Expert Systems With Applications 295 (2026) 128753

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Hierarchical Temporal Sequence Segmentation for weakly supervised video anomaly detection

Nuku Atta Kordzo Abiew [a,b], Lijian Gao [a], Qirong Mao [a,c,d,*]

[a] School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, 212013, China
[b] Faculty of Computing and Information Systems, Ghana Communication Technology University, Accra, Ghana
[c] Jiangsu Engineering Research Center of Big Data Ubiquitous Perception and Intelligent Agriculture Applications, Zhenjiang, Jiangsu, China,
[d] Provincial Key Laboratory of Computational Intelligence and New Technologies in Low-Altitude Digital Agriculture, Zhenjiang, Jiangsu, China

ARTICLE INFO

Keywords:
Anomaly detection
Anomaly localization
Hierarchical Temporal Sequence Segmentation
Video anomaly detection and localization
Weakly supervised video anomaly detection

ABSTRACT

Temporal detection and localization of anomalies with weak labels is challenging due to the limited availability of labeled data and the dynamic nature of anomalies. To address these challenges, we introduce a novel *Hierarchical Temporal Sequence Segmentation (HiTESS)* approach that enhances anomaly detection and localization by leveraging multi-level temporal segmentation. Our method begins by segmenting the video into progressively smaller temporal units, enabling the model to capture anomalies at varying scales. The core contribution of our approach is the hierarchical selection of temporal segments, where multi-level scoring aggregates predictions across different time scales. While Bidirectional LSTMs and Multi-Head Attention refine temporal features, the hierarchical aggregation of predictions significantly enhances anomaly detection accuracy. This hierarchical analysis not only improves feature representation but also enhances the detection accuracy of subtle anomalies that might be overlooked in traditional methods. Experimental results demonstrate that the HiTESS method outperforms existing approaches by 1.32 % AUC and 4.8 % AP on the UCF-Crime and XD-Violence datasets, respectively, highlighting its effectiveness in utilizing aggregated features at varied temporal levels for capturing and localizing anomalies in video sequences.

1. Introduction

Anomaly detection and localization in video sequences present a pivotal challenge in the realm of video surveillance and monitoring, impacting sectors from public safety to industrial operations (Benjdira et al., 2022; Huang et al., 2023; Ullah et al., 2023b). The ability to automatically identify abnormal events in video streams facilitates timely intervention, enhances threat detection, and supports accurate anomaly localization. One key aspect of effective video anomaly detection is the modeling of temporal context, which involves capturing the temporal dependencies and patterns present in video sequences (Himeur et al., 2023; Wen et al., 2023; Wu et al., 2021). Temporal context modeling enables the system to discern between normal and abnormal behaviors based on the temporal dynamics exhibited in the data (Ullah et al., 2023a).

In recent years, deep learning has emerged as a powerful tool for temporal context modeling, offering the capability to learn complex temporal representations directly from raw video data (Feng et al., 2021a; Lv et al., 2021). Traditional supervised methods require extensive labeling of video data, which is both time-consuming and often impractical due to the rare and unpredictable nature of anomalous events. This has spurred interest in weakly supervised learning approaches, where only video-level labels are available (Sultani et al., 2018; Wu et al., 2021; Zhang et al., 2023).

Among these, Multiple Instance Learning (MIL) has emerged as a promising paradigm, treating each video as a bag of instances (frames or segments) (Gao et al., 2022; Lv et al., 2023; Park et al., 2020, 2023; Waqas et al., 2024; Zhang et al., 2019). The key assumption of MIL is that if a video is labeled as anomalous, at least one instance within the bag contains an anomaly, although specific instances containing the anomaly are not labeled (Feng et al., 2021a). Inspired by recent advancements in Multi-Sequence Learning (Li et al., 2022) and hierarchical feature representation (Cheng et al., 2015), which demonstrated improvements in capturing complex temporal dependencies, we extend MIL by integrating hierarchical scoring to enhance anomaly localization.

Despite advancements in MIL for video anomaly detection, the precise localization of temporal anomalies remains largely unresolved. This is primarily due to the inherent challenge of identifying the sparse and

* Corresponding author.
E-mail addresses: nukukordzo@outlook.com (N. A. K. Abiew), ljgao@stmail.ujs.edu.cn (L. Gao), mao_qr@ujs.edu.cn (Q. Mao).

https://doi.org/10.1016/j.eswa.2025.128753
Received 27 October 2024; Received in revised form 16 June 2025; Accepted 23 June 2025
Available online 28 June 2025

subtle features that distinguish normal from abnormal patterns within vast amounts of video data (Abdalla et al., 2024). The difficulty increases when these patterns are embedded in the complex and multi-scale temporal dynamics of real-world data.

These intricate temporal variations make it difficult for traditional methods to capture and localize anomalies effectively with high precision.

Additionally, the effectiveness of MIL can be significantly constrained by its insensitivity to the temporal structure within the video, often leading to imprecise localization and error propagation due to incorrect instance selection (Li et al., 2022). Existing methods struggle to localize anomalies precisely, especially when these anomalies are subtle or sparse, and dispersed across long video sequences (Feng et al., 2021a; Sultani et al., 2018; Zhang et al., 2019; Zhong et al., 2019). The challenge is further compounded by the inability of many models to effectively capture the temporal dependencies between video frames, resulting in inaccurate or missed detections. This can lead to an increase in false positives, particularly in dynamic or unpredictable settings, reducing overall model performance.

The motivation for this work arises from the limitations of existing video anomaly detection methods, which struggle to capture the complex and multi-scale temporal dynamics inherent in real-world events. Video sequences often contain subtle and context-dependent anomalies that manifest at different temporal scales, making it challenging for traditional models to detect these nuances effectively. Existing methods typically treat video sequences as uniform entities, failing to adapt to the varying temporal patterns within.

To address this, we introduce the Hierarchical Temporal Sequence Segmentation (HiTESS) approach, motivated by the need to segment video sequences hierarchically, allowing for fine-grained temporal analysis at multiple resolutions. This enables the detection of both short-term anomalies and long-range dependencies by adapting the level of focus based on the temporal scale.

By leveraging Bidirectional Long Short-Term Memory (BiLSTM) with Multi-Head Attention, HiTESS captures rich temporal features and context, providing a more adaptable and precise method for identifying anomalies in dynamic video data.

Conventional fixed-length window analysis often fails to adapt to the varying temporal patterns of anomalous events. Some anomalies may span only a few frames, while others unfold over longer durations. To address this, we introduce a hierarchical segmentation approach that adaptively captures both short-term and long-term anomalies within a unified framework. In addition, we propose a novel MIL-based formulation that leverages the hierarchical structure to improve anomaly localization and learning efficiency under weak supervision.

The key contributions of this work are summarized as follows:

- We propose **HiTESS**, a novel Hierarchical Temporal Sequence Segmentation framework designed to model multi-scale temporal dynamics in video anomaly detection, addressing the limitations of fixed-window methods,which often struggle to localize abnormal events with varying durations.
- We integrate **Bidirectional LSTM** and **Multi-Head Attention** to effectively capture forward and backward temporal context while emphasizing salient features across segments for fine-grained anomaly localization.
- We introduce a **hierarchical scoring strategy** that aggregates anomaly scores from both segment-level and sequence-level perspectives, enhancing robustness and reducing false positives.
- We conduct extensive experiments on two widely used benchmarks (**XD-Violence** and **UCF-Crime**), demonstrating that HiTESS achieves competitive or superior performance compared to state-of-the-art weakly supervised methods, with improved interpretability and localization accuracy.

The remainder of this paper is structured as follows: Section 2 reviews the related work. Section 3 provides a detailed explanation of the proposed HiTESS algorithm. Section 4 outlines the experimental setup and presents the performance evaluation. Finally, Section 5 concludes the paper and outlines potential avenues for future research in video anomaly detection.

2. Related work

Anomalies in videos refer to unusual events or activities that indicate irregular behavior. The objective of anomaly detection is to identify and pinpoint these anomalous events within video sequences, either temporally or spatially (Zhang et al., 2024). As the volume of video data continues to surge, the demand for effective anomaly detection methods intensifies, prompting researchers to explore innovative techniques that leverage the advancements in deep learning (Sultani et al., 2018; Wang et al., 2020). Central to the success of video anomaly detection and localization is the extraction of spatiotemporal features (Ionescu et al., 2018; Yu et al., 2022), which encapsulate the intricate patterns and details present in each frame of a video sequence. In Baradaran and Bergevin (2023), the study employs an attention mechanism for multi-task learning-based video anomaly detection, integrating proxy tasks, object classes, motions, and contextual information.

The combination of BiLSTM with Multi-head Attention ensures effective capture of temporal dependencies in both forward and backward directions, while also enabling the model to focus on the most relevant spatiotemporal features. This integration enhances the model's ability to detect and localize anomalies by capturing complex patterns and relationships across different scales of the data.

