# A Multi-Head Approach with Shuffled Segments for Weakly-Supervised Video Anomaly Detection

Salem AlMarri Muhammad Zaigham Zaheer Karthik Nandakumar
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
Abu Dhabi, UAE
{salem.almarri, zaigham.zaheer, karthik.nandakumar}@mbzuai.ac.ae

## Abstract

*Weakly-supervised video anomaly detection (WS-VAD) is a challenging task because coarse video-level annotations are insufficient to train fine-grained (segment or frame-level) detection algorithms. Multiple instance learning (MIL) powered by a ranking loss between the highest scoring segments of normal and anomaly videos has become the de-facto standard for WS-VAD. However, ranking loss is not robust to noisy segment-level labels (induced from the video-level labels), which is inherently the case in WS settings. In this work, we propose a new variant of the MIL method that utilizes a margin loss to achieve WS-VAD. The margin loss enables effective training of an anomaly scoring head based on noisy segment-level labels with high data imbalance (large number of normal segments and very few anomalous segments). We also introduce a self-supervised learning paradigm via stochastic shuffling of segments from multiple videos to mimic event changes during training. This forces the model to learn the boundaries between different virtual events (through a boundary localization head) and localizing the center of virtual events (through a center localization head). The efficacy of the proposed multi-head approach in successfully localizing anomalies is demonstrated through experiments on two large-scale VAD datasets (UCF-Crime and XD-Violence).*

## 1. Introduction

Video anomaly detection (VAD) is a well-known computer vision problem with several real-world applications including surveillance, autonomous navigation, and biomedical imaging. VAD methods detect frames in a video that do not conform to the norm, where the norm is determined by the application context. For example, events such as shoplifting or violence can be considered as anomalies in the surveillance context. Since it is prohibitively expensive to obtain fine-grained (frame-level) annotation of anomalies in videos, a weakly-supervised VAD (WS-VAD) setting that aims to learn anomalous events using only video-level binary labels [42] is commonly used. In WS-VAD, a video is labeled as *normal* if no anomalous event is present, whereas it is labeled as *anomaly* if any anomalous event is present.

Typically, WS-VAD has been tackled in the literature using a two-step approach [42, 10, 59, 64, 21, 45]. In the first step, each video is partitioned into a sequence of non-overlapping temporal segments, which are passed through a pre-trained feature extractor to obtain fixed-length representations for each segment. The second step involves the learning of an anomaly scoring model that produces a segment-level anomaly score indicating the strength of the anomaly. The most common strategy for training the anomaly scoring model in the WS-VAD setting is multiple instance learning (MIL) [42, 10]. In MIL, a normal video is considered as a negative bag containing no anomalous segments and an anomaly video is considered as a positive bag containing one or more anomalous segments. The anomaly scoring function is then typically learned by optimizing a ranking loss objective that attempts to ensure that the highest scoring segments in the positive (anomaly) bag have a larger anomaly score compared to the highest scoring segments in the negative (normal) bag [42].

While MIL is an elegant solution, the ranking loss objective commonly used in MIL has several drawbacks. Firstly, since the number of anomalous segments in a video is not known apriori, determining the number of highest scoring segments to compute the loss is cumbersome. Secondly, dependency on a few high scoring segments inherently makes the method less robust to noise in the training data, because it is well known that sample maximum/minimum is the least robust statistic [20] (i.e., maximally sensitive to outliers). Though several variants of the ranking loss have been proposed in the literature [5, 27, 44, 45, 12, 9, 21, 64, 59, 10, 26, 16], most of these heuristic methods do not eradicate the inherent limitations of ranking loss. Furthermore, since anomalous segments in a video tend to be temporally contiguous, additional heuristics are usually required to enforce the contiguity constraint and complement the ranking loss.

In this work, we aim to overcome these limitations of ranking loss by introducing a new variant of MIL for WS-VAD. Similar to existing MIL methods, our approach also utilizes a pair of one normal and one anomaly video to carry out the training. However, instead of relying on the highest scoring segments to drive the training of the anomaly scoring model, we model the distribution of segment-wise anomaly scores in an anomaly video as a mixture distribution and attempt to maximize its distance (Fisher's ratio) to the score distribution of the normal video. This results in a *margin loss*, which is known to be more robust to imbalanced and noisy data [54].

It must be emphasized that the ultimate goal of WS-VAD is precise localization of the anomaly within a given video, which in turn requires accurate detection of the boundaries (start and end) of an anomalous event. Since fine-grained annotations are not available in the WS-VAD setting, it is not possible to directly train a model that detects anomalous event boundaries. Therefore, we propose a novel *segment-level shuffling mechanism* that randomly concatenates segments from the normal-anomaly video pair to simulate "virtual" events within a video. Since these virtual events are composed of varying number of contiguous segments from a video, we hypothesize that learning to detect such virtual events during training will facilitate better detection of real anomalous events (of different lengths) in a test video. Furthermore, this approach also indirectly teaches the model that events are temporally contiguous. In this work, the same backbone encoder used for anomaly scoring is also leveraged to learn these virtual events by introducing additional boundary and center localization heads and jointly training all the three heads (anomaly score, boundary, and center localization) using appropriate loss functions.

Thus, the contributions of this paper are two-fold: 1) We propose a variant of the MIL method for WS-VAD that utilizes a margin loss in lieu of the less robust ranking loss to effectively learn the anomaly scoring model in the presence of noisy and imbalanced labels. 2) We propose a segment-level shuffling mechanism to simulate virtual events and leverage the same backbone used for anomaly scoring (but with different heads) to predict the boundaries and center of these virtual events. We empirically demonstrate that forcing the model to learn these additional pretext tasks enables it to better localize anomalous events in test videos.

## 2. Related Work

**Video Anomaly Detection**: In the computer vision context, instances that do not conform to the norm are often described as *outliers* or *anomalies* [17, 38, 37, 62, 14, 29, 60, 18, 43, 22, 41, 64, 24, 30, 40, 39, 8, 19, 52]. To train a machine learning model that detects such patterns, a large amount of data is required so that a holistic representation of normal patterns [11] can be obtained. In VAD, supervised machine learning methods require fine-grained (frame-level) data annotations [25, 1, 46]. However, such full supervision is not always feasible due to the laborious annotation process and limited availability of the anomalous data. Therefore, weakly-supervised (WS) methods are more popular because a WS model can be trained using only video-level binary labels without explicitly identifying which frames within an anomaly video are anomalous [43, 45, 56]. While several semi-supervised and unsupervised learning approaches [19, 13, 17, 38, 37, 62, 14, 29, 60] have also been proposed, this work deals with WS setting.

