OE-CTST: Outlier-Embedded Cross Temporal Scale Transformer for Weakly-supervised Video Anomaly Detection

Snehashis Majhi[1,2], Rui Dai[1,2], Quan Kong[3], Lorenzo Garattoni[4], Gianpiero Francesca[4], François Brémond[1,2]

[1] INRIA [2] Côte d'Azur University [3] Woven by Toyota [4] Toyota Motor Europe

Abstract

Video anomaly detection in real-world scenarios is challenging due to the complex temporal blending of long and short-length anomalies with normal ones. Further, it is more difficult to detect those due to : (i) Distinctive features characterizing the short and long anomalies with sharp and progressive temporal cues respectively; (ii) Lack of precise temporal information (i.e. weak-supervision) limits the temporal dynamics modeling of anomalies from normal events. In this paper, we propose a novel 'temporal transformer' framework for weakly-supervised anomaly detection: OE-CTST[†]. The proposed framework has two major components: (i) Outlier Embedder (OE) and (ii) Cross Temporal Scale Transformer (CTST). First, OE generates anomaly-aware temporal position encoding to allow the transformer to effectively model the temporal dynamics among the anomalies and normal events. Second, CTST encodes the cross-correlation between multi-temporal scale features to benefit short and long length anomalies by modeling the global temporal relations. The proposed OE-CTST is validated on three publicly available datasets i.e. UCF-Crime, XD-Violence, and IITB-Corridor, outperforming recently reported state-of-the-art approaches.

1. Introduction

Anomaly detection in real-world untrimmed videos is a prominent and active computer vision task, thanks to its inherent applications in smart surveillance systems empowering timely anomaly prevention and investigation. Recently, video anomaly detection has become a demanding task to detect complex and diversified categories of real-world anomalies. Further, the inability to temporally annotate a large amount of untrimmed videos through human effort makes the task more challenging. There exist public datasets [19, 33] which depict the above challenges. In order to address this, recent popular methods [33, 39, 45] adopt a weakly-supervised paradigm which enables superior generalization capabilities than unsupervised uni-class methods [1, 7] to detect real-world video anomalies.

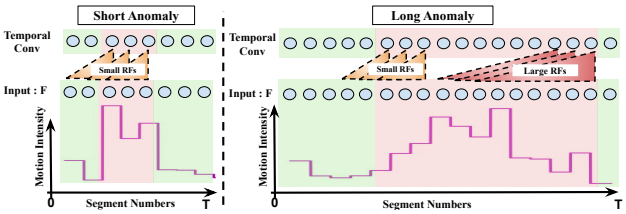

Figure 1. Visualization of distinctive motion cues in short and long-length anomalies. Short anomalies (*left*) e.g. explosions, road accidents contain sharp changes that can easily be captured by temporal convolution with smaller receptive fields (RFs), but long anomalies (*right*) e.g. robbery, shoplifting has progressive changes that need relation modeling with smaller and larger RFs to capture the start and continuity of abnormality respectively.

Despite the prosperity in mainstream weakly-supervised video anomaly detection (WSVAD) approaches [33, 39, 45, 47], their performance is still limited due to the difficulties in detecting both short and long-length anomalies w.r.t. the normal counterparts. Here, by short and long-length anomalies, we mean ones that last below and above a threshold duration (*typically 2 seconds*). From initial analysis, we found that in UCF-Crime [33] (*i.e. a popular dataset in close proximity to real-world scenarios*), out of 128 hrs of videos nearly 33.4 hrs and 17.8 hrs of untrimmed anomaly videos contain long and short length anomaly instances respectively. Further, in real-world, short and long-length anomalies are characterized by divergent temporal cues: *i.e.* short anomalies have distinctive appearance and sharp motion cues which are relatively easy to detect, whereas long anomalies are characterized by subtle and progressive motion cues similar to many normal scenarios as shown in Figure 1. For this, it is more difficult to detect long anomalies than short ones and it requires robust methods to handle them.

To address the above challenges, previous popular methods [33, 39] adopt conventional temporal modeling networks like TCN [13], LSTM [24] to discriminate short anomalies from normal events. Since TCNs are based on 1D Convolutional kernel, they can mostly capture the sharp changes among the temporal neighborhood segments (*i.e.*

[†]Code & Models: https://github.com/snehashismajhi/OECTST

a set of consecutive frames) and not the temporally distant ones. Thus, such methods fail to detect long anomalies as they do not capture the long-range temporal dependencies. With the recent success of transformers in computer vision which are empowered by multi-head self-attention [9], many popular methods in fully-supervised action detection [17, 40] and classification have leveraged temporal transformers for effective global temporal relation encoding. However, weakly-supervised anomaly detection tasks can not get direct benefits from the current temporal transformers, due to **(i) conventional positional encoding:** unlike fully-supervised settings, where the temporal positions have a one-to-one correspondence with the anomaly instances for superior global temporal relation encoding, the weakly-supervised methods do not have such correspondences due to the unavailability of the instance labels; **(ii) naive tokenization scheme:** existing methods follow a fixed-scale tokenization scheme regardless of the action duration, as a result these methods can not accumulate local contextual information for long anomalies. Therefore, we argue that a distinctive design of temporal transformers is necessary for weakly-supervised settings.

To this end, we propose a novel transformer framework that comprises an outlier embedder (OE) and a cross-temporal scale transformer (CTST). Unlike conventional position embedding, the proposed outlier embedder generates anomaly-aware temporal position encoding which enables the transformer to better encode global temporal relations among the normal and abnormal segments (*i.e. temporal tokens*). The anomaly-aware positions are generated by learning the temporal features of a uni-class distribution and treating the outlier as an anomaly. Then, the anomaly-aware position encodings are infused with the temporal tokens and processed by the CTST. The proposed CTST ensures a superior global temporal relation encoding among normal events and anomalies (*i.e. both long and short*) thanks to its two key components: multi-stage design choice, and Cross Temporal Field Attention block (CTFA). The multi-stage design choice allows the CTST to analyze the anomaly-aware position-infused input tokens at different scales by multi-scale tokenization. By this, the transformer encodes the fine-grained temporal relations for the short anomalies at the lower stage and coarse contextual relations for long anomalies at the higher stages. Further, each stage has a CTFA block to effectively encode the correlations between the temporal neighbor and distant tokens, where a stronger neighbor and distant correlations are encoded for short and long anomalies respectively. The main contributions of the work are as follows:

- An Outlier Embedded Cross Temporal Scale Transformer (OE-CTST) to effectively detect long and short anomalies in untrimmed videos.
- A new manner to learn anomaly-aware position embedding to guide the transformer in better global temporal modeling under weak supervision.
- An exhaustive experimental analysis to corroborate the robustness of OE-CTST on three competitive datasets UCF-Crime [33], XD-Violence [39] and IITB-Corridor [30] datasets, outperforming previous approaches.