2.1. MIL For weakly supervised VAD

Weakly supervised video anomaly detection (WS-VAD) first introduced in Sultani et al. (2018) has emerged as a practical approach, leveraging only video-level labels to identify anomalies. This method offers a significant advantage over unsupervised VAD implementations (Chen et al., 2023; Liu et al., 2021; Ristea et al., 2022; Xu et al., 2017; Zaheer et al., 2022), which rely solely on normal videos during training and often result in a high false alarm rate for previously unseen normal events. By utilizing weakly labeled abnormal or normal training videos, WS-VAD provides a more balanced trade-off between detection performance and manual annotation costs (Liu et al., 2024). Unlike unsupervised methods, which have a limited understanding of anomaly data, WS-VAD improves the accuracy and reliability of anomaly detection by incorporating minimal yet informative labels (Feng et al., 2021a; Hussain et al., 2024).

A prominent framework within WS-VAD is MIL, which treats each video as a bag of instances (frames or segments) and has been widely explored by many researchers (Feng et al., 2021a; Kamoona et al., 2023; Lv et al., 2023; Park et al., 2020, 2023). Xia et al. (Sun et al., 2025) revisited instance-level modeling by enhancing temporal discriminability and adaptive selection, improving performance in MIL-based anomaly detection systems. Generally, WS-VAD models produce anomaly scores by utilizing MIL to compare the spatio-temporal features of normal and abnormal events. In Thakare et al. (2022), Wan et al. (2020), a fused dynamic multiple-instance learning loss and a center loss were designed to learn discriminative features for anomaly detection. Zhong et al. (2024) proposed IFS-VAD, which leverages inter-clip feature similarity and a multi-scale temporal MLP to model fine-grained and long-range dependencies, achieving strong performance under weak supervision.

In addition to MIL, attention mechanisms (Singh et al., 2024; Zhang et al., 2022) have emerged as a prominent paradigm for weakly supervised video anomaly detection and localization. Zhang et al. (2024) proposed a self-supervised framework that learns multi-grained spatio-temporal features via continuity discrimination and missing frame estimation, improving anomaly detection robustness. Attention mechanisms dynamically adjust their focus to different segments of a video,

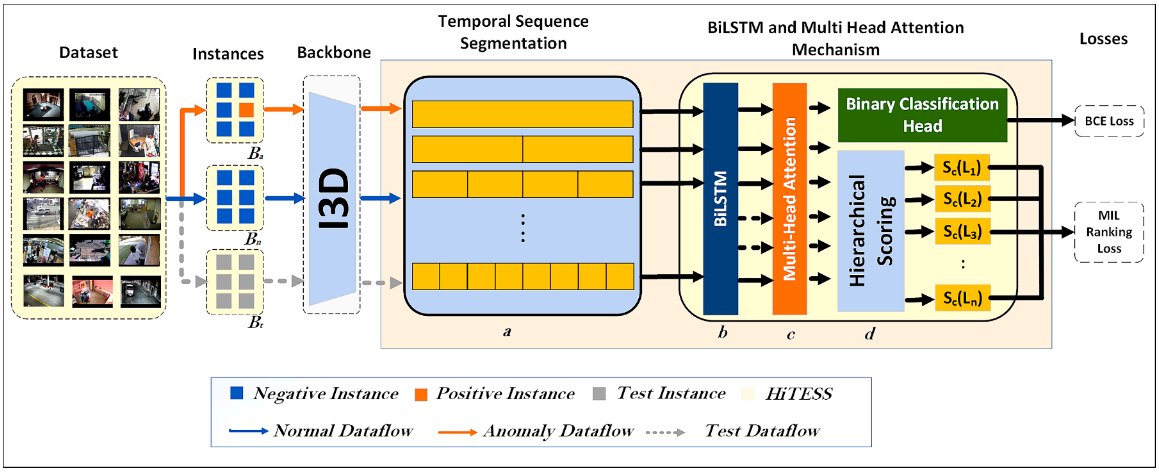

Fig. 1. Conceptual framework for the proposed method. The positive bag B_a (anomalous video) includes at least one anomalous instance, while the negative bag B_n (normal video) contains only non-anomalous instances. Features are first extracted using an Inflated 3D ConvNet(I3D) backbone, providing spatiotemporal representations for each instance. The **HiTESS** framework then integrates: (a) Multi-scale Temporal Segmentation, which captures diverse temporal granularities; (b) BiLSTM for sequential context modeling, enhanced by (c) Multi-Head Attention to encode complex temporal dependencies. Finally, (d) a Hierarchical Scoring Strategy aggregates segment and sequence-level information to refine anomaly detection accuracy.

emphasizing parts that are most indicative of anomalies. This selective focus allows models to highlight significant features while ignoring irrelevant information, enhancing both the accuracy and precision of anomaly detection and localization.

In Baradaran and Bergevin (2023), the study employs an attention mechanism for multi-task learning-based video anomaly detection, integrating proxy tasks, object classes, motions, and contextual information. Also, Watanabe et al. (2022) in 2022, introduced a self-attention mechanism to automatically extract features that are important for determining normal/abnormal from all input segments. Ghadiya et al. (2024) introduced a cross-modal attention mechanism that fuses RGB and optical flow features for weakly supervised video anomaly detection.

TDS-Net (Hussain et al., 2024) proposed a dual-stream transformer architecture for weakly supervised video anomaly detection, leveraging separate motion and appearance branches. While effective, its inference latency (34 ms/sequence) limits deployment for real-time systems. HiTESS, by contrast, achieves lower latency (7.87 ms) using a more efficient BiLSTM-based temporal encoder.

ST-HTAM (Paulraj & Vairavasundaram, 2025) introduced a hierarchical spatiotemporal transformer attention model that captures multiscale feature interactions across frames. While it models long-range dependencies well, its deep transformer stack increases computational cost relative to HiTESS's lightweight temporal segmentation approach.

TTFA (Huang et al., 2022) combines a temporal transformer feature aggregator with a self-guided discriminative feature encoder to improve anomaly localization under weak supervision. Although conceptually aligned with HiTESS in its temporal reasoning goal, our method achieves comparable accuracy with a more compact and interpretable architecture.

TransCNN (Ullah et al., 2023b) demonstrates a hybrid CNN-transformer architecture to enhance spatial representation learning in anomaly detection. Although evaluated on fully supervised datasets such as ShanghaiTech and UCSD, it reflects broader architectural trends that influence WS-VAD design evolution.

Sun and Gong (2024) proposed a multi-scale bottleneck transformer that uses bottleneck token weighting and a temporal contrastive loss to reduce redundancy and modality imbalance, achieving high accuracy on XD-Violence.

Other implementations, such as Feng et al. (2021b), Kaneko et al. (2024), Sun et al. (2023), Yu et al. (2023), Zhang et al. (2019), Zhu and Newsam (2019) show how an attention mechanism can be implemented with other techniques to detect anomalies in video sequences.

By providing a more detailed and context-aware analysis of video data, attention mechanisms complement MIL and have contributed to the improved performance of WS-VAD systems (Wang et al., 2020). While their approach improves detection accuracy, their AP (81.65 %) on XD-Violence is still lower than that of HiTESS (84.91 %), which benefits from hierarchical segmentation and efficient BiLSTM-based temporal reasoning.

3. Methodology

In video anomaly detection, accurately identifying and localizing anomalies is challenging due to their sparse and unpredictable nature. Traditional methods requiring frame-level annotations are impractical. To address this, we propose the HiTESS approach, integrating hierarchical processing with advanced temporal feature extraction. Fig. 1 illustrates the conceptual framework of the proposed HiTESS method.

The framework begins with Hierarchical Temporal Segmentation, laying the foundation for effectively modeling temporal dependencies in video sequences. This is followed by BiLSTM-based Temporal Feature Extraction and the Multi-Head Attention mechanism, both of which are crucial for capturing fine-grained and long-range dependencies. Finally, the framework integrates Hierarchical Scoring with MIL ranking loss to robustly detect and localize anomalies, demonstrating its effectiveness in addressing the challenges posed by the unpredictable nature of anomalous events in videos.

3.1. Hierarchical Temporal Segmentation

HiTESS begins by segmenting video sequences into hierarchical levels, each representing different temporal granularities (Fig. 1). This segmentation allows the model to capture anomalies at various temporal scales. Let $\mathbf{X}$ denote the entire video sequence, which is segmented into hierarchical levels $\{L_1, L_2, \dots, L_H\}$. Each level L_h corresponds to segments of varying lengths, the HiTESS is expressed mathematically as shown in Eq. (1):

$$\mathbf{X} = \bigcup_{h=1}^{H} \mathbf{X}_h \tag{1}$$

where $\mathbf{X}_h$ represents the segments at level h. For each level h, segments $\mathbf{X}_{h,j}$ are defined in Eq. (2) as:

$$\mathbf{X}_h = \{\mathbf{X}_{h,1}, \mathbf{X}_{h,2}, \dots, \mathbf{X}_{h,N_h}\} \tag{2}$$

Here,N_h is the number of segments at level h, allowing the model to process video data at multiple temporal scales.