**MIL based Video Anomaly Detection**: Multiple instance learning (MIL) is a form of WS training that requires forming a bag of instances for each class of video [28, 6, 36]. Following Sultani *et al.* [42], Zhang *et al.* [59] further improve the MIL baseline by introducing an additional loss to train a temporal convolutional network. The inner bag loss encourages separation between the highest and lowest scoring instances within a bag. Tian *et al.* [45] further improve the rank loss by proposing a robust temporal feature magnitude learning method, through which top-k segments are selected and further separated. A binary cross-entropy loss is then used for classifying anomalous segments.

**Noisy Label Training**: Noisily labeled data can often result in overfitting, thereby hindering generalization [31, 7, 23, 35, 15]. Training with noisy labels mandates formalizing a training strategy and introducing appropriate changes to the architecture or the loss function that would prevent overfitting to noisy instances [63]. Yuan *et al.* [54] have shown that a typical square loss function is sensitive to noisy data and is not suitable when AUC maximization is the intended goal. AUC margin loss was proposed by modifying a square loss into a surrogate margin loss thus adding robustness to noisy data through a tunable margin parameter. The anomaly detection method proposed by Zhong *et al.* [61] used a graph convolutional network to clean noisy labels to enable the use of fully supervised action classifiers for WS-VAD. Zaheer *et al.* [56] proposed normalcy suppression for VAD, where the learning model is trained to suppress predictions for the noise-free normal inputs while producing high anomaly scores for anomalous features.

## 3. Proposed Method

**Problem Statement**: The WS-VAD problem can be formally stated as follows. Given a training dataset $\mathcal{D} = \{(\mathbf{V}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{V}_i$ is the $i^{th}$ video, $y_i \in \{normal \equiv 0, anomaly \equiv 1\}$ is the video-level label, and $n$ is the total number of videos in the dataset, the goal is to train an anomaly detector $\mathcal{A}$ that produces fine-grained (frame-level) anomaly predictions for a given test video. We propose an anomaly detector $\mathcal{A}$ consisting of three components - anomaly scoring, boundary localization, and center localization.

## 3.1. Anomaly Scoring Head

We assume that each video $\mathbf{V}_i$ is partitioned into a sequence of $m_i$ non-overlapping segments and these segments are passed through a feature extractor to obtain fixed-length feature representations $\{\mathbf{x}_{ij} \in \mathbb{R}^d\}_{j=1}^{m_i}$, where $d$ is the dimensionality of the feature space. Furthermore, the anomaly scoring model $\mathcal{F} : \mathbb{R}^d \rightarrow [0,1]$ produces a segment-level anomaly score $s_{ij} = \mathcal{F}(\mathbf{x}_{ij})$, where a score closer to 1 indicates higher confidence that the segment is anomalous. In most existing MIL methods, the anomaly scoring model is learned by optimizing a ranking loss objective that attempts to ensure that $max\ \{s_{\ell j}\}_{j=1}^{m_\ell} > max\ \{s_{kj}\}_{j=1}^{m_k}$ for any given normal-anomaly ($y_k = 0, y_\ell = 1$) pair of videos. In this work, the ranking loss is replaced with a margin loss.

**Rationale for Margin Loss**: Let $g_0(s)$ ($g_1(s)$) be the distribution of anomaly scores $s$ for normal (anomalous) segments. While a normal video contains only normal segments, an anomaly video may contain both normal and anomalous segments. Hence,

$$s_{ij}|[y_i = 0] \sim g_0(s) \tag{1}$$

$$s_{ij}|[y_i = 1] \sim (1-\gamma_i)g_0(s) + \gamma_i g_1(s) \equiv \widetilde{g}_1(s), \tag{2}$$

where $\gamma_i$ is the proportion of truly anomalous segments in video $\mathbf{V}_i$. Ideally, $\mathcal{F}$ should attempt to maximize the distance between the distributions $g_0(s)$ and $g_1(s)$. However, since segment-level anomaly labels are not available, we cannot estimate $g_1(s)$ directly. Hence, we attempt to maximize the distance between $g_0(s)$ and $\widetilde{g}_1(s)$. This can be achieved by maximizing the F-ratio between the two distributions, which is defined as:

$$\text{F-ratio} = \frac{(\widetilde{\mu}_1 - \mu_0)^2}{\widetilde{\sigma}_1^2 + \sigma_0^2} \tag{3}$$

where $\widetilde{\mu}_1$ and $\mu_0$ are the expected values of distributions $\widetilde{g}_1$ and $g_0$, respectively, and $\widetilde{\sigma}_1^2$ and $\sigma_0^2$ are the corresponding variances. Intuitively, maximizing the F-ratio implies minimizing the intra-class variance (lower value of denominator) and maximizing inter-class variance (higher value of numerator). Hence, maximization of F-ratio can be reformulated as minimizing the following loss function:

$$\mathcal{L}_{\text{F}} = (\sigma_1^2 + \sigma_0^2) - (\widetilde{\mu}_1 - \mu_0)^2. \tag{4}$$

Note that the above loss function has the following limitations. Firstly, it only attempts to maximize the distance between $\widetilde{\mu}_1$ and $\mu_0$, but does not ensure that $\widetilde{\mu}_1$ is greater than $\mu_0$ (i.e., anomalous segments should have higher scores). This requires minimizing $(1-(\widetilde{\mu}_1 - \mu_0))^2$, instead of maximizing $(\widetilde{\mu}_1 - \mu_0)^2$. Secondly, it has been shown in [54] that the above square loss overfits to easy anomalies and is sensitive to noisily labeled samples. Hence, for better training stability and to account for noisy labels, a tunable margin parameter $\omega$ ($0 < \omega < 1$) is required. Thus, the loss function can be reformulated as:

$$\mathcal{L}_{\text{margin}} = (\sigma_1^2 + \sigma_0^2) + (ReLU(\omega - (\widetilde{\mu}_1 - \mu_0)))^2, \tag{5}$$

where $ReLU(a) = a$, if $a \geq 0$ and 0, otherwise. Thus, when $(\widetilde{\mu}_1 - \mu_0) > \omega$, then $\mathcal{L}_{\text{margin}} = (\sigma_1^2 + \sigma_0^2)$. On the other hand, when $(\widetilde{\mu}_1 - \mu_0) \leq \omega$, then $\mathcal{L}_{\text{margin}} = (\sigma_1^2 + \sigma_0^2) + (\omega - (\widetilde{\mu}_1 - \mu_0))^2$. In contrast to the ranking loss, the above margin loss ignores easy anomalies ($(\widetilde{\mu}_1 - \mu_0) > \omega$) and focuses the model more on hard anomalies ($(\widetilde{\mu}_1 - \mu_0) \leq \omega$). Moreover, since the margin loss is based only on first and second-order statistics of the scores, it is more robust to noisy and imbalanced data compared to ranking loss.