2. Related Work

Video anomaly detection (VAD) is a prominent computer vision task and popular methods either adopt unsupervised (training with only normal videos) or weakly-supervised learning (training using both normal and anomaly videos with video-level labels). As unsupervised methods [1, 7, 12, 12, 16, 20, 28, 34, 44] assume the availability of all possible normal videos for training and as it is quite difficult to collect in one dataset, these methods produce high false alarm in complex real-world environments. In contrast, recent weakly-supervised VAD methods [15, 23–27, 33, 37, 39, 41, 43, 45, 47] overcome the drawback of unsupervised counterparts and ensure greater generalization for real-world settings. Inspired by this, we adopt weakly-supervised settings in our work.

Multiple instance learning (MIL) and self training based methods are two majorly adopted paradigms in weakly-supervised VAD. MIL was first introduced in [33] and has become a main stream paradigm for VAD. It trains a segment level anomaly detector that inputs the global scene based pre-computed deep features and optimized with a classical maximum score based ranking loss. Authors in [43, 47] extend the notion of score based optimization and proposed inner bag loss and motion aware loss to enhance the class separability. As these methods rely on a few high-level segment regression scores, they overlook the low-level feature boundary. For this, Tian *et al.* [36] propose a global temporal feature magnitude-based learning paradigm for better separability between normal and anomaly segments with minimum and maximum feature magnitudes. Following [36], Chen *et al.* [4] enhanced the feature based optimization with contrastive loss. However, the feature magnitudes are influenced by only strong spatio-temporal change across temporal segments leading to ineffective separability for subtle and local anomalies. On the other hand, the Self training based method in VAD aims to iteratively generate pseudo labels for unlabelled data and optimizes a classifier for detection. Zhong *et al.* [45] capture the temporal consistency in a GCN and Feng *et al.* [10] use a deep MIL ranker module to generate pseudo labels and trains a 3D ConvNet with the pseudo labels. Further to enhance its ability authors in [14, 42] aims to refine the pseudo labels iteratively and utilize the uncertainty to reduce the effect of noisy pseudo labels in a multi-stage training paradigm. However, due to the complexity involved in a multi-stage iterative training scheme, Majhi *et al.* [22] recently proposed a MIL based one-step optimization that not only generates pseudo labels

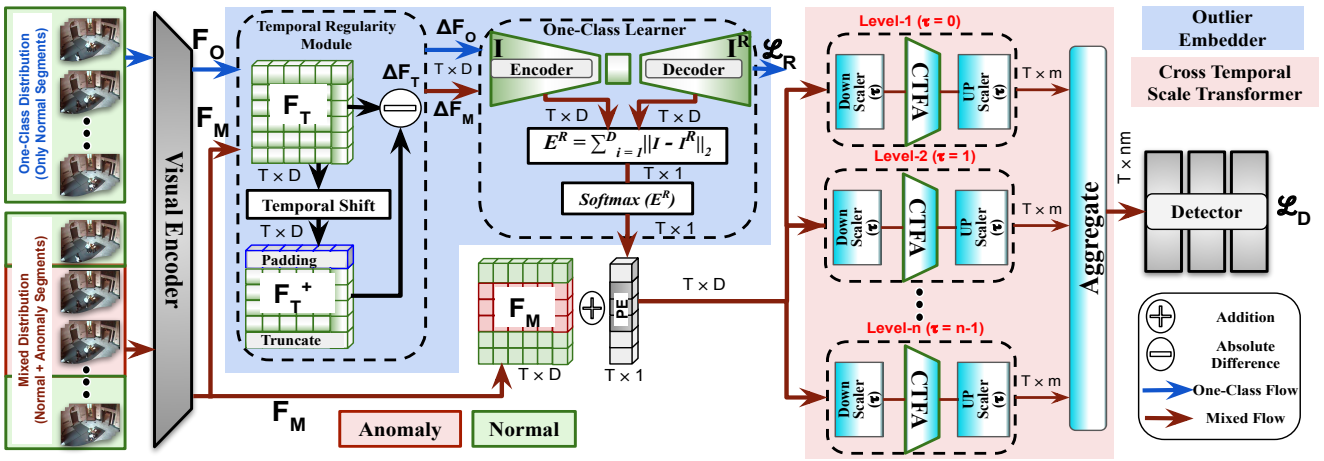

Figure 2. **Outlier-Embedded Cross Temporal Scale Transformer (OE-CTST):** It comprises four major building blocks *i.e.* (A) Visual Encoder, (B) Outlier Embedder, (C) Cross Temporal Scale Transformer, and (D) Detector to detect long and short length anomalies. OE-CTST inputs two dissociative event distributions (*i.e. (i) one-class, (ii) mixed*) during training. However, during inference, the model can correctly detect anomalies for a given untrimmed video. Here, $\boldsymbol{F_O}$= feature map of one-class, $\boldsymbol{F_M}$ = feature map of mixed distribution, $\boldsymbol{F_T}$ = input feature map to temporal regularity module and $\boldsymbol{F_T} \in \{\boldsymbol{F_O}, \boldsymbol{F_M}\}$, $\boldsymbol{F_T^+}$= time-shifted video feature map of $\boldsymbol{F_T}$, $\boldsymbol{\Delta F_T}$= output feature map of temporal regularity module and $\boldsymbol{\Delta F_T} \in \{\boldsymbol{\Delta F_O}, \boldsymbol{\Delta F_M}\}$, **CTFA** = Cross Temporal Field Attention.

but also refines them in an end-to-end manner. Inspired by this, we adopt Majhi *et al.* [22] optimization in our work for anomaly detection.

