupervised one-class learning based anomaly detection methods [1, 6, 3]. Since in weakly-supervised anomaly detection task only video-level labels are provided for learning, authors in [11] only extracts off-the-shelf features from a pre-trained 3D ConvNet backbone and aim at training a classification network through a novel ranking loss function. The optimal separation between normal and anomaly instances is ensured by the ranking loss function by choosing the maximum scoring instances of both normal and anomaly videos for optimization and hence resulting in smaller false positives than that of unsupervised anomaly detection methods. Despite this, Sultani *et al.* were able to produce limited detection performance since they only focus on score level discrimination of the anomaly instances by ignoring the temporal context modeling of videos in order to discriminate anomaly instances at the feature level.

Cleaning Noisy Labels Based Approach was introduced by Zhong *et al.* [16] since they claimed that the MIL based approaches are incompetent in producing higher anomaly detection performances since they are not end-to-end trainable along with the 3D ConvNet backbone due to the unavailability of precise temporal annotations. Thus, Zhong *et al.* [16] aim at training the 3D ConvNet backbone by generating pseudo temporal annotations for untrimmed anomaly videos through a Cleaning Noisy Labels based approach. This enables Zhong *et al.* to learn a discreminative representation for anomaly instances. However, the generation of pseudo temporal annotations is done by training a Graph Convolution Network (GCN). Since training of GCN is computationally complex and can lead to unconstrained latent space, authors in [14] use clustering algorithms for cleaning the noisy labels of untrimmed anomaly videos. Different from [16], Zaheer *et al.* [14] uses the k-means clustering algorithm to produce pseudo temporal annotations for anomalous videos and trained 3D ConvNet backbone for discriminative representation learning of anomaly instances. Since these methods are heavily dependent upon pseudo temporal annotations stage, so a noisy generation of temporal annotations can drastically mislead the training of 3D ConvNet backbone.

Inspired by this, authors in [15] utilize TCN in MIL based approach for obtaining temporal dependency encoding for anomaly instances at the feature level. In addition, they also proposed an inner-bag ranking loss function for improved anomaly detection performance. Another approach [17], combined optical flow features with the RGB feature map for discriminating the anomaly instances that exhibit strong motion in it. Similarly, authors in [7] proposed two-stream framework *i.e. RGB and Social Force* modalities to model the crowd behaviours leading to anomaly patterns. Moreover, they used self-attention mechanism as a feature modulation technique and aggregated the anomaly detection score of both streams to report the final detection performance. Recently, authors in [13] claimed that combining audio features with RGB map can discriminate the anomaly instances effectively at the feature level. In addition, Wu *et al.* [13] also utilize Graph Convolution Network (GCN) for temporal context learning in videos leading to improved anomaly detection performance. Since the usage of TCN and GCN in anomaly detection methods require the whole video for temporal context learning, this makes the system operate on sliding window (offline) mode. Since a major application of anomaly detection methods is to operate on real-time surveillance systems, the usage of optical flow, Social Force, and audio modalities increases the inference time. Thus, we propose a real-time anomaly detection framework composed of a DAM module that not only discriminates