**Practical Implementation of Anomaly Scoring Model**: We employ a neural network consisting of a shared encoder $\mathcal{E}_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^{d*}$ and a regression head $\mathcal{C}_\phi : \mathbb{R}^{d*} \rightarrow [0,1]$ for anomaly scoring, i.e., $\mathcal{F}_{\theta,\phi}(\cdot) = \mathcal{C}_\phi(\mathcal{E}_\theta(\cdot))$. Note that the encoder $\mathcal{E}_\theta$ is also shared by the boundary and center localization heads. For each video $\mathbf{V}_i$ in the training dataset, the statistics $\mu_i = \frac{1}{m_i}\sum_{j=1}^{m_i} \mathcal{F}_{\theta,\phi}(\mathbf{x}_{ij})$ and $\sigma_i^2 = \frac{1}{(m_i-1)}\sum_{j=1}^{m_i}(\mathcal{F}_{\theta,\phi}(\mathbf{x}_{ij}) - \mu_i)^2$ are computed. The margin loss between a given normal-anomaly pair of videos $(\mathbf{V}_k, \mathbf{V}_\ell)$, where ($y_k = 0, y_\ell = 1$), is computed as:

$$\mathcal{L}_{\text{margin}}(\mathbf{V}_k, \mathbf{V}_\ell) = (\sigma_\ell^2 + \sigma_k^2) + (ReLU(\omega - (\widetilde{\mu}_\ell - \mu_k)))^2. \tag{6}$$

The overall margin loss $\mathcal{L}_{\text{margin}}$ over a training batch is computed by summing up the pairwise margin losses among all the normal-anomaly pairs in the batch.

## 3.2. Segment-level Stochastic Shuffling

Since the anomaly scoring model purely focuses on individual segments of a video, it is not well-suited to determine the precise boundaries of an anomalous event. Hence, an additional mechanism is required to enable the WS-VAD model to learn boundaries between events. Given the lack of segment-level labels in the WS setting, we create virtual video sequences as follows. We define a two-state (normal and anomaly) Markov model with transition probability matrix between the two states as shown in Figure 1. Suppose that we have a normal-anomaly pair of videos $(\mathbf{V}_k, \mathbf{V}_\ell)$, where ($y_k = 0, y_\ell = 1$). Let $m_k$ and $m_\ell$ be the number of temporal segments in $\mathbf{V}_k$ and $\mathbf{V}_\ell$, respectively. Starting from the normal state, we perform a random walk for $m_* = (m_k + m_\ell)$ time steps. At each time step, we add the next segment from the video corresponding to the current

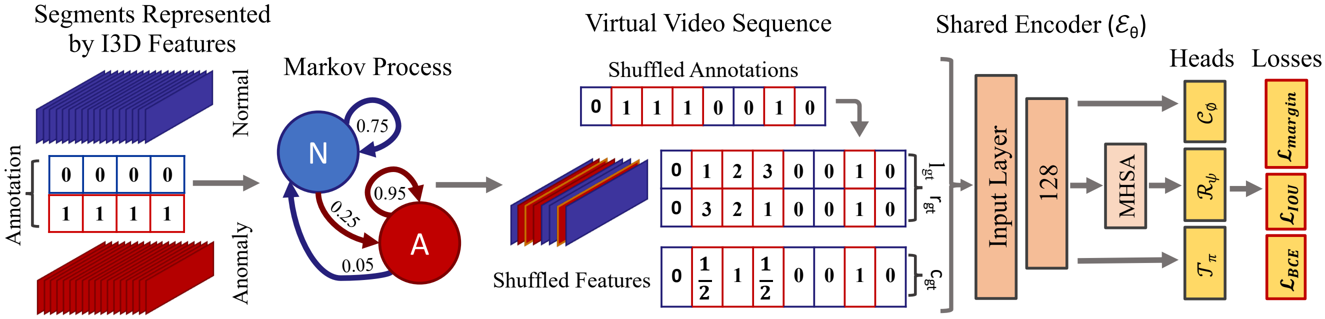

Figure 1. Our proposed multi-head approach for WS-VAD. Given a normal-anomaly pair of videos, we utilize a Markov process to shuffle the videos at the segment level to obtain a virtual video sequence. The features extracted from each segment are passed through an anomaly detector $\mathcal{A}$, which consists of a shared encoder $\mathcal{E}_\theta$ and three heads. The anomaly scoring head $\mathcal{C}_\phi$ is trained using the margin loss and directly outputs the likelihood of a segment being anomalous. The boundary localization head $\mathcal{R}_\psi$ is trained using the intersection-over-union (IOU) loss and predicts the left and right boundaries of the events in a given video sequence. Finally, the center localization head $\mathcal{T}_\pi$ is trained using the binary cross-entropy (BCE) loss and predicts the center of an event in the given sequence. During inference, the final anomaly scores are computed by taking the average of the outputs of the three heads.

state to the model. This results in virtual sequence containing $m_*$ segments containing one or more transitions from normal to anomaly video or vice versa.

A sequence of segments from the anomaly video contained within the virtual video can be considered as a "virtual" event and the ground-truth boundaries of such virtual events are known. Hence, a machine learning model can be trained to detect these virtual events. It must be emphasized that *virtual events do not correspond to real anomalies*. In fact, it is possible that none of the event boundaries in the constructed virtual video may correspond to a true transition between normal and anomaly segments. Yet, we hypothesize that virtual events simulate scene changes by mixing up segments from two different videos and learning to detect abrupt scene changes may benefit the identification of real-world anomalies. Learning to localize these virtual events also indirectly encourages the model to exploit temporal contiguity. Unshuffling a shuffled video sequence can also be considered as a self-supervised pretext task akin to the jigsaw approach in learning better representations [2].