Temporal modeling is another crucial aspect in real-world VAD to learn discriminative features for normal and anomaly segments. As a classical approach, authors in [43] and [47] utilize TCN [13] and optical flow motion [35] cues respectively to capture the short-term sharp temporal variations to aid only short anomalies. In contrast, authors in [36] proposed a multi-scale temporal convolution network (MTN) for global temporal dependency modeling between normal and anomaly segments. Recently, Zhou *et al.* [46] and Chen *et al.* [4] adopt transformer-based global-local and focus-glance blocks respectively to capture long and short-term temporal dependencies in normal and anomalous videos. However, as these methods [4, 36, 46] follow a magnitude-based optimization, they only encourage the sharp abnormal cues of short anomalies to take part in temporal modeling. Thus, they tend to overlook the subtle cues present at the beginning of long anomalies and hence fail to detect them with tight boundaries. Motivated by this, we propose a Outlier-Embedded Cross Temporal Scale Transformer (OE-CTST) that first generates anomaly-aware temporal information for both long and short anomalies and hence allows the transformer to effectively model the global temporal relation among the normal and anomalies.

3. Outlier-Embedded Cross Temporal Scale Transformer (OE-CTST)

Our novel outlier embedded cross-temporal scale transformer (OE-CTST) delineated in Figure 2 aims to temporally detect normal and anomaly segments using weakly-labelled training videos. In this setting, a set of untrimmed videos V with only video-level labels Y is given for training where a video V_i is marked as normal $Y_i = 0$ (*i.e. one-class*) if it has no anomaly and to be anomaly $Y_i = 1$ (*i.e. mixed*) if it contains at least one abnormal clip. OE-CTST has four key building blocks: **(A) Visual Encoder** that extracts initial spatio-temporal representation, **(B) Outlier Embedder (OE)** that learns representations from normal segments and can generate anomaly-aware pseudo temporal position embeddings for long untrimmed anomaly videos, **(C) Cross Temporal Scale Transformer (CTST)** that ensures better global temporal relation modeling by encoding the stronger correlations between the temporally neighbor and distant tokens, **(D) Detector** that estimates anomaly scores for each temporal token to finally detect the anomalies. A concise description of each building block of OE-CTST is given in the following subsections.

3.1. Visual Encoder

Primarily, the objective of visual encoder is to extract *off-the-shelf* spatio-temporal features from long untrimmed videos. At first, the input video V is divided into T non-overlapping contiguous temporal segments, where a segment has a set of consecutive frames. For a given segment, we consider a Video-swin [18] transformer to extract a feature map of dimension $c \times D$, where c is the number of 16-frame clips present inside a segment. Since multiple 16-frame clips can be present inside a segment, we take a `max-pooling` operation along c, to get a $1 \times D$ dimension segment-level feature per segment. Each of the segment-level feature can be seen as a temporal token and for a given V with T segments, the visual encoder results in a video feature map of dimension $T \times D$. During training, the visual encoder outputs two batches (*i.e. each from one-class and mixed distribution*) of video feature maps *i.e.* F_O and F_M to be processed by OE and CTST respectively.

3.2. Outlier Embedder (OE)

In order to generate anomaly-aware pseudo-temporal position embedding for anomaly events in untrimmed videos, it is necessary to learn the normal segment-level representations so that a temporal segment that deviates largely from the learned normal patterns is treated as an outlier (*i.e. anomaly*). For such a case, it is intuitive to learn the spatio-temporal cues of videos pertaining to one-class (*i.e. normal*) distribution which is extensively adopted in unsupervised approaches [29,31,32], but never used in weakly-supervised video anomaly detection. This is due to the existence of a large intra-class variance in spatio-temporal cues of the normal distribution that makes one-class methods ineffective. For this, we propose an outlier embedder that learns the temporal regularity rather than appearance cues in normal videos. The outlier embedder has two functional blocks: *i.e. (i) Temporal regularity module, (ii) One-class learner* as illustrated in Figure 2.

3.2.1 Temporal Regularity Module

The temporal regularity module (TRM) aims at computing the temporal changes among consecutive temporal segments. It is assumed that the temporal regularities for normal videos are consistent whereas their anomaly counterparts are relatively inconsistent. TRM inputs the $T \times D$ dimensional video feature maps $F_T \in \{F_O, F_M\}$ obtained from the visual encoder for temporal regularity computation. At first, it applies a `temporal shift` operation to F_T that principally moves the temporal tokens along the temporal dimension. The outcome of the `temporal shift` operator is also a $T \times D$ dimensional video feature map F_T^+ where the first and last temporal tokens are respectively padded and truncated. Then, an absolute difference between F_T and F_T^+ is computed to denote the temporal regularity ΔF_T. This operation enables to capture the amount of change between consecutive segments. TRM outputs two $T \times D$ dimensional temporal regularity feature maps *i.e.* ΔF_O, ΔF_M for its corresponding input F_O, F_M. Then, ΔF_O is fed to the one-class learner for normality learning and ΔF_M fed to trained one-class learner for computing the anomaly-aware temporal positions.