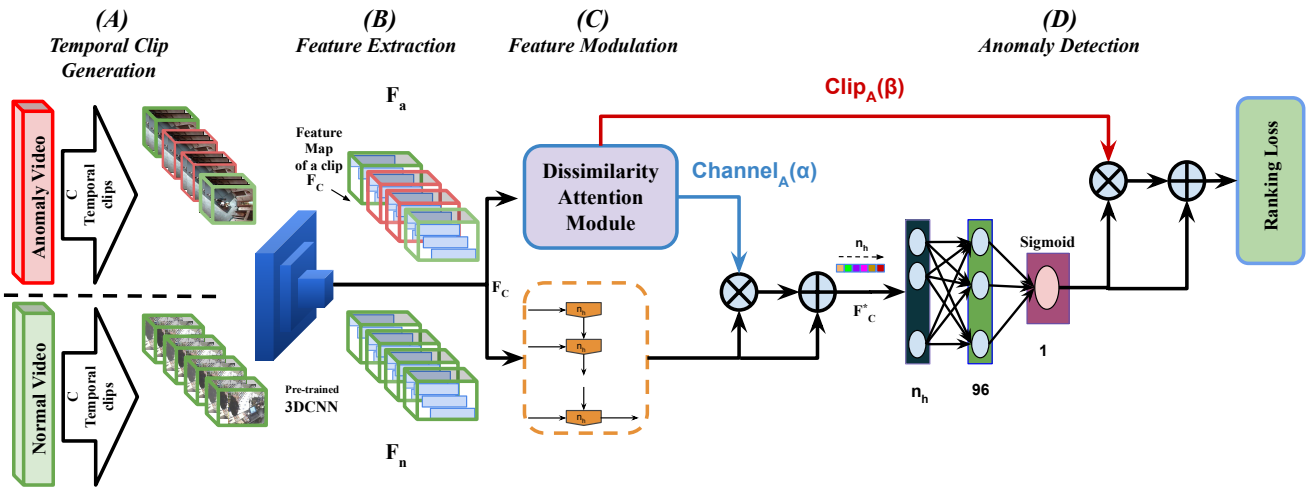


Figure 1: **Proposed Anomaly Detection Framework:** It comprises of four stages. (A) divides an untrimmed video into a fixed number of Temporal clips (*say C*). Subsequently (B) extracts the spatio-temporal feature map for each clip from a pre-trained 3D ConvNet. The main contribution of the proposed framework lies in (C) which modulates the features extracted from the pre-trained 3D ConvNet as well as computes temporal attention weights for each clip through a **Dissimilarity Attention Module (DAM)**. Finally, (D) performs anomaly detection using fully connected networks. The whole framework is optimized using a ranking loss function which takes the detection score from (D) and temporal attention weights from DAM into account for gradient computation.

the anomaly instances at the feature level but also at the score level. This discriminative learning is enabled by the DAM which learns the dissimilarity present in consecutive instances and highlights the temporal salience both at feature and score levels.

3. Proposed Method

The overview of proposed framework is shown in Figure 1 and when the four stages are executed sequentially, it achieves the weakly-supervised anomaly detection task. It can be visualized that, the framework can operate by using RGB videos with only video-level labels to detect anomalies. In addition, Figure 1 presents the training pipeline of the proposed framework, where both normal and anomaly video sequences are taken simultaneously to learn an optimal separation between the classes through a ranking loss function. Whereas, while testing, the framework can take a stream of frames instead of a whole video for real-time anomaly detection. A detailed description of each stage present in the framework is given in the following subsections.

3.1. Temporal Clip Generation

Following earlier MIL based anomaly detection methods [11, 17], the proposed framework divides a long untrimmed video into a fixed number of non-overlapping temporal clips (*say C*) in this stage. The temporal clips obtained from a normal video contains no anomaly clips, however a combination of normal and anomaly clips are obtained from an untrimmed anomaly video. The objective of the temporal clip generation stage is to ensure homogeneity in terms of the number of instances (*i.e.* clips) among the classes.

3.2. Feature Extraction

In *Feature Extraction* stage for each temporal clip, spatio-temporal features are extracted from a pre-trained 3D ConvNet backbone. The 3D ConvNet used in this framework takes a 64 frame snippet (S_i) into account to extract a feature map of dimension $t \times n$, where t denotes the temporal scale and n is the channel size. Since multiple 64-frame snippets $\{S_1, S_2, \ldots, S_m\}$ can be present inside a temporal clip (C_i) and a global feature map per C_i is required, a *max pooling* operation is performed over m. So for a given untrimmed video containing C clips, a spatio-temporal feature map of dimension $C \times t \times n$ is generated from this stage.

3.3. Feature Modulation

The objective of the *feature modulation* stage is to learn a discriminative representation for anomaly and normal instances by enhancing the quality of the feature map generated from the previous stage. This is achieved by two components as shown in Figure 1, Long-Short-Term-Memory (LSTM) and Dissimilarity

Attention Module (DAM). The detailed functionalities of each component is presented below.