Given an initial value n, the sequence is structured into levels such that the sum of terms at each level equals n. At level l, the value of each term is represented by:

$$\text{Term}_l = \frac{n}{2^{l-1}} \tag{3}$$

The number of terms at each level is 2^{l-1}, and the product of the term value and the number of terms results in a total sum of n at every level. Thus, both the number of terms and the total sum of terms can be derived directly from Eq. (3), ensuring consistency across the hierarchical structure.

3.2. BiLSTM for temporal modeling

BiLSTM is a variant of the LSTM architecture that addresses the vanishing gradient problem by updating weights and biases more effectively during training. Unlike standard LSTMs, which process sequences in one direction, BiLSTM operates bidirectionally, capturing dependencies in both forward and backward directions. This allows it to access information from both past and future contexts, making it particularly useful for tasks requiring temporal understanding.

The effectiveness of LSTMs in modeling both spatial and temporal dependencies has been widely demonstrated in various computer vision tasks (Satya Krishna et al., 2022; Srivastava et al., 2015; Tang et al., 2023; Wang et al., 2018, 2019a,b). BiLSTM enhances this capability by processing each segment $\mathbf{X}_{h,j}$ in both directions, enriching feature extraction. This bidirectional processing allows the model to generate a more comprehensive temporal feature representation, crucial for detecting subtle anomalies that may be missed by unidirectional models. The output of the BiLSTM for the segment $\mathbf{X}_{h,j}$ is represented by Eq. (4):

$$\mathbf{F}_{h,j} = \text{BiLSTM}(\mathbf{X}_{h,j}) \tag{4}$$

To balance efficiency and temporal expressiveness, we combine BiLSTM with a Multi-Head Attention mechanism. BiLSTM ensures stable convergence and smooth temporal modeling under weak supervision, while the attention module captures global dependencies. This hybrid design offers a lightweight yet effective alternative to full transformer architectures, which are computationally intensive and prone to overfitting in long , untrimmed videos.

By incorporating bidirectional processing, BiLSTM ensures that both short-term and long-term dependencies are captured. This is particularly beneficial for anomaly detection in complex temporal sequences, such as video data, where considering both preceding and subsequent frames improves the model accuracy. The output $\mathbf{F}_{h,j}$ from the BiLSTM layer serves as a refined temporal representation, which subsequent network layers can further process for more accurate anomaly detection.

3.3. Multi-Head Attention mechanism

The significant performance of attention mechanisms in enhancing the discriminative power of models has brought a paradigm shift toward their usage in recent years. Attention mechanisms allow models to focus on specific regions or frames of a video sequence.

The Multi-Head Attention(MHA) mechanism (Islam et al., 2024; Ullah et al., 2023b; Vaswani et al., 2017) is employed to focus on critical aspects of the temporal features extracted by the BiLSTM network. Let $\mathbf{F}_h$ be the set of aggregated features at level h as shown in Eq. (5):

$$\mathbf{F}_h = \{\mathbf{F}_{h,1}, \mathbf{F}_{h,2}, \dots, \mathbf{F}_{h,N_h}\} \tag{5}$$

The MHA mechanism processes these features to capture long-range dependencies and generate attended representations:

$$\text{Attended Features}_h = \text{MultiHeadAttention}(\mathbf{F}_h) \tag{6}$$

This mechanism allows the model to selectively focus on different segments or temporal positions, enhancing its ability to capture relevant patterns and improve anomaly detection performance.

3.4. Hierarchical scoring with MIL ranking loss

The hierarchical scoring method integrates outputs from various temporal levels to enhance anomaly localization by evaluating both local (segment-level) and global (sequence-level) information. For each segment $X_{h,j}$, the anomaly score $S_{h,j}$ is computed as:

$$S_{h,j} = \text{Score}(\text{Attended Features}_{h,j}) \tag{7}$$

The function **Score** in Eq. (7) represents the process of calculating the anomaly score for a given segment $X_{h,j}$. It takes the *Attended Features*, generated through Multi-Head Attention applied to the output of the BiLSTM, and processes them to produce a score that reflects the likelihood of the segment containing an anomaly.

The scoring process is implemented in the HiTESS Scoring Module algorithm (Algorithm 1). This algorithm uses a recursive method to compute anomaly scores for each segment and combine them across temporal levels. At each step, BiLSTM and Multi-Head Attention are applied to capture both local temporal dependencies and contextual information across time. The algorithm processes the sequence and its sub-sequences, recursively calculating scores and pooling them into a global anomaly score.

Algorithm 1 HiTESS scoring module with MIL ranking loss.

```
Require: Sequence $x$, Pooling $pooling$, Labels $y$, Margin $\Delta$
Ensure: Global score
function ScoreSequence($x$, $pooling$)
    // Step 1: LSTM and Attention
    $lstm\_output \leftarrow$ BidirectionalLSTM($x$)
    $att\_output \leftarrow$ MultiHeadAttention($lstm\_output$)
    $pro\_sequence \leftarrow$ Normalize and adjust($att\_output$)
    // Step 2: Calculate the current sequence score
    $initial\_score \leftarrow pooling(pro\_sequence)$
    // Step 3: Recursively process sub-sequences
    $left\_seg, right\_seg \leftarrow$ split sequence($pro\_sequence$)
    $scores \leftarrow [initial\_score]$
    $stack \leftarrow [left\_seg, right\_seg]$
    while stack is not empty do
        $cur\_seg \leftarrow$ stack.pop()
        if $cur\_seg$ is not empty then
            $sub\_score \leftarrow$ score sequence($cur\_seg, pooling$)
            $scores$.append($sub\_score$)
        end if
    end while
    // Step 4: Combine scores using the pooling method
    $global\_score \leftarrow pooling$(stack($scores$))
    // Step 5: Calculate MIL ranking loss (during training)
    if training phase then
        // Apply MIL ranking loss
        $mil\_loss \leftarrow 0$
        for $i \leftarrow 1$ to $N$ do
            for $j \leftarrow 1$ to $N$ do
                if $y_i > y_j$ then
                    $\delta \leftarrow \Delta - (score_i - score_j)$
                    $mil\_loss \leftarrow mil\_loss + \max(0, \delta)$
                end if
            end for
        end for
        $mil\_loss \leftarrow mil\_loss/(N \times N)$
    end if
    return $global\_score$
end function
```

The final anomaly score S_{final} is obtained by combining scores from all hierarchical levels:

$$S_{\text{final}} = \text{Combine}(S_1, S_2, \dots, S_H) \tag{8}$$

The **Combine** function in Eq. (8) merges the anomaly scores from different hierarchical levels. These scores, derived from segment-wise evaluations, are pooled together to generate a final global anomaly score. Average Pooling, identified as the optimal pooling function for this approach, computes the mean of the scores, providing a balanced representation by smoothing extreme values and effectively capturing the overall behavior of the sequence. In contrast, Max Pooling, which selects the maximum value among the segment scores, tends to overemphasize the most significant anomaly, leading to a loss of important contextual information. Similarly, Attention Pooling, which dynamically assigns weights to different parts of the sequence based on relevance, introduces bias by overfocusing on certain segments, resulting in suboptimal detection.

To optimize these scores within the MIL framework, the MIL ranking loss is used. This loss function ensures that segments containing anomalies receive higher scores than those from normal segments. The MIL ranking loss $\mathcal{L}_{\text{MIL}}$ is defined as:

$$\mathcal{L}_{\text{MIL}} = \sum_{i=1}^{N} \sum_{j=1}^{N} \max\left(0, 1 - S_{\text{final}}^{(i)} + S_{\text{final}}^{(j)}\right) \tag{9}$$

where $S_{\text{final}}^{(i)}$ is the score for an anomalous segment and $S_{\text{final}}^{(j)}$ is the score for a normal segment. This approach helps in improving detection precision by optimizing anomaly scores and addressing the limitations of traditional MIL techniques.

In summary, the hierarchical scoring method, combined with MIL ranking loss, enhances anomaly detection by effectively integrating temporal information from various levels and ensuring that anomalous segments are prioritized over normal ones. This approach addresses the limitations of traditional MIL techniques, improving both the accuracy and precision of anomaly localization.

3.5. Anomaly detection

For predicting whether a frame is normal or anomalous, the model uses a sigmoid function:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{10}$$

where $\sigma(z)$ produces probabilities from the model's output. To handle class imbalances, a weighted Binary Cross-Entropy (BCE) loss function is employed. The Binary Cross-Entropy (BCE) loss incorporates class weights to adjust the loss calculation and averages the results to effectively evaluate model performance. This weighted BCE loss ensures that the model balances the influence of normal and anomalous frames:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} \left[w_{\text{pos}} y_i \log(\hat{y}_i) + w_{\text{neg}} (1 - y_i) \log(1 - \hat{y}_i) \right] \tag{11}$$

where y_i is the true label, $\hat{y}_i$ is the predicted probability, and w_{pos} and w_{neg} are the class weights for positive and negative classes, respectively.

In summary, combining hierarchical scoring with MIL ranking loss during training improves anomaly localization and detection precision, while BCE ensures balanced classification and effective evaluation of frame-level predictions. After training, the final detection results were obtained by selecting the segments with the highest anomaly scores in the hierarchy. This approach ensures that the most anomalous segments are prioritized while also considering both local and global patterns in the data, resulting in more accurate and reliable anomaly detection and localization.