3.2.2 One-class Learner (OC-L)

The purpose of the one-class learner (OC-L) module is to explicitly learn the normal cues. Thus, our OC-L takes the architectural configuration of previous popular unsupervised anomaly detection methods (*i.e.(i) temporal autoencoders [11], (ii) spatiotemporal autoencoders [6], (iii) U-Net [16]*) which learn the normal latent space representation by feeding and reconstructing the frames pertaining to the normal scene into an auto-encoder architecture. However, our configured OC-L learns the normality by reconstructing the temporal regularity token of one-class distribution *i.e* ΔF_O. Since our OC-L learns the $T \times D$ dimensional temporal regularity video feature map ΔF_O instead of $T \times H \times W \times C$ (*as in previous unsupervised methods*), our method is computationally less expensive while learning the one-class distribution effectively. Further, any unsupervised encoder-decoder structure can be easily configured and embedded into OC-L in a plug-and-play manner to enhance the normality learning of ΔF_O. The OC-L is optimized with a `reconstruction loss` for a normal one-class distribution temporal regularity map (ΔF_O) as input, whereas it generates anomaly-aware position embeddings (PE) for a mixed distribution temporal regularity map (ΔF_M). The PE is obtained by first computing the temporal token-wise error ($E^R \in \mathbb{R}^{T\times 1}$) b/w input ($I$) and output ($I^R$) of encoder and decoder respectively; $E^R = \sum_{j=1}^{D}(\|I - I^R\|_2)$. Then, to obtain the temporal PE at i^{th} timestep, the E^R are normalized with *softmax* activation: $PE_i = \frac{\exp(E_i^R)}{\sum_{i=1}^{T}\exp(E_i^R)}$. As real-world distribution has large variance in anomaly types due to the presence of sharp (e.g. explosion) and subtle (e.g. shoplift) cues, the token-wise error E^R is less salient and leads to ambiguities between normal events and subtle anomalies. Hence, *softmax* normalizes the E^R across temporal locations and assigns a higher likelihood to anomaly temporal tokens with higher E^R w.r.t. normal ones as the position information. This anomaly-aware PE is infused with the mixed distribution video feature map F_M and fed to cross temporal scale transformer for global temporal modeling of long-short anomalies.

3.3. Cross Temporal Scale Transformer (CTST)

The goal of the cross temporal scale transformer (CTST) is to learn discriminative representations for long-short length anomalies w.r.t. normal counterparts. Since the short and long anomalies are characterized by disjoint cues (*i.e.* sharp and progressive spatio-temporal cues respectively), it is beneficial to encode the temporal relations at multiple semantic levels (*i.e. temporal scale*). For this, as shown in Figure 2, our CTST follows a multi-level architecture based on a temporal feature pyramid to benefit both long and short length anomalies. The lower levels of CTST encode the fine-grained sharp temporal change for short anomalies and the higher levels gather the contextual temporal evolution of long anomalies.

Each level of CTST processes the input video feature map (F_M) at a particular temporal scale with its three major components: **(i) down scaler** performs the temporal downsampling for neighborhood context aggregation, **(ii) Cross Temporal Field Attention** (CTFA) estimates the pairwise correlation between temporally neighbor and distant tokens for enhanced global temporal relation modeling, and **(iii) up scaler** performs the temporal upsampling to regain the original temporal resolution (T). The temporal down and up sampling of F_M are done with a scaling factor (τ), where $\tau \in 0, 1, \ldots, n-1$ and $n =$ no. of levels in CTST. Thanks

to the down and up scalers, the CTFA processes the F_M at various temporal semantic levels for superior pairwise correlation estimation among the temporal tokens. For instance, level-1 of CTST has $\tau = 0$, which implies that the temporal relation modeling in CTFA takes place at a semantic-level equivalent to the original temporal resolution (T) (*i.e. fine-grained*), whereas level-n has $\tau = n - 1$, that implies the coarse level semantic relation building in CTFA at a temporal resolution of $T/2^{(n-1)}$.

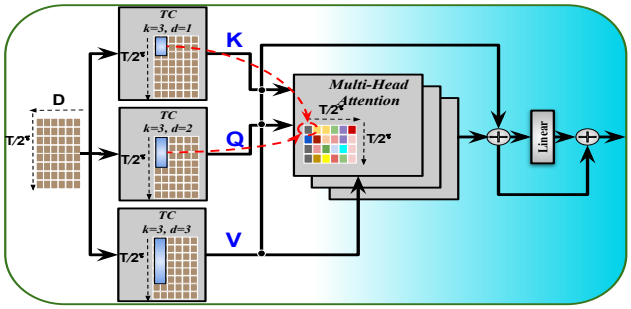

Figure 3. The **Cross Temporal Field Attention (CTFA)** encodes the correlation between the neighborhood and distant temporal token. Here, TC, k, and d denote 1D convolution, kernel size, and dilation rate respectively.

Cross Temporal Field Attention (CTFA): In each level of CTST, the cross-temporal field attention (CTFA) block (shown in Figure 3) inputs the $T/2^{\tau} \times D$ video feature map (X) generated by down scaler for capturing the semantic correlation between temporally neighbor and distant tokens. While vanilla transformers typically use self-similarity, wherein query, key, and value are based on the same temporal semantics, we extend this notion through the concept of cross-temporal semantic similarities. First, the input video feature map is projected into three parallel temporal convolutions (TC) layers, each has m conv filters with kernel size $k \in \{3\}$ and dilation rate $d \in \{1, 2, 3\}$ respectively to capture the short-term temporal consistency among neighbor (*TC with d=1*) and distant (*TC with d=2,3*) tokens. Similar to [3], these local projections are made to benefit both short and long anomalies by infusing contextual cues from tokens at distant locations. Second, to model the global temporal relation, these projections are fed to standard multi-head attention [9] in terms of $K \in \mathbb{R}^{(T/2^{\tau}) \times m}$, $Q \in \mathbb{R}^{(T/2^{\tau}) \times m}$, and $V \in \mathbb{R}^{(T/2^{\tau}) \times m}$ as shown in Figure 3. The self-attention map for i_{th} head can be computed as below:

$$ATT_i(K, Q, V) = softmax(\frac{QK^T}{\sqrt{m_h}})V \tag{1}$$

where m_h = feature dimension in each head. $ATT_i \in \mathbb{R}^{(T/2^{\tau}) \times m}$ generated from each head can encode the correlation between the neighbor and distant temporal consistency, where a stronger neighbor and distant correlations are encoded for short and long anomalies respectively. Next, the attention maps generated from all heads are `added` with the K, Q, and V through a skip-connection to retain the video feature map inductive bias. Then, it is projected to a `linear` layer for local feature mixing and it is `added` to the original inductive bias to locally mix features through a skip connection. The output of CFTA is a $T/2^{\tau} \times m$ dimensional temporally encoded video feature map, which is then upsampled with the upscaler to retain the original temporal scale *i.e.* T. At the end of the cross-temporal scale transformer, the $T \times m$ dimensional video feature maps obtained from n levels are combined by concatenating them along the m-axis. This results in $T \times nm$ dimensional video feature maps which are fed to the detector for anomaly detection.