Long-Short-Term-Memory (LSTM) With an objective of clip-level temporal contextual learning that can distinguish an anomaly clip from the normal ones, a vanilla *many-to-one* LSTM module is used on top of the clip-level feature map obtained from 3D ConvNet. The LSTM $f()$ having θ_h parameters takes the feature map of a clip $F_C \in \mathbb{R}^{t \times n}$ and outputs a n_h dimensional temporally encoded feature vector $F_C^{'}$ at the final time step. The LSTM has n_h number of hidden neurons and the output encoding is *tanh* squashed.

Dissimilarity Attention Module (DAM) In order to learn a discriminative representation for the normal and anomaly instances (*i.e. clips*) in long untrimmed videos, a Dissimilarity Attention Module (DAM) is proposed in this work. The DAM has two branches (*i.e.* $Channel_A$ *and* $Clip_A$) as shown in Figure 2 which perform two major functionalities. Firstly, the $Channel_A$ branch highlights the salient channels in the clip-level encoded feature vector $F_C^{'}$ obtained from the LSTM network. Secondly, $Clip_A$ branch highlights the salient temporal clips in an untrimmed video that contains anomaly patterns and hence, it guides the ranking loss function for effective optimization between normal and anomaly classes.

Unlike [7, 17], where optical flow and social force modalities are used for attention map generation, DAM uses only clip-level RGB feature map obtained from 3D ConvNet to compute $Channel_A$ and $Clip_A$. In addition, DAM neither requires the global temporal information of a video, rather, it uses the local temporal dissimilarity among two consecutive clips for computing the $Channel_A$ and $Clip_A$. Depending upon the dissimilarity measures used to generate $Channel_A$ and $Clip_A$, two variants of DAM are proposed *i.e. Dist-DAM and Cov-DAM*. In *Dist-DAM*, Manhattan distance measure is used as a dissimilarity measure and likewise, the diagonal elements of the cross-covariance matrix is used as a dissimilarity measure in *Cov-DAM* . This mechanism of generating attention map from local contextual information enables the DAM module to perform on online scenarios.

It can be visualized from Figure 2 that for computing the $Channel_A$ and $Clip_A$ for C_i clip, the feature map of C_i and C_{i-1} having dimension $t \times n$ each is required. At first, to capture the change in the distribution of channels for each t_j where $j \in [1, t]$ between C_i and C_{i-1}, a dissimilarity measure (D_n) is computed across n as presented in 1.

$$D_n = \begin{cases} \sum_{j=1}^{t} |(C_i)_{j,k} - (C_{i-1})_{j,k}|, & \text{if Dist-DAM} \\ diag([(C_i)_k - \overline{(C_i)_k}] & \\ [(C_{i-1})_k - \overline{(C_{i-1})_k}]^T), & \text{if Cov-DAM} \end{cases} \quad (1)$$

where $\forall k \in [1, n]$. Similarly, to capture the change in the distribution of temporal scale for each n_j where $j \in [1, n]$ between C_i and C_{i-1}, a dissimilarity measure (D_t) is computed across t as presented in 2.

$$D_t = \begin{cases} \sum_{j=1}^{n} |(C_i)_{j,k} - (C_{i-1})_{j,k}|, & \text{if Dist-DAM} \\ diag([(C_i)_k - \overline{(C_i)_k}] & \\ [(C_{i-1})_k - \overline{(C_{i-1})_k}]^T), & \text{if Cov-DAM} \end{cases} \quad (2)$$

where $\forall k \in [1, t]$. The dissociated dissimilarity computation across channel and temporal scales allows the DAM module to capture those channels and temporal scales which significantly change between two consecutive clips (*i.e.* C_i *and* C_{i-1}). The next step is to bind the D_n and D_t in a single dissimilarity feature map (D_{nt}) by performing Hadamard Product of D_n and D_t. Since, the dimensions of D_n and D_t are $1 \times n$ and $1 \times t$ respectively, D_n and D_t are inflated across temporal scale (t) and channel (n) to perform the Hadamard Product as shown in 3.