4. Experiments and results

In this section, we describe the datasets and the baseline methods used in our study. We provide a comparative analysis of our method against other architectures. In all experiments, accuracy, recall, precision, specificity, F1-Score, and Frame-Level AUC(Area Under Curve) metrics were reported.

4.1. Benchmark datasets

In this work, we evaluated our approach on two datasets: UCF-Crime and XD-Violence.

The **UCF-Crime** (Sultani et al., 2018) dataset spans a total of 128 hours and comprises 1,900 long and untrimmed real-world surveillance videos across 13 categories of anomalous events. Of these, 1610 videos with video-level labels are designated for training, while 290 videos with frame-level labels are reserved for testing

XD Violence (Wu et al., 2020) is a large-scale dataset that contains 4754 untrimmed videos with a total duration of 217 hours and collected from multiple sources, such as movies, sports, surveillance, and CCTVs organized into 7 classes. The training set contains 3954 videos with video-level labels, and the test set contains 800 videos with frame-level labels. To model our approach in a multiple-instance learning paradigm, we considered all 13 anomalous event classes on UCF-Crime and 6 violence classes on the XD-Violence as a positive bag, and the normal activity classes from both datasets as negative bags. Following the Weakly Supervised learning, we leverage on only the video-level labels.

4.2. Implementation details

The objective of anomaly detection is to assess the existence of anomalies in a video and, if present, identify their temporal location within the sequence. In a weakly supervised setting, a video sequence V and its associated video-level label $z \in \{0, 1\}$ are provided, where $z = 1$ signifies the presence of an anomaly, and $z = 0$ indicates that the sequence is anomaly-free. The video sequences are further grouped into positive and negative bags. The positive bag B_p (anomalous video) contains at least one anomalous instance, while the negative bag B_n (normal video) contains only non-anomalous instances. The challenge lies in temporally localizing anomalous segments within the sequence since the video-level labels do not provide precise information about the anomaly's position within the video.

Our implementation of the MIL pipeline for video anomaly detection begins with data preprocessing, in which videos are categorized and frames are extracted, resized, and normalized for consistency. In recent times, pre-trained networks such as Inflated 3D (I3D) (Carreira & Zisserman, 2017), and Convolutional 3D (C3D) (Tran et al., 2015) have become the de facto feature extractors for video data. In our implementation, we chose the I3D network for its strong ability to capture fine-grained spatio-temporal features. Video clips are treated as instances within a bag, with features extracted via the RGB and flow networks of the I3D model.

Feature extraction is a crucial step in the pipeline for modeling temporal context in video anomaly detection and localization. The objective of feature extraction is to transform raw input data into a more compact and informative representation that captures relevant patterns and structures. In the context of video data, feature extraction is particularly important for capturing spatio-temporal information from individual frames.

The model was implemented in Keras 2.0.8 with TensorFlow as the backend, running on a GTX 1050 Ti GPU with 32GB memory. We utilized the I3D network pre-trained on the Kinetics-400 dataset, where each feature in Tian et al. (2021) corresponds to approximately 16 frames of the UCF-Crime dataset. The training process employed the Adam optimizer with a learning rate of 1e-5 and a batch size of 32. A sliding-window approach, with a segment length of 16 frames and a step size of 8, was used to create overlapping windows, ensuring comprehensive coverage of the dataset and enhancing the model's ability to learn detailed temporal relationships.

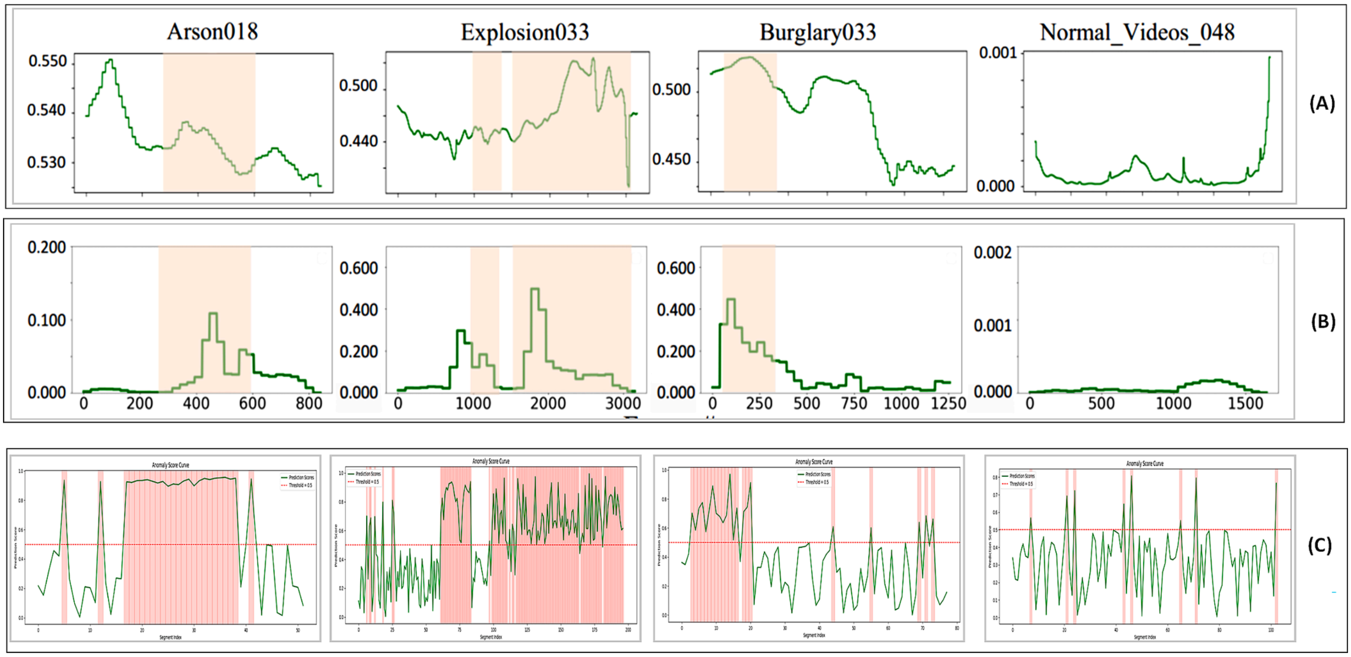

Fig. 2. Temporal localization of the testing results on UCF-Crime. In the figure, (a), (b), and (c) represent the results of Lv et al. (2021), Sultani et al. (2018) and our model, respectively. The light orange shaded regions, thus in (a) and (b) and the light red shaded region in (c) represent the temporal localized anomalies.

4.3. Experimental evaluation

We visualize the localization effects of our approach in Fig. 2 on several test videos from the UCF-Crime dataset. During inference, features extracted from video segments were processed by the model, which outputs prediction scores indicating the likelihood of anomalies in each segment. By applying a predefined threshold to these scores, segments were classified as normal or anomalous, allowing precise location of unusual events within the video.

The presented method demonstrates its strength in localizing anomalous events within video sequences, as effectively illustrated in Fig. 2(c). The plot utilizes the number of frames along the horizontal axis and the anomaly scores along the vertical axis, with curves indicating the predicted anomaly scores over time. The presence of abnormal events is clearly marked by the red-shaded regions, providing a direct and intuitive visualization of the temporal localization of anomalies. This approach is crucial as it not only detects the presence of anomalies but also highlights their precise temporal location within the video sequence, which is essential for applications such as video surveillance or activity recognition.

A key aspect of the figure is the red horizontal line representing the anomaly score threshold. When the anomaly score curve exceeds this threshold, an abnormal event is flagged, and the specific video frames where this occurs are highlighted. The method thus offers a robust mechanism for identifying deviations from normal behavior in real-time video analysis, providing both clear detection and temporal localization. By setting a clearly defined threshold and marking frames that surpass it, the system ensures an efficient and precise detection of anomalies. This is evident from the red-shaded regions in the figure, which correlate well with abnormal events, further affirming the method's reliability. This visualization offers a transparent and accessible means to track the model's performance in identifying anomalies. HiTESS achieves 7.87 ms/sequence inference time with batch processing (batch size = 32), enabling real-time performance at 127 FPS–well beyond the 30 FPS requirement for surveillance systems. This highlights the efficiency gains from GPU parallelization and demonstrates the framework's potential for practical deployment in real-time anomaly detection systems.

The method leverages the model's ability to detect anomalies in a straightforward and efficient manner, suitable for analyzing complex video data.

4.3.1. Quantitative performance

Following previous methods in video anomaly detection (Feng et al., 2021a; Liu & Ma, 2019; Liu et al., 2018; Sultani et al., 2018; Wan et al., 2020), we adopted a weakly supervised learning framework for training our models on both the UCF-Crime and XD-Violence datasets. Weak supervision allows the use of video-level annotations, where only high-level labels (anomalous or normal) are provided without frame-level specificity. This approach is not only practical for large-scale datasets but also essential, as obtaining frame-level labels can be prohibitively expensive and time-consuming.