### 3.3. Boundary Localization Head

Given a shuffled virtual video sequence, the goal of the boundary localization model is to detect the transitions from normal to anomaly video (corresponding to the left boundary of the virtual event) and from anomaly to normal video (denoting the right boundary of the virtual event). To achieve this goal, we define two vectors $\mathbf{l}_{gt}$ and $\mathbf{r}_{gt}$ of length $m_*$ representing the left and right boundary offsets for a virtual video sequence. The $i^{th}$ elements of $\mathbf{l}_{gt}$ and $\mathbf{r}_{gt}$ are set to zero if the corresponding element/segment in the virtual video sequence comes from the normal video. Otherwise, the value of the $i^{th}$ element of $\mathbf{l}_{gt}$ is set to the offset (distance) from the nearest left boundary (number of elapsed time steps since the last state transition from normal to anomaly). Similarly, the value of the $i^{th}$ element of $\mathbf{r}_{gt}$ is set to the offset (distance) from the nearest right boundary (number of time steps before the next state transition from anomaly to normal). Note that $\mathbf{t}_{gt} = (\mathbf{l}_{gt} + \mathbf{r}_{gt})$ is a vector containing the length of the virtual events at each location.

The boundary localization model consists of the same shared encoder $\mathcal{E}_\theta$ used for anomaly scoring, but is followed by a multi-head self-attention attention (MHSA) block and a multi-variate regression head (denoted together as $\mathcal{R}_\psi$) that directly attempts to predict the left and right boundary offsets. Let $\mathbf{l}_{pred}$ and $\mathbf{r}_{pred}$ be the left and right boundary offsets predicted by the boundary localization model $\mathcal{R}_\psi(\mathcal{E}_\theta(\mathbf{X}))$, where $\mathbf{X}$ is the set of feature representations of the segments contained in the virtual video sequence. Let $\mathbf{t}_{pred} = (\mathbf{l}_{pred} + \mathbf{r}_{pred})$. The accuracy of boundary predictions can be evaluated based on the intersection-over-union (IOU) value, which is computed as:

$$\begin{aligned} \text{intersection} &= \min(\mathbf{l}_{gt}, \mathbf{l}_{pred}) + \min(\mathbf{r}_{gt}, \mathbf{r}_{pred}), \\ \text{union} &= \mathbf{t}_{gt} + \mathbf{t}_{pred} - \text{intersection}, \\ \text{IOU} &= \frac{\text{intersection}}{\text{union}}. \end{aligned} \tag{7}$$

Finally, the boundary localization model is learned by minimizing $\mathcal{L}_{iou} = E[-\log(\text{IOU})]$, where $E$ denotes the expectation operator. Since IOU $\in [0,1]^{m_*}$, minimizing $\mathcal{L}_{iou}$ is equivalent to maximizing the IOU value.

### 3.4. Center Localization Head

Boundary localization is suitable for detecting state transitions in most cases. However, one limitation of the IOU loss is that if the virtual events are very short, even a small displacement may drastically reduce the IOU score. Since the virtual events are created stochastically, shorter segments are likely to exist. Therefore, to complement the boundary localization, we introduce a center localization model which learns to predict the center of a virtual event. The center labels for a video sequence is obtained as:

$$\mathbf{c}_{gt} = \frac{min(\mathbf{l}_{gt}, \mathbf{r}_{gt})}{max(\mathbf{l}_{gt}, \mathbf{r}_{gt})}. \tag{8}$$

The above center labeling approach assigns a label of 1 to the center segment within a virtual event and progressively lower values for segments of the virtual event as we move away from the center. Segments corresponding to the normal video are assigned a value of 0. The center localization model consists of the same shared encoder $\mathcal{E}_\theta$ used for anomaly scoring and boundary localization, but consists of a separate regression head $\mathcal{T}_\pi : \mathbb{R}^{d*} \rightarrow [0, 1]$ to obtain the predicted center label $\mathbf{c}_{pred} = \mathcal{T}_\pi(\mathcal{E}_\theta(\mathbf{X}))$. The center localization model is trained using the standard binary cross entropy loss function $\mathcal{L}_{bce}(\mathbf{c}_{gt}, \mathbf{c}_{pred})$.

### 3.5. Training the Multi-Head Anomaly Detector

To summarize, the proposed anomaly detector $\mathcal{A}$ consists of an encoder $\mathcal{E}_\theta$, an anomaly scoring head $\mathcal{C}_\phi$, a boundary localization head $\mathcal{R}_\psi$ (which includes a MHSA layer), and a center localization head $\mathcal{T}_\pi$. The shared encoder shown in Figure 1 consists of a fully connected layer, a ReLU activation function, and a dropout layer followed by layer normalization. While the individual heads are learned by minimizing the corresponding loss functions described earlier, the shared encoder $\mathcal{E}_\theta$ is jointly optimized by minimizing the following objective:

$$\min_{\theta} \left(\lambda_{margin}\mathcal{L}_{margin} + \lambda_{iou}\mathcal{L}_{iou} + \lambda_{bce}\mathcal{L}_{bce}\right), \tag{9}$$

where $\lambda_{margin}$, $\lambda_{iou}$, and $\lambda_{bce}$ are hyperparameters that can be adjusted to control the relative importance of the losses.

### 3.6. Inference using Multi-Head Anomaly Detector

Given a new test video $\mathbf{V}_t$ during inference, we partition it into a sequence of $m_t$ non-overlapping segments and extract features $\{\mathbf{x}_{tj}\}_{j=1}^{m_i}$ for each segment. These features are passed through the anomaly detector to obtain $\mathbf{s}_{pred}$, $\mathbf{t}_{pred}$ and $\mathbf{c}_{pred}$ from the three heads. Note that $\mathbf{s}_{pred} = [\mathcal{F}_{\theta,\phi}(\mathbf{x}_{t1}), \cdots, \mathcal{F}_{\theta,\phi}(\mathbf{x}_{tm_t})]$. The final anomaly score is computed as $(\mathbf{s}_{pred}+(\mathbf{t}_{pred}/m_t)+\mathbf{c}_{pred}))/3$. Since all the three components of the above score are already normalized between 0 and 1, with 1 being anomalous, their average gives a normalized anomaly score between 0 and 1 for each segment. If the anomaly score is greater than a threshold, the segment is categorized as anomalous and all the frames within the segment are labeled as anomalous.

Note that the above inference scoring approach is different from several existing state-of-the-art (SOTA) methods that utilize video-level normalization at inference time [3, 19]. We argue that using normalization at inference time limits the practical applicability of the scheme in real-world scenarios because it hinders the computation of anomaly score for one or several frames immediately, rather than waiting for a video to be completed. We also conduct experiments based on anomaly scoring using individual heads, details of which are discussed in the ablation section.

## 4. Experiments and Results

### 4.1. Datasets

Our proposed approach has been evaluated on two WS-VAD benchmark datasets, UCF-Crime [42] and XD-Violence [51], which include multiple classes of anomalous activities taking place in multiple scenes.