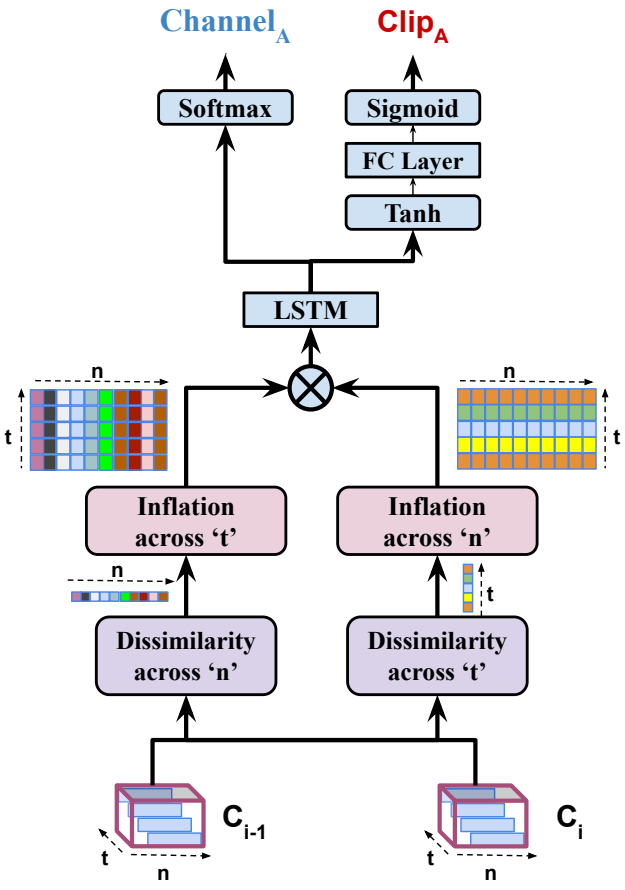


Figure 2: **Dissimilarity Attention Module(DAM)** highlights the salient channels ($Channel_A$) in the feature map and computes the temporal attention weights ($Clip_A$) for a clip (C_i), given the feature map of C_i and C_{i-1} as input.

$$D_{nt} = Inflate(D_n) \otimes Inflate(D_t) \quad (3)$$

The dissimilarity feature map (D_{nt}) is then passed on to a *many-to-one* vanilla LSTM network with n_h hidden neurons to encode those channels which significantly change over the temporal scale. The output of LSTM network is further normalized with *softmax* activation to generate the channel attention maps ($Channel_A$). However, for computing the clip attention weights ($Clip_A$), the LSTM output map is first *tanh* squashed and then projected into a fully connected (FC) layer having 1 neuron with *sigmoid* activation. The usage of *sigmoid* activation in $Clip_A$, makes the clip attention map mutually exclusive of other clip attention maps. The functional representation for computing the $Channel_A$ and $Clip_A$ is given in 4 and 5 respectively.

$$Channel_A = Softmax(LSTM(D_{nt})) \tag{4}$$

$$Clip_A = Sigmoid(FC(tanh(LSTM(D_{nt})))) \tag{5}$$

Finally, to obtain a modulated feature vector (F_C^*) per clip, first a channel attention mask is computed by performing Hardamad Product between the channel attention weights ($Channel_A$) and LSTM network output ($F_C^{'}$). Subsequently the channel attention mask is added with ($F_C^{'}$) to compute (F_C^*).

3.4. Anomaly Detection

The *anomaly detection* stage detects the temporal clips that contain anomaly patterns using a multi-layer perceptron (MLP) as shown in Figure 1. The MLP takes the modulated features of a clip (F_C^*) as input and assigns independent detection scores (D) in the final layer through a *sigmoid* activation.