3.4. Detector

The detector is a multi-layer perceptron (MLP) with three `fully-connected` (FC) layers which input the $T \times nm$ dimensional video feature maps to assign anomaly ranks (*or* scores) to each temporal token. For this, the final layer of MLP has a single neuron with *sigmoid* activation to rank each temporal token independently. Finally, the detector outputs a score map S of dimension $T \times 1$ to be used in anomaly detection.

Optimization: The proposed framework containing an outlier embedder (OE) and a cross temporal scale transformer (CTST) with detector is jointly trainable with two disjoint batches of input video feature maps. Here, similar to [33, 36], the visual encoder is a pre-trained frozen module which is only used for feature extraction. The OE takes only the normal video feature maps (F_O) and is optimized with `reconstruction-loss` as shown in (2). CTST with detector takes both normal and anomaly video feature maps $F_M \in \mathbb{R}^{T \times nm}$ into account to compute the normal ($S_n \in \mathbb{R}^T$) and anomaly ($S_a \in \mathbb{R}^T$) temporal token wise scores and optimizes itself with a `self-rectifying loss` proposed by [22] as shown in (3) and (4).

$$\mathscr{L}_R(F_O) = \|F_O - F_O^R\|_2 \tag{2}$$

$$\begin{aligned}\mathscr{L}_D(S_a, S_n) = \lambda_1 \max(0, 1 - \sum_{i=1}^{T}(S_a^i) + \sum_{i=1}^{T}(S_n^i)) \\ +\lambda_2 \|Err(Correct) - Err(Noisy)\|\end{aligned} \tag{3}$$

$$Err(X) = \begin{cases} \underbrace{\frac{1}{T}\sum_{i=1}^{T}(S_n^i - Y_n^i)^2,}_{MSE(S_n)} & \text{if } X = Correct \\ \quad \forall i, \quad Y_n^i = Normal & \\ \underbrace{\frac{1}{T}\sum_{i=1}^{T}(S_a^i - Y_a^i)^2,}_{MSE(S_a)} & \text{if } X = Noisy \\ \quad \forall i, \text{ if } S_a^i \le S_{ref} \text{ then } Y_a^i = Normal, & \\ \quad \forall i, \text{ if } S_a^i > S_{ref} \text{ then } Y_a^i = Anomaly & \end{cases} \tag{4}$$

The `self-rectifying loss` ensures both video context level and temporal instance (*i.e. token*) level score maximization between normal and anomaly. This is done

by generating a pseudo-temporal annotation (*i.e.* Y_n *and* Y_a *for normal and anomaly respectively*) for each token and refining it by minimizing $\|Err(Correct) - Err(Noisy)\|$. Here, $Err()$ is `mean-squared-error (MSE)`. In S_a, the pseudo temporal labels Y_a^i are computed by comparing their prediction scores (S_a^i) to a dynamic reference point (S_{ref}), where $S_{ref} = (\max(S_a) + \min(S_a))/2$. Where as in S_n, Y_n is always set to 0 (*i.e. normal*) as it contains no anomaly. The overall objective function of our OE-CTST is defined as $\mathscr{L}_{total} = \beta_1 \mathscr{L}_R(F_O) + \beta_2 \mathscr{L}_D(S_a, S_n)$, where β_1 and β_2 are the loss weighting factors. This $\mathscr{L}_{total}$ is used to train our model in an end-to-end manner.

4. Experiments

4.1. Datasets and Evaluation Matrics

The experiments are conducted on three public anomaly detection datasets which have adequate samples from both long and short-duration anomalies, namely UCF-Crime (UCF-C) [33], XD-Violance (XD-V) [39], and IITB-Corridor (IITB-C) [30]. In this work, we evaluate our framework in terms of (i) *overall* and (ii) *long-short anomaly* performance. For *overall* performance, we use the official test set of UCF-C, XD-V, and IITB-C datasets given by [33], [39], and [22] respectively. Since the official test set of UCF-C and IITB-C are biased towards certain abnormal categories, we also evaluate the overall performance in the K-Fold test set of UCF-C and IITB-C datasets for a robust evaluation. As the temporal annotation of the complete dataset is required in K-fold evaluation, we obtained it from Wan *et al* [38]. Similarly, for *long-short anomaly* performance, we use K-Fold test set in UCF-C and IITB-C datasets. However, due to the unavailability of temporal annotations for the complete XD-V dataset, we evaluate the long-short anomaly performance in the official test-set given by [39]. We consider anomalies longer than 2 seconds in duration as long anomalies and others as short ones. In UCF-C and IITB-C datasets, we use frame-level AUC as the performance indicator, but for XD-V dataset we follow [39] to use frame-level average precision (AP) as the performance indicator. In addition, for all K-Fold evaluations, we report the mean AUC (mAUC). **Kindly refer to supplementary material for complete dataset descriptions and implementation details**.

4.2. Ablation Study

A detailed study is carried out in this section to quantify the robustness and novelty of the OE-CTST framework. For all ablation studies the official test-sets of UCF-C, XD-V and IITB-C datasets are choosen.

Effectiveness of OE-CTST: In order to show the necessity of key components present in OE-CTST, each component is evaluated in terms of anomaly detection performance as shown in Table 1. First, as a baseline experiment the detector is stacked on top of a visual encoder (*video swin*) to report initial detection performance and afterward, the remaining components are added. To begin with the temporal modeling by multi-level design in CTST, we first use a vanilla transformer [8] block (*i.e.* K,Q,V from same temporal token). This significantly improves performance compared to the baseline which shows the need for multi-scale temporal modeling in anomaly detection. Then, we replace the vanilla block with CTFA in CTST and found a performance gain of $+3.29\%$, $+2.36\%$, $+6.86\%$ in UCF-C, XD-V, IITB-C datasets respectively w.r.t. baseline. This outlines the superiority of CTFA blocks in temporal modeling compared to vanilla ones. Further, we infuse the vanilla sine-cosine temporal position embeddings with the input temporal tokens to CTST and we obtain a slight performance gain. But, when we infuse the position embedding from OE in place of sine-cosine, a performance gain of $+4.49\%$, $+6.36\%$, $+2.05\%$ is achieved in UCF-C, XD-V, IITB-C datasets respectively w.r.t. our CTST (*row-3*). This performance boost corroborates the potentiality of OE in generating anomaly-aware position information to benefit global temporal modeling in CTST.