**UCF-Crime:** UCF-Crime is one of the large-scale real-world VAD datasets, comprising of 1900 videos with a total of 128 hours of life-threatening anomalous activities captured via surveillance cameras. The training set contains only video-level labeling. The dataset includes 13 anomaly categories such as assault, burglary, and robbery.

**XD-Violence:** XD-Violence is a relatively more recent VAD dataset comprising of 4754 videos with a total length of 217 hours of multi-scene footage taken from movies, sport events, hand-held cameras, CCTV, etc. The dataset also introduces audio modality to complement video features, enabling researchers to work on multi-modal approaches for anomaly detection [51, 53, 48]. Furthermore, in contrast to the limit of up to two anomalous events per test video in the UCF-Crime dataset, a test video of XD-Violence may contain more than two anomalous events.

### 4.2. Evaluation Metrics

We have utilized the well-known Area Under the receiver operating characteristic Curve (AUC) metric for evaluating our proposed approach on UCF-Crime, following several existing methods [42, 45, 5, 49, 12]. In addition, we have also utilized the mean average precision (mAP) evaluation metric for our approach on XD-Violence, following previous publications [48, 33, 32, 51]. However, it may be noted that we do not optimize our model for mAP, as we compute the mAP based on the same model optimized for AUC.

### 4.3. Implementation Details

We use I3D features [4] to train our model. The features are extracted by using RGB and flow modalities. The code is written in Pytorch. PESG optimizer [54] is used to carry out the training with learning rate of $8 \times 10^{-2}$ and weight decay of $1 \times 10^{-3}$. Margin $\omega$ in Eq. 6 is set to 1. Moreover, $\lambda_{margin}$, $\lambda_{iou}$, and $\lambda_{bce}$ in Eq. 9 are set to 1, 1, and 0.5 respectively. The state transition probability used to control the frequency of shuffling between normal to anomaly state in Section 3.2 is set to 0.05 and 0.75 for anomaly to normal

| Method | Features | Params | AUC |
|---|---|---|---|
| Sultani *et al.* [42] | C3D-RGB | 2.11M | 75.41 |
| Sultani *et al.* [42] | I3D-RGB | - | 77.92 |
| Zaheer *et al.* [58] | ResNext | 6.5M | 79.84 |
| Feng *et al.* [12] | I3D-RGB | 31M | 82.30 |
| Wu *et al.* [51] | I3D-RGB | 0.76M | 82.44 |
| Zaheer *et al.* [56] | C3D-RGB | - | 83.03 |
| Tian *et al.* [45] | I3D-RGB | 24.76M | 84.30 |
| Watanabe *et al.* [47] | I3D-RGB | 0.33M | 84.91 |
| Chen *et al.* [5] | I3D-RGB | 28.6M | **86.98** |
| **Ours \w rank loss** | I3D-(RGB+FLOW) | **0.26M** | 84.70 |
| **Ours** | I3D-(RGB+FLOW) | **0.26M** | 85.47 |

Table 1. AUC performance and number of trainable parameters comparison of our approach with state-of-the-art anomaly detection approaches on UCF crime dataset. Best and second best values are highlighted as bold and underlined.

| Method | Features | Modality | AUC | AP |
|---|---|---|---|---|
| Sultani *et al.* [42] | C3D | R | - | 73.20 |
| Tian *et al.* [45] | I3D | R | **89.30** | 77.81 |
| Wei *et al.* [48] | I3D+VGGish | R+A | - | 80.13 |
| Wu *et al.* [49] | I3D | R | - | 80.26 |
| Pang *et al.* [34] | I3D+VGGish | R+A | - | **81.69** |
| Panariello *et al.* [32] | I3D | R | 90.23 | 71.68 |
| Wu *et al.* [50] | I3D+VGGish | R+A | 89.75 | 75.90 |
| Wu *et al.* [51] | I3D+VGGish | R+A | - | 78.64 |
| **Ours** | I3D | R | 88.80 | 66.90 |
| **Ours** | I3D+VGGish | R+F | 90.60 | 70.20 |
| **Ours** | I3D+VGGish | R+A | 91.10 | 71.40 |
| **Ours** | I3D+VGGish | R+F+A | **91.53** | **75.45** |

Table 2. AUC and AP performance comparison of our approach with state-of-the-art anomaly detection approaches on XD-violence dataset. Best and second best AUCs are highlighted as bold and underlined, respectively.

state. Following the standard setting widely used in existing approaches [42, 57, 45, 55], each feature vector is computed by using 16 frames as input. Then, following [42, 45], the feature vectors are averaged to obtain 32 features per video. Each training batch includes feature vectors of 60 videos, half of which are normal and the rest of the half anomalous. Moreover, as mentioned in Section 3, the videos are shuffled in pairs. This way, in total, a single batch consists of 1920 feature vectors.

### 4.4. Quantitative Results on UCF-Crime Dataset

The AUC performance of our proposed approach on UCF-Crime is reported in Table 1 and compared against several existing SOTA WS-VAD methods. Methods proposed in [45, 12] adopt the ranking loss introduced in [42], augment additional components, and report overall performance improvements of 8.89% and 6.89% in AUC score. Other methods such as [56, 51, 58] do not particularly adopt a ranking loss function for their training but still follow the baseline of [42]. In our approach, instead of the conventional ranking loss, we utilize margin loss to drive the overall training. Our method demonstrates 85.47% AUC score, which is comparable to several existing SOTA methods. In addition, we provide a comparison between the conventional ranking loss [42] and the margin loss used in our approach by modifying our approach to train on rank loss. This results in a lower performance of 84.70, demonstrating that the margin loss is more suitable for the WS-VAD task. We also provide a comparison of the number of model parameters in Table 1. It can be observed that our approach has the smallest network size with 0.26M parameters. Notably, compared to Chen *et al.* [5], our approach demonstrates a drop of 1.51% AUC but with 24.2M lesser parameters, providing a reasonable balance between the performance and the model size.

### 4.5. Quantitative results on XD-Violence Dataset

Following existing SOTA methods, the performance of our proposed approach using mAP and AUC metrics on XD-Violence dataset are reported in Table 2. While our approach outperforms all compared method in terms of AUC by achieving 91.53%, it yields slightly lower mAP of 75.45 compared to some existing methods. However, given the smaller model size and good AUC performance, our approach is competitive.

## 5. Analysis and Discussion

We perform extensive ablation studies to evaluate the importance of different components used in our approach as well as analyze the sensitivity of our method to various hyperparameters.