The models compared in this study were selected because they also operate under weak supervision. This ensures a fair evaluation across models that are all learning to localize and detect anomalous events without detailed, frame-level guidance. As presented in Table 1, we compare our approach to other state-of-the-art methods using Frame-Level AUC as the primary metric for the UCF-Crime dataset. AUC measures how well the model distinguishes between normal and anomalous frames across the video.

For the XD-Violence dataset, we evaluated our model using the Average Precision (AP) metric as proposed in Wu et al. (2020), Yang et al. (2024), as shown in Table 2, to quantify the effectiveness of our approach in detecting anomalous events across a wide range of scenarios.

By aligning with the weak supervision paradigm, we ensure that our comparisons are consistent with prior methods and demonstrate how different modeling choices affect the ability to generalize from coarse labels to fine-grained anomaly detection. The performance of each model is closely tied to the type of features used for anomaly detection.

C3D-RGB features extract both spatial and temporal information from video frames using 3D convolutional neural networks.

TSN-RGB and TSN-Flow use Temporal Segment Networks to capture long-term temporal dependencies from RGB frames or optical flow, respectively. I3D-RGB inflates standard 2D convolutions to 3D, making it effective in capturing detailed spatiotemporal patterns from RGB video

Table 1
Comparative analysis of Frame-Level AUC with other benchmark methods under weakly supervised learning on UCF-Crime dataset.

Method	Feature	AUC(%)
Sultani et al. (2018)	C3D-RGB	75.41
TCN-IBL (Zhang et al., 2019)	C3D-RGB	78.66
GCN (Zhong et al., 2019)	TSN-Flow	78.08
GCN (Zhong et al., 2019)	C3D-RGB	81.08
GCN (Zhong et al., 2019)	TSN-RGB	82.12
MIST (Feng et al., 2021a)	C3D-RGB	81.40
MIST (Feng et al., 2021a)	I3D-RGB	82.30
RTFM (Tian et al., 2021)	C3D-RGB	83.28
RTFM (Tian et al., 2021)	I3D-RGB	84.03
BN-WVAD (Yi et al., 2022)	I3D-RGB	84.29
CRFD (Wu & Liu, 2021)	I3D-RGB	84.89
MSL (Li et al., 2022)	I3D-RGB	85.30
NG-MIL (Park et al., 2023)	I3D-RGB	85.63
GLFE (Basak & Gautam, 2024)	I3D-RGB	86.12
MGFN (Chen et al., 2023)	VS-RGB	86.67
MGFN (Chen et al., 2023)	I3D-RGB	86.98
ST-HTAM (Paulraj & Vairavasundaram, 2025)	Transformer	81.42
TDS-Net (Hussain et al., 2024)	I3D	84.5
Ours	**I3D-RGB**	**88.30**

Table 2
Comparative analysis of Average Precision (AP) performance with other benchmark methods under weakly supervised learning on XD-Violence dataset.

Method	Feature	AP(%)
Wu et al. (2020)	I3D-RGB	78.64
RTFM (Tian et al., 2021)	I3D-RGB	77.81
MSL (Li et al., 2022)	I3D-RGB	78.28
NG-MIL (Park et al., 2023)	I3D-RGB	78.51
CRFD (Wu & Liu, 2021)	I3D-RGB	75.90
MGFN (Chen et al., 2023)	I3D-RGB	79.19
MGFN (Chen et al., 2023)	VS-RGB	80.11
MGFN (Chen et al., 2023)	VS-RGB	80.11
ST-HTAM (Paulraj & Vairavasundaram, 2025)	Transformer	78.06
Ours	**I3D-RGB**	**84.91**

Table 3
Evaluation performance analysis of **Our Approach** on UCF-Crime and XD-Violence datasets.

Metric	UCF-CRIME(%)	XD-Violence(%)
Accuracy	80.68	83.95
Precision	78.00	84.91
Recall	83.57	83.95
Specificity	78.00	86.46
F1-Score	80.68	84.13
AUC	88.30	91.41

frames. Lastly, VS-RGB focuses on spatial features from RGB frames to detect anomalies in static images or frame transitions.

These features play a pivotal role in the models' abilities to detect and localize anomalies effectively. As seen in Tables 1 and 2, the I3D-RGB features, used in several state-of-the-art approaches, including our own, consistently yield higher AUC and AP scores on both datasets, demonstrating their robustness in capturing complex spatiotemporal patterns in video data.

4.3.2. Comparative analysis of model performance metrics

In Table 3 and Fig. 3, we presented a comparative evaluation metric performance analysis of our model on both datasets. The metrics include Accuracy, Precision, Recall, Specificity, F1-Score, and AUC (Area Under the Curve), which together provide a detailed assessment of the model's performance.

On the UCF-Crime dataset, our model achieves an accuracy of 80.68 %, precision of 78.00 %, recall of 83.57 %, and an F1-score of 80.68 %, with specificity of 78.00 % and AUC of 88.30 %, demonstrating strong discriminatory power. For XD-Violence, the model shows consistent performance with 83.95 % accuracy, 83.95 % recall, 84.91 % precision, an F1-score of 84.13 %, 86.46 % specificity, and an AUC of 91.41 %, indicating high effectiveness in violence detection.

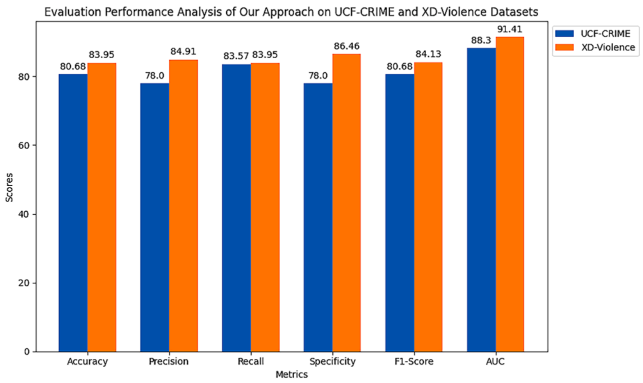

Fig. 3. Evaluation metric performance analysis on UCF-Crime and XD-Violence datasets.

This evaluation showcases the generalizability and robustness of our approach across different datasets, with strong performance in both anomaly detection and classification.

4.4. Ablation studies

To further understand the contribution of the individual components of our proposed HiTESS architecture, we conducted a series of ablation studies. The goal was to isolate and assess the impact of each major architectural component. Throughout these experiments, all hyperparameters, including learning rate, batch size, step size, overlap flag, segment length, and the number of units, were held constant to ensure consistency and reliability in the results. To verify the effectiveness of the complete HiTESS model, we systematically constructed and evaluated three baseline models by progressively removing or modifying specific architecture components. The complete HiTESS model serves as the benchmark for comparison. This model includes all three core components: BiLSTM layers, Multi-Head Attention, and the custom HiTESS scoring mechanism. The baseline models and their corresponding analyses are as follows:

- **HiTESS without BiLSTM**: In this model, we removed the BiLSTM while retaining both the Multi-Head Attention and the HiTESS scoring mechanism. The goal was to assess the importance of BiLSTM in capturing temporal dependencies within the input sequences.This model is denoted as Model A in Table 4.
- **HiTESS without Multi-Head Attention**: The Multi-Head Attention mechanism was excluded for this variant, leaving the BiLSTM and the HiTESS intact. This baseline tests the contribution of the attention mechanism in modeling long-range dependencies between elements within the input sequence.This model is denoted as Model B in Table 4.
- **BiLSTM with Multi-Head Attention**: In this final baseline, the HiTESS scoring mechanism was replaced with a simpler global pooling operation max(average) pooling represented as Model C in Table 4. Both BiLSTM and Multi-Head Attention remained part of the architecture, allowing us to isolate and measure the contribution of the HiTESS scoring strategy to the overall performance.

The results demonstrated that the HiTESS scoring mechanism plays a crucial role in enhancing model performance. As seen in Model C, replacing HiTESS with simpler global pooling (such as max or average) resulted in a performance drop. This highlights the importance of HiTESS's ability to capture fine-grained, segment-level patterns across multiple temporal scales. This segmentation approach is critical for effectively

Table 4
Ablation study results for different model configurations on the UCF-Crime dataset. **HiTESS**: Full model with all components, **Model A**: HiTESS w/o BiLSTM, **Model B**: HiTESS w/o Multi-Head Attention, **Model C**: BiLSTM + Multi-Head Attention (No HiTESS). All models were evaluated on the UCF-Crime dataset.