Baseline	CTST		PE		AUC(%)		AP(%)
	Vanilla	CTFA	Vanilla	OE	D1	D3	D2
✓	-	-	-	-	79.21	80.35	73.06
✓	✓	-	-	-	81.78	82.89	74.93
✓	-	✓	-	-	82.50	87.21	75.42
✓	-	✓	✓	-	83.79	87.97	76.21
✓	-	✓	-	✓	**86.99**	**89.26**	**81.78**

Table 1. Ablation to show the impact of each component in OE-CTST framework on UCF-Crime (D1), XD-Violence (D2), and IITB-Corridor (D3) datasets.

OC-L	w/o Pre-train			w Pre-train		
	UCF-C	XD-V	IITB-C	UCF-C	XD-V	IITB-C
T-AE	85.02	**81.78**	87.62	**86.99**	80.96	89.04
ST-AE	85.31	81.54	87.79	**86.99**	80.91	**89.26**
UNet	85.37	81.31	87.88	86.94	80.62	89.18

Table 2. Ablation to study various one-class learner (OC-L) designs in terms of anomaly detection performance.

Study of Various OC-L: The one-class learner (OC-L) block in outlier embedder (OE) is flexible to adapt state-of-the-art unsupervised encoder-decoder approach for enhanced normal temporal regularity learning. Here, we study three popular architectures: (i) temporal autoencoder (T-AE) [11], (ii) spatio-temporal autoencoder (ST-AE) [6], (iii) UNet [16] as reported in Table 2 to show the impact of various design choices in anomaly-aware positions. Further, we also study the impact of pre-training the OC-L on all three datasets. From Table 2, it can be observed that all three design choices achieve similar performance with a marginal difference. This is due to the learning of more salient features (*i.e* temporal regularities) rather than appearance cues of raw video frames. Further, it can be seen that pre-training OC-L (particularly ST-AE) improves the detection perfor-

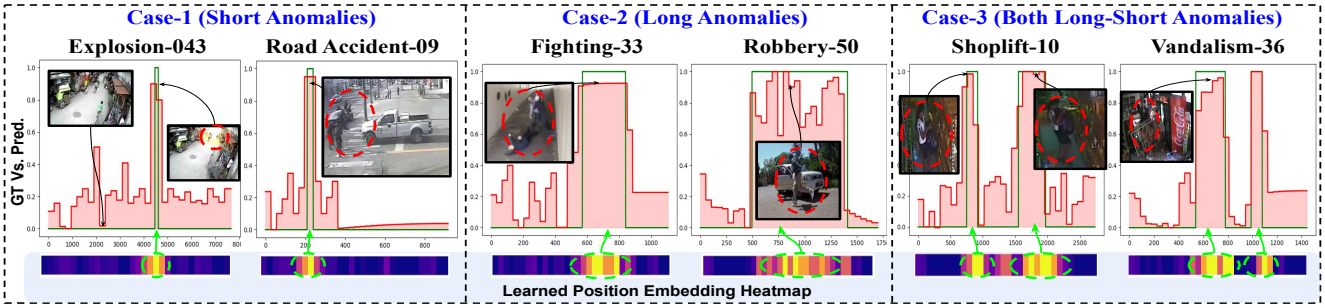

Figure 4. Visualization of ground truth (green shed) vs. prediction scores (red shed) for various cases in Row-1. For each plot in Row-1, the X and Y axis denotes the number of frames and corresponding scores respectively. Row-2 shows the learned anomaly-aware position embedding by outlier embedder in terms of heatmap, where the lighter region corresponds to an anomaly event.

Datasets	Levels (n) in CTST				
	n=0	n=1	n=2	n=3	n=4
UCF-C	80.42	81.92	84.7	**86.99**	85.3
XD-V	75.16	78.41	**81.78**	80.90	80.42
IITB-C	83.35	86.78	**89.26**	88.64	88.17

Table 3. Ablation to show the upper bound of CTST by increasing the number of levels (n) in all three datasets

mance in UCF-C and IITB-C datasets by $+1.68\%$, $+1.47\%$ respectively. However, for XD-V datasets, T-AE without pre-training works well compared to with pre-training. This is due to the existence of large intra-class variance in the normal distribution of XD-V dataset.

Upper Bound of CTST: In order to check the sensitivity of CTST hyper-parameters, we linearly increase the number of levels (n) and analyze its impact on overall detection performance as reported in Table 3. It is found that the performance tends to decline after $n = 3, 2, 2$ in UCF-C, XD-V, and IITB-C datasets respectively which denotes the upper bound. This observation is reasonable as the average length of videos in XD-V and IITB-C is smaller than that of UCF-C, so CTST requires less temporal decomposition in its transformer block w.r.t. the UCF-C counterpart.

4.3. Qualitative Analysis

In Figure 4, we show the anomaly detection performance of our OE-CTST in terms of prediction scores (Row-1) in major three cases (**case-1: short anomalies, case-2: long anomalies, and case-3: both long and short anomalies**). For each case, we consider two examples to quantify the robustness of our method. Further, we show the learned position embedding by outlier embedder in terms of heatmap in Row-2. From Figure 4, it can be observed that our method is able to effectively detect the short anomalies (case-1) in both "Explosion-043" (*a bomb explodes in a street*) and "RoadAccident-09" (*a car runs over a biker*) videos by generating high scores for abnormal segments. Similarly, for long anomalies (case-2) our model also precisely detects the abnormality in both "Fighting-33" (people fight in a pathway) and "Robbery-50" (robbing a car driver). For both cases, the learned position embedding also corresponds to the abnormal temporal position which outlines the effectiveness of our method. Further, we analyze our method on case-3 (*i.e. both long and short anomalies*). "Shoplift-10" and "Vandalism-36" videos capture such a situation where short and long anomalies are paired. In "Shoplift-10", a girl tries to steal things from a shop in a repetitive manner. Similarly, in "Vandalism-36" a person destroys property inside a shop. For both examples, our method precisely detects the long-short anomalies thanks to the anomaly-aware position information from OE followed by effective coarse-fine temporal modeling in CTST.