### 5.1. Ablation study

A detailed ablation study is provided in Table 3 and each set of experiments is discussed below:
**On using margin loss for noisy labels**: For our anomaly scoring head ($\mathcal{C}_\phi$), we compare the performance difference between the commonly used mean squared error ($\mathcal{L}_{MSE}$) loss and our margin loss ($\mathcal{L}_{margin}$) designed specifically to handle label noise. Training only with $\mathcal{C}_\phi$, the margin loss $\mathcal{L}_{margin}$ outperforms MSE loss ($\mathcal{L}_{MSE}$) by 11.21% on the UCF-Crime dataset, which demonstrates its utility. On XD-Violence dataset, the improvement in performance using $\mathcal{L}_{margin}$ is not as noticeable. It may be attributed to the smaller length of videos in this dataset, which corresponds to less noise in the labels.
**On impact of boundary localization and center localization losses**: We evaluate the performance of our approach with and without boundary localization and center localization heads. It may be noted that using these additional heads without shuffling (by simply concatenating the segments from the normal and anomaly videos one after the other) negatively impacts the performance compared to only the anomaly scoring head. However, shuffling results in notable performance gains, which demonstrates that the stochastic shuffling mechanism goes hand in hand with our boundary

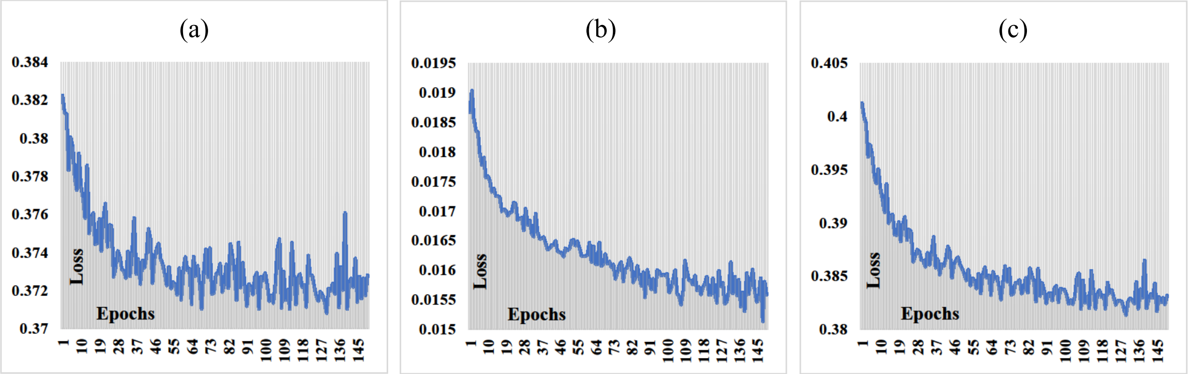

Figure 2. Training loss visualizations of our approach. (a) $\mathcal{L}_{margin}$ is based on the weak labels available for training therefore demonstrates less fluctuations. (b) $\mathcal{L}_{IOU}$ and (c) $\mathcal{L}_{BCE}$ losses are computed based on sef-supervisory signals, thus demonstrating slight fluctuations. However, all losses converge reasonably.

and center localization heads and contributes positively towards the overall performance.

### 5.2. Sensitivity to Hyperparameters

In this section, we analyze and discuss several design choices, parameter selection, and the empirical justification of other settings of our approach.

**Relative importance of the three losses**: In Table 4, we summarize the study on utilizing different weight values for the boundary and center localization losses used in the training of our approach, while fixing the value of $\lambda_{margin}$ to 1. The optimal performance is achieved when localization loss weight $\lambda_{iou}$ is set to 1 and center localization loss weight $\lambda_{bce}$ is set to 0.5. This demonstrates that our losses not only align well with the overall training objective but also play a strong role in successful training of the proposed approach.

**Impact of state transition probability**: The stochastic shuffling strategy is a vital part of our training mechanism and is controlled by the transition probability from anomaly to normal state (say $\gamma$). We fix the normal-anomaly transition probability at 0.05. When $\gamma$ is low, lesser mixing occurs and consequently lower number of virtual events created. On the other hand, a higher value of $\gamma$ will result in boundaries being created too frequently. We conduct experiments to identify the optimal $\gamma$ value and the results are summarized in Figure 4. Horizontal axis represents varying values of $(1 - \gamma)$ whereas the left vertical axis represents AUC% score on UCF-crime dataset. The right vertical axis represents the total number of virtual events generated as a result of the respective $\gamma$ values. The best performance is achieved when $\gamma$ is set to $0.25$, which corresponds to an average of 3 virtual events per shuffled sequence created using two videos.

**Impact of using videos from different classes for shuffling**: In MIL methods, it is a common practice to have one positive and one negative bag to carry out the training. Our method also carries out shuffling between a normal and an anomaly video to ensure that the labels from both classes are present during each iteration (Section 3.2). To validate the importance and correctness of this configuration, we carry out experiments using (A) 'anomaly-anomaly/normal-normal' and (B) 'no-constraint' configurations. In configuration (A), an anomaly video is always paired with another anomaly video and a normal video is always paired with another normal video. This configuration results in a noticeably lower performance of 61.77%. Configuration (B), which does not enforce any constraint, demonstrates a performance of 81.83%. However, the best performance of 85.47% is achieved when pairing is done across classes.

**Impact of multi-modality on our approach**: Although our approach is not specifically designed for multimodal training, we have carried out a set of experiments to observe the performance when trained on different modalities of XD-Violence dataset. The results of these experiments are summarized in Table 5. While RGB modality is the most important for VAD, combining flow or audio individually with RGB improves the performance. However, the best performance is achieved when all three modalities are combined.

### 5.3. Qualitative Analysis

**Anomaly Scores:** Anomaly score predictions of our approach on various videos of UCF-Crime are presented in Figure 3. It can be observed that in cases such as Stealing and Burglary, our approach produces higher anomaly scores corresponding to the anomaly ground-truth. Moreover, in the case of two normal videos, our approach correctly produces significantly lower anomaly scores. A slight failure case can be observed in *Stealing079*, where the model keeps predicting high anomaly scores after the stealing event is over. Careful analysis of this video shows that the stolen item was left behind by the thief and a commotion occurred at the scene. Although this was labeled as normal, it is clearly not normal and this explains the higher anomaly scores produced by our model.