Optimization of the Framework: The proposed framework along with the DAM module is end-to-end trainable excluding the 3DConvNet feature extractor. During training, following earlier method [11], a MIL based ranking loss function is adopted for optimization. The ranking loss operates between two bags of instances, namely D_a and D_n, where D_a and D_n is a collection of detection scores (D) corresponding to the temporal clips obtained from anomaly and normal video sequences respectively. Furthermore, to maximize the margin of separation among the classes, identifying the anomaly and normal instances in D_a and D_n is a crucial task. Since with video-level labels it is difficult to categorize the normal and anomaly instances, authors in [11] select the maximum scoring instances of D_a and D_n as anomaly and normal and optimized the separation among them as shown in 6.

$$R_L(D_a, D_n) = \max(0, 1 - \max_{i \in D_a}(D_a^i) + \max_{i \in D_n}(D_n^i)) \tag{6}$$

However, a maximum detection score instance of D_a may not persist anomaly in all types of abnormal scene and this inappropriate selection of anomaly instances can lead to non-maximal margin of separation.

To address this, we take the clip attention weights ($Clip_A$) obtained from DAM into account to identify the normal and anomaly instances effectively in D_n and D_a for optimization. Since, $Clip_A$ is generated by learning the dissimilarity between two consecutive clips (*i.e.* C_i *and* C_{i-1}), an abrupt change in spatio-temporal space can associate higher clip attention weights. Now the $Clip_A$ are modulated with D obtained from the final layer of MLP to compute the weighted detection scores (wD) given by $wD = (D \otimes Clip_A) \oplus D$. These wD are then used in 6 to formulate weighted ranking loss (wR_L) as shown in 7. Further more, the wR_L contains a temporal smoothing and a sparsity constraint as proposed by [11].

$$\begin{aligned} wR_L(wD_a, wD_n) = \max(0, 1 - \max_{i \in wD_a}(wD_a^i) + \max_{i \in wD_n}(wD_n^i)) \\ +\lambda_1 \sum_{i}^{(N-1)} (wD_a^i - wD_a^{i+1})^2 + \lambda_2 \sum_{i}^{N} (wD_a^i) \end{aligned} \tag{7}$$

where N = T $\times$ batchsize, λ_1 and λ_2 are the weighting factors of the temporal smoothing and sparsity constraints respectively. This weighted ranking loss (wR_L) is employed in the proposed method to maximize the margin of separation between normal and anomaly instances.

4. Experimental Analysis

Datasets - The experiments are conducted on two widely used anomaly detection datasets, namely, UCF-Crime[11] and ShanghaiTech [8]. **UCF-Crime** is a diverse and large-scale dataset containing 1900 real-world surveillance videos from 13 types of anomaly activities. In this dataset anomaly activities may occur for longer duration or for shorter duration, which makes the problem of detection more challenging. It has 1610 videos for training out of which 810 and 800 videos belongs to anomaly and normal classes respectively. Similarly, for testing there are 290 videos containing 140 anomaly and 150 normal videos. **ShanghaiTech** is a medium scale dataset recorded in a University campus. Originally [8] was designed for unsupervised anomaly detection task, but we follow a recent train-test protocol designed by Zhong *et.al.* [16] for weakly-supervised settings. This contains, 175 normal and 65 anomaly videos for training as well as 155 normal and 44 anomaly videos for testing.

Evaluation Metric - Following earlier approaches [11, 16], frame-level Receiver Operating Characteristics (ROC) and its corresponding Area Under the Curve (AUC) is used to evaluate the anomaly detection performance. In addition, a false alarm rate(FAR) measure introduced by Sultani *et.al.* [11] is also used to evaluate the robustness of the proposed method.

Table 1: Sequential Ablation Studies of the proposed method in UCF-Crime and ShanghaiTech dataset to quantify the importance of different modules in terms of AUC(%)

Method	Components	UCF-Crime	ShanghaiTech
l_1	I3D	77.42	76.19
l_2	l_1 + LSTM	79.55	77.96
Dist-DAM	l_2 + Channel_A	81.49	80.1
	l_2 + Channel_A + Clip_A	**82.57**	**88.86**
Cov-DAM	l_2 + Channel_A	81.51	80.32
	l_2 + Channel_A + Clip_A	**82.67**	**88.22**

Table 2: Anomaly detection performance comparisons of the proposed method with state-of-the-art online and offline methods in terms of frame-level AUC and FAR on UCF-Crime and ShanghaiTech dataset.