4.4. State-of-the-art Comparison

In Table 4, we compare our method with the recently reported competitive methods for UCF-C, XD-V and IITB-C datasets. The comparison is made upon two indicators *i.e. overall and long-short performance* to justify our claim. For a fair comparison, we use two popular visual encoders *i.e.* I3D ResNet50 (I3D-Res) [2] and video-swin transformer (V-Swin) [18]. Further, as the K-Fold evaluation in overall and long-short performance is introduced by us, we re-implement previous approaches [22, 33, 36, 39] for the K-fold evaluation.

Overall Performance: The overall performance comparison is made with the official split for the three datasets, and also with the K-Fold split for two datasets (*i.e.* UCF-C and IITB-C, shown in Table 4) to get a better understanding. The unavailability of temporal annotations in the complete XD-V dataset, limits the K-fold evaluation. In UCF-C, our method outperforms the previous I3D-Res based method Majhi *et al.* [22] by a $+0.92\%$ and $+0.42\%$ margin in official and K-fold test sets respectively. As [22] has a complex network that considers human trajectories for relation modeling with the scene, it can not be applied to XD-V dataset due to the unavailability of human trajectories. Further, considering V-Swin as visual encoder, our method has gained 0.32% and 1.07% performance compared to Chen *et al.* [4] in official and K-fold test set of UCF-C dataset. Although authors in [4] utilized transformer blocks to enhance temporal features, their temporal modeling ability remains limited due to feature-magnitude based optimization which overlooks the subtle cues and enhances the sharp cues. Further, in overall performance comparison, the performance gain in K-fold has much more relevance, since it covers

Methods	Feature	Overall Performance					Long-Short Performance					
		UCF-C		XD-V	IITB-C		UCF-C (K-Fold)		XD-V (Official)		IITB-C(K-Fold)	
		Official	K-Fold	Official	Official	K-Fold	Long	Short	Long	Short	Long	Short
Sultani *et al.* [33]	C3D	75.41	-	73.20	-	-	-	-	-	-	-	-
	I3D-Inc	77.42	78.89	75.68	74.59	65.82	42.31	57.41	68.51	80.06	72.23	76.8
Wu *et al.* [39]	I3D-Inc	82.44	83.01	75.41	79.46	73.28	48.70	62.38	70.36	83.21	74.19	78.29
Tian *et al.* [36]	I3D-Res	84.30	84.62	77.81	81.12	74.34	49.82	63.22	71.90	85.76	75.33	80.42
Majhi *et al.* [22]	I3D-Inc	84.33	-	-	84.12	-	-	-	-	-	-	-
	I3D-Res	85.45	86.50	-	86.98	72.40	52.21	63.01	73.22	84.13	74.60	79.60
Chen *et al.* [4]	V-Swin	86.67	86.89	80.11	88.17	77.28	52.16	63.46	73.96	86.62	77.02	81.67
Ours	**I3D-Res**	**86.37**	**86.92**	**80.56**	**88.02**	**77.46**	**54.79**	**63.63**	**74.96**	**86.13**	**76.72**	**81.17**
	V-Swin	**86.99**	**87.96**	**81.78**	**89.26**	**79.22**	**55.31**	**64.01**	**75.32**	**87.65**	**78.05**	**82.14**

Table 4. State-of-the-art comparisons in terms of overall and long-short anomaly performance in UCF-C, XD-V, and IITB-C datasets.

the complete dataset with diverse and unbiased categories of anomalies during evaluation. Unlike [22] and [4], ours is a generic method and has a more robust temporal relation modeling ability. Thus, in XD-V dataset our method outperforms [36] by a significant ($+2.75\%$ and $+1.67\%$ margin in I3D-Res and V-Swin respectively). Further, we achieve a $+1.09\%$ and $+1.94\%$ performance boost on offical and K-fold test set of IITB-C dataset with V-Swin as a visual encoder. This shows the potentiality of our method in overall performance gain.

Long-short Performance: The long-short performance comparison only includes anomaly videos in the test set and this is done for the K-fold split of UCF-C, IITB-C datasets. However, for XD-V dataset, we use the official split by excluding the normal videos from it. It can be observed from Table 4 that our method outperforms in long anomalies by a $+3.15\%$, $+1.36\%$ and $+1.03\%$ margin in UCF-C, XD-V, and IITB-C datasets respectively with V-Swin as a visual encoder. Since UCF-C contains more long anomaly instances (*shoplifting, robbery, fighting, vandalism*) compared to the other two datasets, the performance boost for UCF-C is larger. Similarly, for short anomalies, our method boosts the performance by $+0.55\%$, $+1.03\%$ and $+0.47\%$ on UCF-C, XD-V and IITB-C datasets respectively with V-Swin as visual encoder. As, XD-V has more short anomalies (*accident, explosion*), thus it has a higher performance gain compared to UCF-C and IITB-C datasets. These results also show that among long and short anomalies, precisely detecting long anomalies with tight boundaries is more challenging compared to short anomalies. This is due to the existence of progressive temporal evolution in long anomalies that mainly lie in close proximity to many normal scenarios.