| $\mathcal{L}_{MSE}$ | $\mathcal{L}_{margin}$ | $\mathcal{L}_{IOU} + \mathcal{L}_{BCE}$ | Shuffle ($\gamma$) | AUC (%) - UCF | AUC (%) - XD | AP (%) - XD |
|---|---|---|---|---|---|---|
| ✓ | × | × | × | 72.24 | 91.14 | 74.16 |
| ✓ | × | × | ✓ | 71.51 | 91.10 | 74.66 |
| ✓ | × | ✓ | × | 82.79 | 90.46 | 71.03 |
| ✓ | × | ✓ | ✓ | 83.76 | 91.21 | 73.08 |
| × | ✓ | × | × | 83.45 | 91.51 | 71.37 |
| × | ✓ | × | ✓ | 82.96 | 91.44 | 72.03 |
| × | ✓ | ✓ | × | 81.36 | 89.59 | 73.49 |
| × | ✓ | ✓ | ✓ | **85.47** | **91.53** | **75.45** |

Table 3. Ablation studies of our method based on multiple training configurations. Different components of our approach are removed to observe the performance gains.

| $\lambda_{margin}$ | $\lambda_{IOU}$ | $\lambda_{BCE}$ | AUC% |
|---|---|---|---|
| 1 | 1 | 1 | 84.90 |
| 1 | 0.5 | 1 | 84.80 |
| 1 | 1 | 0.5 | 85.47 |
| 1 | 0.5 | 0.5 | 81.00 |
| 1 | 0.1 | 0.1 | 81.60 |
| 1 | 0 | 0 | 82.96 |

Table 4. Weight balancing different values of losses. As seen, too high or too low values result in degraded performances.

|  | **Feature Dimension** | **AP** | **AUC** |
|---|---|---|---|
| RGB+Flow+Audio | 2048 | **75.5** | **91.5** |
| RGB | 1024 | 66.9 | 88.8 |
| RGB+Flow | 2048 | 70.2 | 90.6 |
| RGB+Audio | 1152 | 71.4 | 91.1 |
| Flow+Audio | 1152 | 66.7 | 88.7 |
| Flow | 1024 | 60.7 | 85.6 |
| Audio | 128 | 56.5 | 82.9 |

Table 5. Analysis on different modalities of XD-Violence dataset.R:RGB, F:Flow, A:Audio.

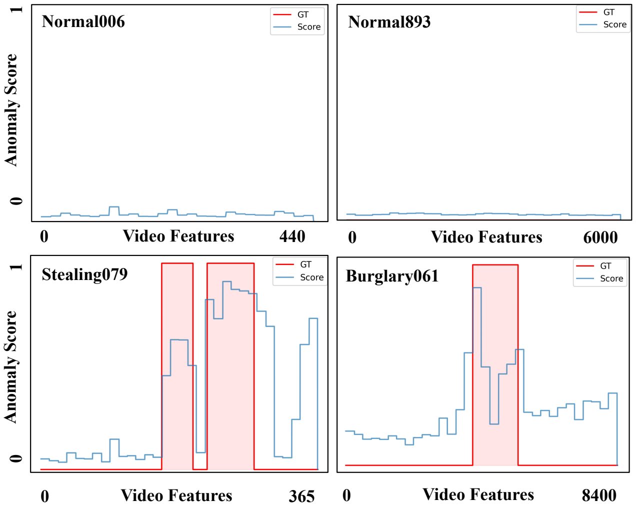

Figure 3. Test prediction plots generated by our model on different videos of UCF-Crime. Shaded areas indicate ground-truth annotation depicting anomalous segments.

**Losses and Convergence:** In Figures 2, we provide the overall training progress by plotting all three losses and the AUC of the model. It can be seen that overall model converges reasonably with the three losses gradually decreasing. The AUC performance of our approach drops initially. However, it increases afterwards and gradually peaks after about 100 epochs. The initial decrease in performance may be attributed to the pseudo-labels that we create to train the boundary localization and center localization losses.

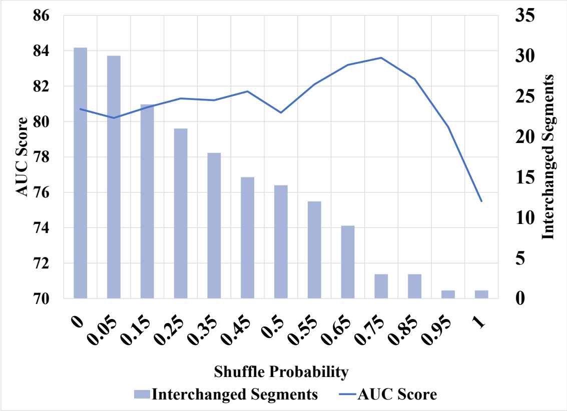

Figure 4. Impact of shuffling probability $\gamma$. Note that the x-axis is $(1-\gamma)$ and the right y-axis is the average number of virtual events obtained after shuffling.

## 6. Conclusion

In this work, we presented a new variant of MIL based weakly-supervised video anomaly detection that avoids the pitfalls of ranking loss. Apart from leveraging a margin loss to optimally train the model on noisy labels, we have proposed a self-supervised feature shuffle mechanism and introduced center and boundary localization losses to regularize the training. Despite the simplicity of the proposed approach, we have demonstrated that it works well on two large-scale video anomaly datasets.

## References

[1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. *CoRR*, abs/2111.08644, 2021.

[2] Unaiza Ahsan, Rishi Madhok, and Irfan Essa. Video jigsaw: Unsupervised learning of spatiotemporal context for video

action recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 179–189. IEEE, 2019.

[3] Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. Pseudobound: Limiting the anomaly reconstruction capability of one-class classifiers using pseudo anomalies. *Neurocomputing*, 534:147–160, 2023.

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[5] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. *arXiv preprint arXiv:2211.15098*, 2022.

[6] Veronika Cheplygina, David MJ Tax, and Marco Loog. Multiple instance learning with bag dissimilarities. *Pattern recognition*, 48(1):264–275, 2015.

[7] MIT Critical Data, Cátia M Salgado, Carlos Azevedo, Hugo Proença, and Susana M Vieira. Noise versus outliers. *Secondary analysis of electronic health records*, pages 163–183, 2016.

[8] Fernando Pereira dos Santos, Leonardo Sampaio Ferraz Ribeiro, and Moacir Antonelli Ponti. Generalization of feature embeddings transferred from different video anomaly detection domains. *CoRR*, abs/1901.09819, 2019.

[9] Shikha Dubey, Abhijeet Boragule, Jeonghwan Gwak, and Moongu Jeon. Anomalous event recognition in videos based on joint learning of motion and appearance with multiple ranking measures. *Applied Sciences*, 11(3):1344, 2021.

[10] Shikha Dubey, Abhijeet Boragule, and Moongu Jeon. 3d resnet with ranking loss function for abnormal activity detection in videos. In *2019 International Conference on Control, Automation and Information Sciences (ICCAIS)*, pages 1–6. IEEE, 2019.