Metric	HiTESS(%)	ModelA(%)	ModelB(%)	ModelC(%)
Accuracy	80.68	80.00	77.24	72.41
Precision	78.00	78.08	76.42	65.15
Recall	83.57	81.42	76.42	92.14
Specificity	78.00	78.66	78.00	54.00
F1-Score	80.68	79.72	76.42	76.33
AUC-ROC	88.30	81.89	81.55	79.94

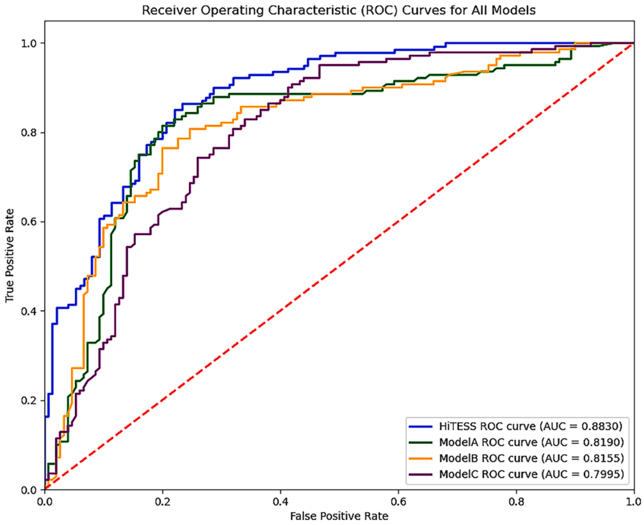

Fig. 4. Combined ROC curve for the full HiTESS model and its ablated variants (Model A, Model B, and Model C). The ROC curve highlights the model's performance in terms of true positive and false positive rates, with AUC values corresponding to each configuration as listed in Table 4.

detecting complex events, particularly in video anomaly detection tasks where temporal precision is essential. HiTESS, by providing robust temporal segment assessment, contributes significantly to the model's superior ability to detect subtle anomalies.

In these ablation experiments, we systematically assessed the performance of each model configuration on the UCF-Crime dataset. The results of these experiments are summarized in Table 4. The comparison between the full HiTESS model and the ablated versions provides insights into the relative importance of each architectural component. It was observed that removing any single component led to a degradation in performance, demonstrating the complementary nature of the BiLSTM layers, the Multi-Head Attention mechanism, and the HiTESS. The HiTESS model has significantly outperformed the three baseline models used for the ablation studies. Fig. 4 presents the combined ROC curve for the full HiTESS model and its ablated variants, illustrating the model's discriminative capability across different configurations based on their respective AUC scores.

Table 5
Impact of pooling functions on HiTESS model performance.

Metrics	UCF-Crime(%)			XD-Violence(%)		
Pooling	Avg	Max	Attn	Avg	Max	Attn
Accuracy	80.68	70.68	76.55	83.95	73.95	80.75
Precision	78.00	70.78	76.56	84.91	74.91	81.65
Recall	83.57	70.68	76.55	83.95	73.95	80.75
Specificity	78.00	69.33	79.33	86.46	76.46	83.46
F1-Score	80.68	70.69	76.52	84.13	74.13	80.92
AUC	88.30	77.96	82.48	91.41	81.41	87.96

Table 6
Joint effect of hierarchy depth and attention configuration on the UCF-Crime dataset.

Hierarchy depth	Attention heads	AUC (%)	Inference time (ms)
3	2	87.1	12.3
6	2	87.9	15.1
3	4	88.3	14.2
6	4	**89.5**	20.3

Table 7
Impact of pooling mechanisms across different attention head configurations.

Attention heads	Pooling	UCF-Crime (AUC %)	XD-Violence (AP %)
2	Average	86.40	89.12
2	Max	84.01	87.05
2	Attention	85.20	88.34
4	Average	**88.30**	**91.41**
4	Max	85.70	88.60
4	Attention	86.92	90.03

To further investigate the role of different pooling mechanisms within the HiTESS model, we conducted a study evaluating the effectiveness of average, max, and attention pooling. These pooling functions were combined with the HiTESS scoring mechanism, and the model's performance was assessed while retaining the BiLSTM and Multi-Head Attention layers. The results, summarized in Table 5, revealed that HiTESS with average pooling provided the most balanced and consistent results, effectively capturing both broad and detailed temporal patterns. Max pooling led to a performance decline due to its inability to capture subtle features, while attention pooling performed similarly to average pooling but did not fully leverage HiTESS's fine-grained temporal segmentation (Table 8).

Overall, HiTESS paired with average pooling was identified as the best approach for anomaly detection tasks, offering superior performance across multiple metrics such as accuracy, precision, and AUC-ROC.

To evaluate the combined impact of hierarchy depth and attention configuration, we conducted joint experiments summarized in Table 6. The results show that both increased depth and attention heads contribute to improved performance. Notably, the combination of 4 attention heads and deeper hierarchies yields higher AUC scores, though at the cost of increased inference time. For instance, while 6 layers with 4 heads produce the highest AUC (89.5 %), they also raise inference latency to 20.3 ms. In contrast, 3-layer configurations with 4-heads provide a better balance for deployment scenarios.

We further analyzed the interaction between pooling mechanisms and attention configurations. As shown in Table 7, average pooling consistently outperforms both max and attention pooling across 2-head and 4-head configurations. This suggests that pooling effectiveness is influenced by attention settings and reinforces the conclusion that average pooling paired with 4 attention heads is the most effective configuration.

Taken together, these results demonstrate that optimal performance is achieved by a synergistic configuration consisting of 4 attention heads, average pooling, and moderate hierarchical depth. This setup maximizes anomaly detection accuracy while maintaining computational efficiency.

4.5. Rationale for baseline method selection

The baseline methods were chosen to benchmark HiTESS against standard approaches in temporal modeling and anomaly detection. By removing either BiLSTM or Multi-Head Attention, we can isolate the

Table 8
Computational efficiency of HiTESS under batch inference. Real-time capability is evaluated based on the 30 FPS threshold (i.e., 33.3 ms per frame).

Metric	UCF-Crime	XD-Violence
FLOPs	5.42 GFLOPs	5.95 GFLOPs
Inference time (batch)	7.87 ms/sequence	8.21 ms/sequence
Model size	45.67 MB	46.80 MB
Real-t ime	Yes	Yes

importance of each module in capturing temporal dependencies and refining long-range features. This helps us understand their individual contributions to the overall model performance.

Additionally, replacing the HiTESS scoring mechanism with traditional global pooling (e.g., max/average pooling) allows us to evaluate the advantages of hierarchical multi-scale aggregation. This comparison highlights the impact of HiTESS's scoring on improving anomaly detection over conventional pooling techniques.

These baselines were selected to validate how each component–BiLSTM, Multi-Head Attention, and hierarchical scoring–works together to enhance the model's ability to capture temporal dynamics and improve anomaly detection accuracy.

4.6. *Computational efficiency*

Beyond accuracy and localization performance, we analyze the computational efficiency of the proposed HiTESS framework to assess its applicability in real-time surveillance systems. All experiments were conducted using an NVIDIA GTX 1050 Ti GPU with 32GB system memory.

HiTESS achieves an average inference speed of 127 frames per second (FPS) and operates with a computational complexity of approximately 5.42 GFLOPs per video sequence as shown in Table 8.This efficiency is primarily enabled by the hierarchical segmentation strategy, which avoids redundant temporal processing, and the use of a lightweight BiLSTM backbone instead of heavier transformer-based architectures.

While most recent methods such as RTFM (Tian et al., 2021) and WSAL (Lv et al., 2021) do not report computational metrics such as FLOPs or inference latency, we provide this analysis to highlight the real-time feasibility of HiTESS. With an inference speed of 127 FPS and lightweight design, HiTESS offers a favorable balance between efficiency and accuracy, reinforcing its practical value for real-time anomaly detection in surveillance scenarios.

5. Conclusion

In this work, we introduced Hierarchical Temporal Sequence Selection (HiTESS) to tackle the challenge of temporal anomaly localization in videos with weak labels. HiTESS employed a multi-level temporal segmentation framework that effectively addressed the challenges posed by limited labeled data, temporal localization, and the dynamic nature of anomalies. By integrating Bidirectional LSTM networks with Multi-Head Attention mechanisms, our method captured and represented temporal features more precisely, leading to improved detection and localization of subtle frame-level anomalies.

Experimental results on the UCF-Crime and XD-Violence datasets show that HiTESS outperforms existing methods, improving AUC by 1.32 % and AP by 4.8 %. While the method meets real-time requirements under batch inference, future work will explore further latency optimization for low-resource environments such as edge devices.

These results highlight its ability to reduce detection errors and enhance accuracy by effectively capturing and localizing anomalies across varied temporal scales in complex video sequences. The hierarchical sequence processing and scoring method refines anomaly localization by combining local and global evaluations, significantly reducing false positives. This approach advances video anomaly detection by setting a new standard for precision and robustness, effectively overcoming the limitations of traditional multiple-instance learning techniques.

Despite the lack of computational metrics reported in most existing WSAD benchmarks, we provide FLOPs and latency evaluations to promote reproducibility and highlight HiTESS's suitability for real-time scenarios. Our analysis demonstrates that HiTESS offers a favorable balance between accuracy and efficiency, making it practical for deployment in latency-sensitive environments.

In future work, we plan to extend HiTESS by incorporating multimodal features (e.g., audio, textual captions), integrating transformer-based temporal modeling for better generalization, and exploring its application in open-set and continual anomaly detection scenarios. We also aim to optimize HiTESS for low-resource environments, including deployment on edge devices.