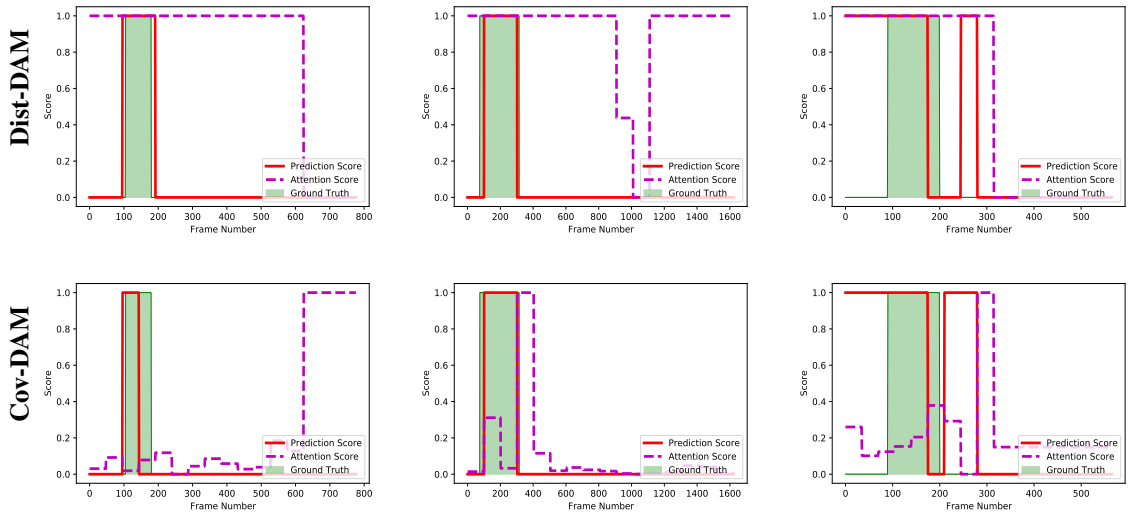
Mode of Operation	Methods	Feature	UCF-Crime		ShanghaiTech	
			AUC(%)	FAR (%)	AUC(%)	FAR (%)
Offline	Zhong *et al.*[16]	TSN ^{RGB}	82.12	0.1	84.44	-
	Majhi *et al.*[10]	I3D ^{RGB}	82.12	-	-	-
	Wu *et al.*[13]	I3D ^{RGB}	82.44	-	-	-
Online	SVM Baseline	-	50	-	-	-
	Hasan *et al.*[5]	AE ^{RGB}	50.6	27.2	-	-
	Lu *et al.*[9]	C3D ^{RGB}	65.51	3.1	-	-
	Sultani *et al.*[11]	C3D ^{RGB}	75.41	1.9	-	-
		I3D ^{RGB}	77.42	1.4	-	-
	Zhang *et al.*[15]	C3D^{RGB}	78.66	-	-	-
	Lin *et al.*[7]	C3D^{RGB}	78.28	-	-	-
	Zhu *et al.*[17]	C3D ^{RGB}	79	-	-	-
	Zaheer *et al.*[14]	C3D ^{RGB}	79.54	-	84.16	-
	Wu *et al.*[13]	I3D^{RGB}	-	-	85.38	-
	Zhong *et al.*[16]	C3D ^{RGB}	81.08	2.8	76.44	-
	Proposed Method w **Dist-DAM**	I3D ^{RGB}	**82.57**	**0.3**	**88.86**	**2.1**
	Proposed Method w **Cov-DAM**	I3D ^{RGB}	**82.67**	**0.3**	**88.22**	**2.3**