[11] Ted Dunning and Ellen Friedman. Practical machine learning: A new look at anomaly detection. 2014.

[12] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14009–14018, 2021.

[13] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning, 2021.

[14] 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. *CoRR*, abs/1904.02639, 2019.

[15] Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5138–5147, 2019.

[16] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016.

[17] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 733–742, 2016.

[18] Xing Hu, Shiqiang Hu, Yingping Huang, Huanlong Zhang, and Hanbing Wu. Video anomaly detection using deep incremental slow feature analysis network. *IET Computer Vision*, 10(4):258–267, 2016.

[19] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video, 2019.

[20] Wen Jin, Anthony KH Tung, and Jiawei Han. Mining top-n local outliers in large databases. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 293–298, 2001.

[21] Ammar Mansoor Kamoona, Amirali Khodadadian Gostar, Alireza Bab-Hadiashar, and Reza Hoseinnezhad. Multiple instance-based video anomaly detection using deep temporal encoding–decoding. *Expert Systems with Applications*, 214:119079, 2023.

[22] Federico Landi, Cees G. M. Snoek, and Rita Cucchiara. Anomaly locality in video surveillance. *CoRR*, abs/1901.10364, 2019.

[23] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE international conference on computer vision*, pages 1910–1918, 2017.

[24] Shuheng Lin, Hua Yang, Xianchao Tang, Tianqi Shi, and Lin 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, 2019.

[25] Kun Liu and Huadong Ma. Exploring background-bias for anomaly detection in surveillance videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1490–1499, New York, NY, USA, 2019. Association for Computing Machinery.

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

[27] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *IEEE transactions on image processing*, 30:4505–4515, 2021.

[28] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997.

[29] Jefferson Ryan Medel and Andreas E. Savakis. Anomaly detection in video using predictive convolutional long short-term memory networks. *CoRR*, abs/1612.00390, 2016.

[30] Medhini Narasimhan and S. SowmyaKamath. Dynamic video anomaly detection and localization using sparse denoising autoencoders. *Multimedia Tools and Applications*, 77:13173–13195, 2017.

[31] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

[32] Aniello Panariello, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Consistency-based self-supervised learning for temporal anomaly localization. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 338–349. Springer, 2023.

[33] Wenfeng Pang, Wei Xie, Qianhua He, Yanxiong Li, and Jichen Yang. Audiovisual dependency attention for violence detection in videos. *IEEE Transactions on Multimedia*, 2022.

[34] Wen-Feng Pang, Qian-Hua He, Yong-jian Hu, and Yan-Xiong Li. Violence detection in videos based on fusing visual and audio information. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2260–2264. IEEE, 2021.

[35] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017.

[36] Gwenolé Quellec, Guy Cazuguel, Béatrice Cochener, and Mathieu Lamard. Multiple-instance learning for medical image and video analysis. *IEEE reviews in biomedical engineering*, 10:213–234, 2017.

[37] Manassés Ribeiro, André Eugênio Lazzaretti, and Heitor Silvério Lopes. A study of deep convolutional auto-encoders for anomaly detection in videos. *Pattern Recognition Letters*, 105:13–22, 2018. Machine Learning and Applications in Artificial Intelligence.

[38] M. Sabokrou, M. Fathy, and M. Hoseini. Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. *Electronics Letters*, 52(13):1122–1124, 2016.

[39] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, and Reinhard Klette. Fully convolutional neural network for fast anomaly detection in crowded scenes. *CoRR*, abs/1609.00866, 2016.

[40] M. Sabokrou, Mohsen Fayyaz, Mahmood Fathy, and Reinhard Klette. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing*, 26:1992–2004, 2017.

[41] Behnam Sabzalian, Hossein Marvi, and Alireza Ahmadyfard. Deep and sparse features for anomaly detection and localization in video. In *2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, pages 173–178, 2019.

[42] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.

[43] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. *CoRR*, abs/1801.04264, 2018.

[44] Kamalakar Vijay Thakare, Nitin Sharma, Debi Prosad Dogra, Heeseung Choi, and Ig-Jae Kim. A multi-stream deep neural network with late fuzzy fusion for real-world anomaly detection. *Expert Systems with Applications*, 201:117030, 2022.

[45] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4975–4986, October 2021.

[46] Boyang Wan, Wenhui Jiang, Yuming Fang, Zhiyuan Luo, and Guanqun Ding. Anomaly detection in video sequences: A benchmark and computational model. *CoRR*, abs/2106.08570, 2021.

[47] Yudai Watanabe, Makoto Okabe, Yasunori Harada, and Naoji Kashima. Real-world video anomaly detection by extracting salient features in videos. *IEEE Access*, 10:125052–125060, 2022.

[48] Dong-Lai Wei, Chen-Geng Liu, Yang Liu, Jing Liu, Xiao-Guang Zhu, and Xin-Hua Zeng. Look, listen and pay more attention: Fusing multi-modal information for video violence detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1980–1984. IEEE, 2022.

[49] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 729–745. Springer, 2022.

[50] Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, 30:3513–3527, 2021.

[51] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 322–339. Springer, 2020.

[52] Ke Xu, Tanfeng Sun, and Xinghao Jiang. Video anomaly detection and localization based on an adaptive intra-frame classification network. *IEEE Transactions on Multimedia*, 22(2):394–406, 2020.

[53] Jiashuo Yu, Jinyu Liu, Ying Cheng, Rui Feng, and Yuejie Zhang. Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection, 2022.

[54] Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3040–3049, 2021.

[55] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, Arif Mahmood, and Seung-Ik Lee. Cleaning label noise with

clusters for minimally supervised anomaly detection. *arXiv preprint arXiv:2104.14770*, 2021.

[56] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 358–376. Springer, 2020.

[57] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Clustering aided weakly supervised training to detect anomalous events in surveillance videos. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[58] Muhammad Zaigham Zaheer, Arif Mahmood, Muhammad Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection, 2022.

[59] Jiangong Zhang, Laiyun Qing, and Jun 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.

[60] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 1933–1941, New York, NY, USA, 2017. Association for Computing Machinery.

[61] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1237–1246, 2019.

[62] Joey Tianyi Zhou, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, and Rick Siow Mong Goh. Anomalynet: An anomaly detection network for video surveillance. *IEEE Transactions on Information Forensics and Security*, 14(10):2537–2550, 2019.

[63] Wencheng Zhu, Jiwen Lu, Jiahao Li, and Jie Zhou. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30:948–962, 2021.

[64] Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *arXiv preprint arXiv:1907.10211*, 2019.