CRediT authorship contribution statement

Nuku Atta Kordzo Abiew: Conceptualization, Methodology, Software, Investigation, Visualization, Writing – original draft; **Lijian Gao:** Writing – review & editing; **Qirong Mao:** Supervision, Methodology, Validation, Writing – review & editing, Funding acquisition.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Acknowledgements

This work was supported by the General Project of National Nature Science Foundation of China, Grant/Award Number: 62176106; the Jiangsu Key Research and Development Plan, Grant/Award Number: BE2020036; the Key Project of the National Science Foundation of China, Grant/Award Number: U1836220; the Project of Faculty of Agricultureal Engineering of Jiangsu University, Grant/Award Number: NGXB20240101; and the Special Scientific Research Project of School of Emergency Management of Jiangsu University, Grant/Award Number: KY-A-01.

References

Abdalla, M., Javed, S., Radi, M. A., Ulhaq, A., & Werghi, N. (2024). Video anomaly detection in 10 years: A survey and outlook. https://arxiv.org/abs/2405.19387.

Baradaran, M., & Bergevin, R. (2023). Multi-task learning based video anomaly detection with attention. In *2023 IEEE/CVF Conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 2886–2896). Los Alamitos, CA, USA: IEEE Computer Society. https://doi.org/10.1109/CVPRW59228.2023.00290

Basak, S., & Gautam, A. (2024). Diffusion-based normality pre-training for weakly supervised video anomaly detection. *Expert Systems with Applications*, *251*, 124013. https://doi.org/10.1016/j.eswa.2024.124013

Benjdira, B., Koubaa, A., Azar, A. T., Khan, Z., Ammar, A., & Boulila, W. (2022). Tau: A framework for video-based traffic analytics leveraging artificial intelligence and unmanned aerial systems. *Engineering Applications of Artificial Intelligence*, *114*, 105095. https://doi.org/10.1016/j.engappai.2022.105095

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on computer vision and pattern recognition (CVPR)* (pp. 4724–4733). https://doi.org/10.1109/CVPR.2017.502

Chen, Y., Liu, Z., Zhang, B., Fok, W., Qi, X., & Wu, Y.-C. (2023). Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the thirty-seventh AAAI conference on artificial intelligence and thirty-fifth conference on innovative applications of artificial intelligence and thirteenth symposium on educational advances in artificial intelligence* AAAI'23/IAAI'23/EAAI'23. AAAI Press. https://doi.org/10.1609/aaai.v37i1.25112

Cheng, K. W., Chen, Y. T., & Fang, W. H. (2015). Video anomaly detection and localization using hierarchical feature representation and gaussian process regression. In *2015 IEEE Conference on computer vision and pattern recognition (CVPR)* (pp. 2909–2917). https://doi.org/10.1109/CVPR.2015.7298909

Feng, J. C., Hong, F. T., & Zheng, W. S. (2021a). Mist: Multiple instance self-training framework for video anomaly detection. In *2021 IEEE/CVF Conference on computer vision and pattern recognition (CVPR)* (pp. 14004–14013). https://doi.org/10.1109/CVPR46437.2021.01379

Feng, X., Song, D., Chen, Y., Chen, Z., Ni, J., & Chen, H. (2021b). Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection. In *Proceedings of the 29th ACM international conference on multimedia* MM '21 (p. 5546-5554). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3474085.3475693

Gao, L., Zhou, L., Mao, Q., & Dong, M. (2022). Adaptive hierarchical pooling for weakly-supervised sound event detection. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 1779–1787).

Ghadiya, A., Kar, P., Chudasama, V., & Wasnik, P. (2024). Cross-modal fusion and attention mechanism for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops* (pp. 1965–1974).

Himeur, Y., Al-Maadeed, S., Kheddar, H., Al-Maadeed, N., Abualsaud, K., Mohamed, A., & Khattab, T. (2023). Video surveillance using deep transfer learning and deep domain adaptation: Towards better generalization. *Engineering Applications of Artificial Intelligence*, *119*, 105698. https://doi.org/10.1016/j.engappai.2022.105698

Huang, C., Liu, C., Wen, J., Wu, L., Xu, Y., Jiang, Q., & Wang, Y. (2022). Weakly supervised video anomaly detection via self-guided temporal discriminative transformer. *IEEE Transactions on Cybernetics*, *54*(5), 3197–3210.

Huang, X., Zhao, C., & Wu, Z. (2023). A video anomaly detection framework based on appearance-motion semantics representation consistency. In *ICASSP 2023 - 2023 IEEE International conference on acoustics, speech and signal processing (icassp)* (pp. 1–5). https://doi.org/10.1109/ICASSP49357.2023.10097199

Hussain, A., Ullah, W., Khan, N., Khan, Z. A., Kim, M. J., & Baik, S. W. (2024). Tds-net: Transformer enhanced dual-stream network for video anomaly detection. *Expert Systems with Applications*, *256*, 124846. https://doi.org/10.1016/j.eswa.2024.124846

Ionescu, R. T., Khan, F. S., Georgescu, M. I., & Shao, L. (2018). Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 7834–7843). https://api.semanticscholar.org/CorpusID:54475483.

Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., & Pedrycz, W. (2024). A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications*, *241*, 122666. https://doi.org/10.1016/j.eswa.2023.122666

Kamoona, A. M., Gostar, A. K., Bab-Hadiashar, A., & Hoseinnezhad, R. (2023). Multiple instance-based video anomaly detection using deep temporal encoding-decoding. *Expert Systems with Applications*, *214*, 119079. https://doi.org/10.1016/j.eswa.2022.119079

Kaneko, Y., Miah, A.S.M., Hassan, N., Lee, H. S., Jang, S. W., & Shin, J. (2024). Multimodal attention-enhanced feature fusion-based weekly supervised anomaly violence detection. 2409.11223.

Li, S., Liu, F., & Jiao, L. (2022). Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *AAAI Conference on artificial intelligence*. https://api.semanticscholar.org/CorpusID:248982052.

Liu, K., & Ma, H. (2019). Exploring background-bias for anomaly detection in surveillance videos. In *Proceedings of the 27th ACM international conference on multimedia* MM '19 (p. 1490-1499). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3343031.3350998

Liu, W., Luo, W., Lian, D., & Gao, S. (2018). Future frame prediction for anomaly detection - A new baseline. In *2018 IEEE/CVF Conference on computer vision and pattern recognition* (pp. 6536–6545). https://doi.org/10.1109/CVPR.2018.00684

Liu, Y., Yang, D., Wang, Y., Liu, J., Liu, J., Boukerche, A., Sun, P., & Song, L. (2024). Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. *ACM Computing Surveys*, *56*(7). https://doi.org/10.1145/3645101

Liu, Z., Nie, Y., Long, C., Zhang, Q., & Li, G. (2021). A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *2021 IEEE/CVF International conference on computer vision (ICCV)* (pp. 13568–13577). https://doi.org/10.1109/ICCV48922.2021.01333

Lv, H., Yue, Z., Sun, Q., Luo, B., Cui, Z., & Zhang, H. (2023). Unbiased multiple instance learning for weakly supervised video anomaly detection. In *2023 IEEE/CVF Conference on computer vision and pattern recognition (CVPR)* (pp. 8022–8031). https://doi.org/10.1109/CVPR52729.2023.00775

Lv, H., Zhou, C., Cui, Z., Xu, C., Li, Y., & Yang, J. (2021). Localizing anomalies from weakly-labeled videos. *IEEE Transactions on Image Processing*, *30*, 4505-4515. https://doi.org/10.1109/TIP.2021.3072863

Park, H., Noh, J., & Ham, B. (2020). Learning memory-guided normality for anomaly detection. In *2020 IEEE/CVF Conference on computer vision and pattern recognition (CVPR)* (pp. 14360–14369). https://doi.org/10.1109/CVPR42600.2020.01438

Park, S., Kim, H., Kim, M., Kim, D., & Sohn, K. (2023). Normality guided multiple instance learning for weakly supervised video anomaly detection. In *2023 IEEE/CVF Winter conference on applications of computer vision (WACV)* (pp. 2664–2673). https://doi.org/10.1109/WACV56688.2023.00269

Paulraj, S., & Vairavasundaram, S. (2025). Transformer-enabled weakly supervised abnormal event detection in intelligent video surveillance systems. *Engineering Applications of Artificial Intelligence*, *139*, 109496.

Ristea, N. C., Madan, N., Ionescu, R. T., Nasrollahi, K., Khan, F. S., Moeslund, T. B., & Shah, M. (2022). Self-supervised predictive convolutional attentive block for anomaly detection. In *2022 IEEE/CVF Conference on computer vision and pattern recognition (CVPR)* (pp. 13566–13576). https://doi.org/10.1109/CVPR52688.2022.01321

Satya Krishna, N., Nagesh Bhattu, S., Somayajulu, D. V. L. N., Narendra Kumar, N. V., & Jaya Shankar Reddy, K. (2022). GssMILP for anomaly classification in surveillance videos. *Expert Systems with Applications*, *203*, 117451. https://doi.org/10.1016/j.eswa.2022.117451

Singh, R., Sethi, A., Saini, K., Saurav, S., Tiwari, A., & Singh, S. (2024). Attention-guided generator with dual discriminator GAN for real-time video anomaly detection. *Engineering Applications of Artificial Intelligence*, *131*, 107830. https://doi.org/10.1016/j.engappai.2023.107830

Srivastava, N., Mansimov, E., & Salakhutdinov, R. (2015). Unsupervised learning of video representations using LSTMs. In *Proceedings of the 32nd international conference on international conference on machine learning - volume 37* ICML'15 (p. 843-852). JMLR.org.

Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on computer vision and pattern recognition* (pp. 6479–6488). https://doi.org/10.1109/CVPR.2018.00678

Sun, S., & Gong, X. (2024). Multi-scale bottleneck transformer for weakly supervised multimodal violence detection. In *2024 IEEE International conference on multimedia and expo (ICME)* (pp. 1–6). IEEE.

Sun, S., Hua, J., Feng, J., Wei, D., Lai, B., & Gong, X. (2025). Delving into instance modeling for weakly supervised video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Sun, W., Cao, L., Guo, Y., & Du, K. (2023). Cross-modal attention mechanism for weakly supervised video anomaly detection. In *Biometric recognition: 17th Chinese conference, CCBR 2023, Xuzhou, China, December 1–3, 2023, proceedings* (p. 437–446). Berlin, Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-981-99-8565-4_41

Tang, S., Li, C., Zhang, P., & Tang, R. (2023). SwinLSTM: Improving spatiotemporal prediction accuracy using swin transformer and LSTM. In *2023 IEEE/CVF International conference on computer vision (ICCV)* (pp. 13424–13433). https://doi.org/10.1109/ICCV51070.2023.01239

Thakare, K. V., Sharma, N., Dogra, D. P., Choi, H., & Kim, I. J. (2022). A multi-stream deep neural network with late fuzzy fusion for real-world anomaly detection. *Expert Systems with Applications*, *201*, 117030. https://doi.org/10.1016/j.eswa.2022.117030

Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J. W., & Carneiro, G. (2021). Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *2021 IEEE/CVF International conference on computer vision (ICCV)* (pp. 4955–4966). https://doi.org/10.1109/ICCV48922.2021.00493

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *2015 IEEE International conference on computer vision (ICCV)* (pp. 4489–4497). https://doi.org/10.1109/ICCV.2015.510

Ullah, W., Hussain, T., Ullah, F. U. M., Lee, M. Y., & Baik, S. W. (2023a). TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection. *Engineering Applications of Artificial Intelligence*, *123*, 106173. https://doi.org/10.1016/j.engappai.2023.106173

Ullah, W., Min Ullah, F. U., Ahmad Khan, Z., & Wook Baik, S. (2023b). Sequential attention mechanism for weakly supervised video anomaly detection. *Expert Systems with Applications*, *230*, 120599. https://doi.org/10.1016/j.eswa.2023.120599

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., & Polosukhin, I. (2017). Attention is all you need. In I.Guyon, U.V. Luxburg, S.Bengio, H.Wallach, R.Fergus, S.Vishwanathan, & R.Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. (*vol. 30*). https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wan, B., Fang, Y., Xia, X., & Mei, J. (2020). Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE International conference on multimedia and expo (ICME)* (pp. 1–6). https://doi.org/10.1109/ICME46284.2020.9102722

Wang, Y., Gao, Z., Long, M., Wang, J., & Yu, P. S. (2018). PredRNN + +: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *International conference on machine learning*. https://api.semanticscholar.org/CorpusID:4946865.

Wang, Y., Jiang, L., Yang, M. H., Li, L. J., Long, M., & Fei-Fei, L. (2019a). Eidetic 3D LSTM: A model for video prediction and beyond. In *International conference on learning representations*. https://api.semanticscholar.org/CorpusID:86785011.

Wang, Y., Zhang, J., Zhu, H., Long, M., Wang, J., & Yu, P. S. (2019b). Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In *2019 IEEE/CVF Conference on computer vision and pattern recognition (CVPR)* (pp. 9146–9154). https://doi.org/10.1109/CVPR.2019.00937

Wang, Z., Zou, Y., & Zhang, Z. (2020). Cluster attention contrast for video anomaly detection. *Proceedings of the 28th ACM International Conference on Multimedia*, . https://api.semanticscholar.org/CorpusID:222278287.

Waqas, M., Tahir, M.A., Author, M. D., Al-Máadeed, S., Bouridane, A., & Wu, J. (2024). Simultaneous instance pooling and bag representation selection approach for multiple-instance learning (MIL) using vision transformer. *Neural Computing & Applications*, *36*, 6659–6680. https://api.semanticscholar.org/CorpusID:267732664.

Watanabe, Y., Okabe, M., Harada, Y., & Kashima, N. (2022). Real-world video anomaly detection by extracting salient features. In *2022 IEEE International conference on image processing (ICIP)* (pp. 891–895). https://doi.org/10.1109/ICIP46576.2022.9897864

Wen, X., Lai, H., Gao, G., Xiao, Y., Wang, T., Jia, Z., & Wang, L. (2023). Video anomaly detection based on cross-frame prediction mechanism and spatio-temporal memory-enhanced pseudo-3D encoder. *Engineering Applications of Artificial Intelligence*, *126*, 107057. https://doi.org/10.1016/j.engappai.2023.107057

Wu, J., Zhang, W., Li, G., Wu, W., Tan, X., Li, Y., Ding, E., & Lin, L. (2021). Weakly-supervised spatio-temporal anomaly detection in surveillance video. In *International joint conference on artificial intelligence*. https://api.semanticscholar.org/CorpusID:236098982.

Wu, P., & Liu, J. (2021). Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, *30*, 3513–3527. https://doi.org/10.1109/TIP.2021.3062192

Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., & Yang, Z. (2020). Not only look, but also listen: Learning multimodal violence detection under weak supervision. In A.Vedaldi, H.Bischof, T.Brox, & J.-M. Frahm (Eds.), *Computer vision – ECCV 2020* (pp. 322–339). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-58577-8_20.

Xu, D., Yan, Y., Ricci, E., & Sebe, N. (2017). Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, *156*, 117–127. Image and Video Understanding in Big Data https://doi.org/10.1016/j.cviu.2016.10.010

Yang, Z., Liu, J., & Wu, P. (2024). Text prompt with normality guidance for weakly supervised video anomaly detection. *2024 IEEE/CVF Conference on computer vision and pattern recognition (CVPR)*, (pp. 18899–18908). https://api.semanticscholar.org/CorpusID:269137657.

Yi, S., Fan, Z., & Wu, D. (2022). Batch feature standardization network with triplet loss for weakly-supervised video anomaly detection. *Image and Vision Computing*, *120*(C). https://doi.org/10.1016/j.imavis.2022.104397

Yu, G., Wang, S., Cai, Z., Liu, X., & Wu, C. (2022). Effective video abnormal event detection by learning a consistency-aware high-level feature extractor. In *Proceedings of the 30th ACM international conference on multimedia* MM '22 (p. 6337-6346). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3503161.3547944. https://doi.org/10.1145/3503161.3547944

Yu, J. H., Moon, J. H., & Sohn, K. A. (2023). Attention-guided residual frame learning for video anomaly detection. *Multimedia Tools and Applications*, *82*(8), 12099–12116. https://doi.org/10.1007/s11042-022-13643-z

Zaheer, M. Z., Mahmood, A., Khan, M. H., Segu, M., Yu, F., & Lee, S. I. (2022). Generative cooperative learning for unsupervised video anomaly detection. In *2022 IEEE/CVF Conference on computer vision and pattern recognition (CVPR)* (pp. 14724–14734). https://doi.org/10.1109/CVPR52688.2022.01433

Zhang, C., Chen, P., Lei, T., Wu, Y., & Meng, H. (2022). What-where-when attention network for video-based person re-identification. *Neurocomputing*, *468*, 33–47. https://doi.org/10.1016/j.neucom.2021.10.018

Zhang, C., Li, G., Qi, Y., Wang, S., Qing, L., Huang, Q., & Yang, M. H. (2023). Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *2023 IEEE/CVF Conference on computer vision and pattern recognition (CVPR)* (pp. 16271–16280). https://doi.org/10.1109/CVPR52729.2023.01561

Zhang, J., Qing, L., & Miao, J. (2019). Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. *2019 IEEE International Conference on Image Processing (ICIP)*, (pp. 4030–4034). https://api.semanticscholar.org/CorpusID:202787315.

Zhang, M., Wang, J., Qi, Q., Sun, H., Zhuang, Z., Ren, P., Ma, R., & Liao, J. (2024). Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 17385–17394).

Zhong, J.X., Li, N., Kong, W., Liu, S., Li, T.H., & Li, G. (2019). 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 (CVPR)*.

Zhong, Y., Zhu, R., Yan, G., Gan, P., Shen, X., & Zhu, D. (2024). Inter-clip feature similarity based weakly supervised video anomaly detection via multi-scale temporal MLP. *IEEE Transactions on Circuits and Systems for Video Technology*, .

Zhu, Y., & Newsam, S. (2019). Motion-aware feature for improved video anomaly detection. *ArXiv*, *1907.10211*. https://api.semanticscholar.org/CorpusID:198229642.