4.1. Implementation Details

Training-At first spatio-temporal features are extracted from the I3D backbone [2] for each temporal clips obtained from a long untrimmed video. The number of temporal clips (C) is set to 16 in these experiments. The input to I3D is the center cropped image of dimension 224 $\times$ 224 from the full frame for features extraction. The spatio-temporal features are extracted for from the `Global Average Pooling` layer of I3D which yields a feature map of dimension $t \times n$, where t = 7 and n = 1024. The number of hidden neurons used in *many-to-one* vanilla LSTM is set to 1024 ($= n_h$). Then the proposed framework is trained using Adam optimizer at a learning rate 0.0001 and with the loss weighting factors $\lambda_1 = \lambda_2 = 8 \times 10^{-5}$. We also randomly select 10 anomaly and 10 normal videos as a mini-batch and compute the gradient using reverse mode automatic differentiation on computation graph using Tensorflow. Then the loss is computed and back-propagated for the whole batch. **Testing-** We empirically found that center crop and the four corner crop is suitable for model testing and results in superior performance. This is to cover the spatial fine detail of the anomaly, as in [16].

4.2. Ablation Study

A detailed ablation study in UCF-Crime and ShanghaiTech datasets is given in Table 1 to observe the importance of each modules present in the proposed framework. At first, experiment is carried out with I3D feature followed by 3-layer MLP anomaly detection module to define the baseline condition of the proposed framework. Subsequently to learn the intra-clip temporal dependency in I3D feature map, a *many-to-one* LSTM module is added and it is found to improve around 2% detection performance in both the datasets shown in l_2 of Table 1. Then the major component of the proposed method, DAM is invoked to the framework for verifying the influence of Channel_A and Clip_A in anomaly detection



(Explosion27) (Shooting8) (Road Accidents128)

Figure 3: Visualization of Clip_A score along with the prediction score with respect to the ground truth during testing. The Row-1 and Row-2 shows a comparison of Clip_A and the prediction scores when Dist-DAM and Cov-DAM modules are invoked separately in the proposed framework. The comparision is made for three videos of UCF-Crime dataset, namely, "Explosion27", "Shooting8", and "Road Accidents128".

performance. Since we propose two variants of DAM, *i.e.* *Dist-DAM* and *Cov-DAM* in this paper, firstly *Dist-DAM* is considered for experimentation and it is evident that when both Channel_A and Clip_A are added to l_2 a substantial performance improvement of 10.9% and 3.02% is achieved in ShanghaiTech and UCF-Crime datasets respectively. Likewise, when *Cov-DAM* is added to the framework in place of *Dist-DAM*, a huge performance improvement of 3.12% and 10.9% is achieved in UCF-Crime and ShanghaiTech datasets respectively than that of l_2 in Table 1.

For qualitative analysis, we visualize the Clip_A weights along with the prediction scores of different videos and showcase a comparison between *Dist-DAM* and *Cov-DAM* in Figure 3. Surprisingly, we found that the Clip_A weights learned by *Dist-DAM* are more sensitive to the post anomaly scenarios by producing higher weights due to a panic like situation than that of *Cov-DAM*. However, both *Dist-DAM* and *Cov-DAM* based frameworks are capable of detecting anomalies precisely in many videos as shown for "Explosion27" for "Shooting8". In addition, there are also few special cases like "Road Accidents128" video where the scene changes abruptly and the post anomaly scenarios are very similar to a abnormal situation where the framework containing either *Dist-DAM* or *Cov-DAM* produces false positives.