Discussion: Additionally, we found that on official test sets of D1:UCF-C and D2:XD-V datasets, our method OE-CTST (D1:86.99, D2:81.78) achieves better overall performance compared to other recent state-of-the-art methods such as MSL [14] (D1:85.30, D2:78.59), MGFN [4](D1:86.67,D2:80.11), ECU [42] (D1:86.22, D2:78.74), URDMU (D1:86.97, D2:81.66), UMIL [21] (D1:86.75, D2:N/A), LAA [5] (D1:86.10, D2:81.30). From overall and long-short performance comparisons, it can be seen that in UCF-C dataset there exists a big gap between long-short and overall performance. Since the overall performance considers both normal and anomaly videos for evaluation, the performance gets elevated by accurately predicting many normal videos. As a result, state-of-the-art methods performing well on overall performance may still struggle in long-short anomaly detection. For this, we focus on evaluating methods on long and short anomaly videos. To perform these comparisons, we have to choose state-of-the-art methods (e.g. [4])that have sufficient training information on the public domain to carry out long-short performance evaluation. The significant performance gain (in Table 4) and qualitative results (in Figure 4) shown by our method quantify the robustness towards long anomalies while improving the short anomalies performance as well. This is due to the accurate temporal modeling by OE-CTST, which is missing in previous works.

5. Conclusion

In this work, we propose a novel temporal transformer, OE-CTST, for weakly-supervised anomaly detection. The proposed method generates anomaly-aware temporal position information thanks to an outlier embedder that enables the transformer to better model the global temporal relations between normal and anomaly segments under weak supervision. Further, the cross-temporal scale transformer effectively learns the correlation between temporal neighbors and distant tokens to precisely detect long and short anomalies. From extensive experimentation, we found that OE-CTST achieves superior performance than the competitive methods on three widely used datasets in terms of overall and long-short anomaly performance indicators.

Acknowledgements: This work was supported by Toyota Motor Europe (TME) and the French government, through the 3IA Cote d'Azur Investments managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002

References

[1] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv 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] Joao Carreira and Andrew 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] Agniv Chatterjee, Snehashis Majhi, Vincent Calcagno, and François Brémond. Trichanet: An attentive network for trichogramma classification. In *VISIGRAPP (4: VISAPP)*, pages 864–872, 2023.

[4] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 387–395, 2023.

[5] MyeongAh Cho, Minjung Kim, Sangwon Hwang, Chaewon Park, Kyungjae Lee, and Sangyoun Lee. Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12137–12146, 2023.

[6] Yong Shean Chong and Yong Haur Tay. Abnormal event detection in videos using spatiotemporal autoencoder. In *International symposium on neural networks*, pages 189–196. Springer, 2017.

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

[8] Rui Dai, Srijan Das, Kumara Kahatapitiya, Michael S Ryoo, and Francois Bremond. Ms-tct: Multi-scale temporal convtransformer for action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20041–20051, 2022.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] 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.

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

[12] Jaechul Kim and Kristen 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.

[13] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks: A unified approach to action segmentation. In *European conference on computer vision*, pages 47–54. Springer, 2016.

[14] Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1395–1403, 2022.

[15] 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. IEEE, 2019.

[16] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018.

[17] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31:5427–5441, 2022.

[18] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022.

[19] 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.

[20] 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.

[21] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8022–8031, 2023.

[22] Snehashis Majhi, Rui Dai, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Human-scene network: A novel baseline with self-rectifying loss for weakly supervised video anomaly detection. *arXiv preprint arXiv:2301.07923*, 2023.

[23] Snehashis Majhi, Srijan Das, François Brémond, Ratnakar Dash, and Pankaj Kumar Sa. Weakly-supervised joint anomaly detection and classification. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–7. IEEE, 2021.

[24] Snehashis Majhi, Srijan Das, and François Brémond. Dam: Dissimilarity attention module for weakly-supervised video anomaly detection. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2021.

[25] Snehashis Majhi, Ratnakar Dash, and Pankaj Kumar Sa. Temporal pooling in inflated 3dcnn for weakly-supervised video anomaly detection. In *2020 11th International Confer-*

ence on Computing, Communication and Networking Technologies (ICCCNT), pages 1–6. IEEE, 2020.

[26] Snehashis Majhi, Deepak Ranjan Nayak, Ratnakar Dash, and Pankaj Kumar Sa. Multi-level 3dcnn with min-max ranking loss for weakly-supervised video anomaly detection. In *International Conference on Neural Information Processing*, pages 25–37. Springer, 2022.

[27] Seongheon Park, Hanjae Kim, Minsu Kim, Dahye Kim, and Kwanghoon Sohn. Normality guided multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2665–2674, 2023.

[28] Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2569–2578, 2020.

[29] Adín Ramírez Rivera, Adil Khan, Imad Eddine Ibrahim Bekkouch, and Taimoor Shakeel Sheikh. Anomaly detection based on zero-shot outlier synthesis and hierarchical feature distillation. *IEEE Transactions on Neural Networks and Learning Systems*, 33(1):281–291, 2020.

[30] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[31] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

[32] Mohammad Sabokrou, Mahmood Fathy, Guoying Zhao, and Ehsan Adeli. Deep end-to-end one-class classifier. *IEEE transactions on neural networks and learning systems*, 32(2):675–684, 2020.

[33] 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.

[34] Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu. Scene-aware context reasoning for unsupervised abnormal event detection in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 184–192, 2020.

[35] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018.

[36] 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*, pages 4975–4986, 2021.

[37] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie 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.

[38] Boyang Wan, Wenhui Jiang, Yuming Fang, Zhiyuan Luo, and Guanqun Ding. Anomaly detection in video sequences: A benchmark and computational model. *IET Image Processing*, 15(14):3454–3465, 2021.

[39] 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 *European Conference on Computer Vision*, pages 322–339. Springer, 2020.

[40] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. *Advances in Neural Information Processing Systems*, 34:1086–1099, 2021.

[41] Muhammad Zaigham Zaheer, Arif Mahmood, Hochul Shin, and Seung-Ik Lee. A self-reasoning framework for anomaly detection using video-level labels. *IEEE Signal Processing Letters*, 27:1705–1709, 2020.

[42] Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16271–16280, 2023.

[43] 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.

[44] B. Zhao, L. Fei-Fei, and E. P. Xing. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR 2011*, pages 3313–3320, June 2011.

[45] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[46] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. *arXiv preprint arXiv:2302.05160*, 2023.

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