4.3. State-of-the-art Comparison

To check the robustness of the proposed framework, a comparison is made with the recently reported state-of-the-art methods in UCF-Crime and ShanghaiTech datasets as shown in Table 2 and Figure 4. For UCF-Crime dataset, anomaly detection performance comparison is made upon three indicators *i.e.* ROC, AUC and FAR. However, for shanghaiTech dataset only two indicators *i.e.* ROC and AUC is used for the performance comparison due to the unavailability of protocol for computing FAR. It can be seen from Table 2, the proposed framework containing either *Dist-DAM* or *Cov-DAM* outperforms the state-of-the-art online anomaly detection methods by a larger margin in both UCF-Crime and ShanghaiTech dataset and also produce significantly smaller false positive than that of online methods in UCF-Crime dataset. It is majorly due to the capabilities of the DAM in discriminating the normal and anomaly instances in real-world scenarios by considering the dissimilarity present in consecutive instances. In addition, to detect anomalies in real-time scenario, we compare speed of the proposed framework with earlier methods in terms of FPS and it is evident from Table 3, the proposed framework invoking either *Dist-DAM* or *Cov-DAM* is much faster that Zhong *et.al.* due to the reduced number of parameters.

Table 3: Testing Speed Comparison of the proposed method with existing online anomaly detection methods.

Methods	Param	Speed(FPS)
Zhong *et al.*[16] - C3D^{RGB}	78M	130
Proposed- Dist-DAM + I3D^{RGB}	**29M**	**282**
Proposed- Cov-DAM + I3D^{RGB}	**29M**	**267**

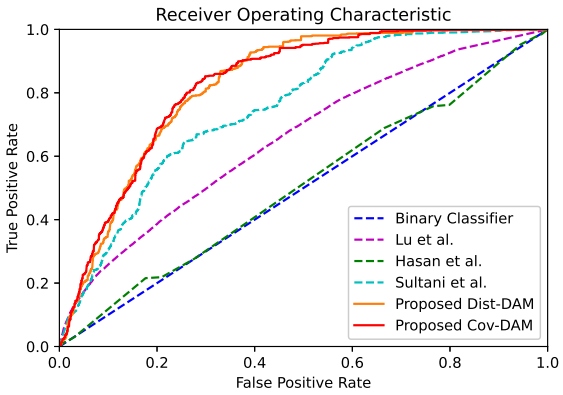


Figure 4: State-of-the-art ROC curve comparison with the proposed method.

5. Conclusion

In this work, a anomaly detection framework containing DAM is proposed to discriminate the anomaly instances from normal ones during training with video level labels. The discrimination is ensured by the Channel_A and Clip_A of DAM by taking the local contextual information into account rather than the global context, enabling the framework to detect the anomalies in *real-time* scenarios. From experimentation, it is evident that the proposed framework outperforms the state-of-the-art online anomaly detection methods in UCF-Crime and ShanghaiTech datasets. In addition, it is also considerably more faster than earlier online methods in order to meet the real-time requirements. However, the proposed framework is still affected by the post anomaly scenarios, resulting in false positives.

Acknowledgements. This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. The authors are also grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

References

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, 30(3):555–560, 2008.

[2] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[3] Y. Cong, J. Yuan, and J. Liu. Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognition*, 46(7):1851–1864, 2013.

[4] S. Dubey, A. Boragule, and M. Jeon. 3d resnet with ranking loss function for abnormal activity detection in videos. *arXiv preprint arXiv:2002.01132*, 2020.

[5] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[6] J. Kim and K. Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2928. IEEE, 2009.

[7] S. Lin, H. Yang, X. Tang, T. Shi, and L. Chen. Social mil: Interaction-aware for crowd anomaly detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019.

[8] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.

[9] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.

[10] S. Majhi, S. Das, F. Bremond, R. Dash, and P. K. Sa. Weakly-supervised joint anomaly detection and classification. *arXiv preprint arXiv:2108.08996*, 2021.

[11] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018.

[12] B. Wan, Y. Fang, X. Xia, and J. Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.

[13] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision*, pages 322–339. Springer, 2020.

[14] M. Z. Zaheer, A. Mahmood, H. Shin, and S.-I. Lee. A self-reasoning framework for anomaly detection using video-level labels. *IEEE Signal Processing Letters*, 27:1705–1709, 2020.

[15] J. Zhang, L. Qing, and J. Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4030–4034. IEEE, 2019.

[16] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[17] Y. Zhu and S. Newsam. Motion-aware feature for improved video anomaly detection. *arXiv preprint arXiv:1907.10211*, 2019.