Weakly-Supervised Abnormal Event Detection in Surveillance Video

Dual-backbone fusion of TimesFormer and I3D Features with Multiple-Instance Learning (MIL)

Afeka College

Internet of Things — Final Project

Noam Tsfaty — 314952912 Liav Cohen — 209454693

July 20, 2025

Abstract

With the proliferation of IoT-enabled surveillance systems in modern cities, the volume of footage far exceeds human monitoring capabilities. Automatic abnormal event detection is therefore indispensable for public safety, yet it remains challenged by three enduring obstacles: anomalies are both rare and heterogeneous, precise frame-level annotation is prohibitively expensive, and effective detection demands modeling both rapid motions and extended temporal dependencies. In this paper, we introduce a weakly-supervised framework that relies exclusively on video-level normal/abnormal labels but achieves precise frame-level localization. Each video is uniformly divided into 32 segments. From each segment, we extract (i) a 768-dimensional **TimesFormer** embedding to capture long-range spatiotemporal patterns; and (ii) a 1 024-dimensional **I3D** descriptor encoding fine-grained motion cues. These are concatenated into a 1 792-dimensional feature vector, ℓ_2-normalized, and fed to a streamlined multiple-instance learning (MIL) head, where a single max-pooling attention layer—trained with a top-K multiple instance binary cross-entropy loss—highlights anomalous segments without requiring any frame-level supervision On the UCF-Crime dataset, our approach outperforms the previous state of the art by 15 AUC points (0.90 vs. 0.75), while its real-time design makes it well suited for deployment in resource-constrained IoT surveillance systems.

1 Introduction

Closed-circuit television (CCTV) networks now underpin the public-safety infrastructure of modern cities, shopping centres, and transport hubs. A mid-sized city may deploy on the order of 10^4 cameras[1] —producing far more footage than human operators can review. The rise of IoT-enabled edge devices further amplifies this data stream, imposing strict compute and bandwidth constraints on real-time analysis.
Automatic abnormal-event detection must therefore overcome three fundamental challenges

1. **Sparsity and heterogeneity.** Anomalous events occur infrequently and exhibit vast variation in appearance, motion, and duration. Models trained on a limited set of examples often fail to generalise to unseen behaviours or environments.

2. **Annotation bottleneck.** Precise frame- or pixel-level labelling costs minutes per second of video. Thus, most public benchmarks supply only a single *video-level* label $y \in \{0, 1\}$.

3. **Multi-scale temporal context.** Events such as traffic collisions occupy only a few frames, while others—e.g. loitering or preparation for theft—can span thousands. Capturing both rapid, local motion and long-range dependencies is essential.

The model takes as input a video X, which is temporally segmented into S non-overlapping fixed-length clips $\{x_1, x_2, \ldots, x_S\}$. These segments are processed individually to obtain semantically rich representations that preserve both spatial appearance and temporal motion characteristics. Each segment x_i is passed through two distinct backbone networks. One encoder is based on 3D convolutional operations that are sensitive to local motion cues and short-term dynamics. This encoder outputs a descriptor h_i, which captures temporally localized patterns—essential for detecting brief yet critical events like sudden running or abrupt changes in posture. The second encoder adopts a transformer-based architecture that models spatial and temporal dependencies using global attention mechanisms. It produces a feature t_i, which encodes long-range interactions and high-level semantic information, useful for identifying more contextual or slowly developing anomalies.
To unify the information from both sources, the two features h_i and t_i are first normalized and then concatenated into a single descriptor $v_i = [t_i, h_i]$ of fixed dimension d. This fusion step creates a representation that is simultaneously aware of localized dynamics and long-term scene structure, which is particularly advantageous for capturing the multiscale nature of anomalies in real-world scenarios. For instance, a person suddenly sprinting in a quiet area constitutes a brief anomaly, while loitering or trespassing may evolve over a longer temporal window. Both types of events can now be captured within the same learning framework through this feature unification.
Each segment-level vector v_i is passed to a segment scorer $f(v_i)$, which is implemented as a two-layer multilayer perceptron (MLP). The first layer applies a linear transformation with weights $W \in \mathbb{R}^{512\times d}$ and bias $b \in \mathbb{R}^{512}$, followed by a LeakyReLU nonlinearity with negative slope coefficient set to 0.1, allowing small gradient flow even when activation is

[1] "CCTV Surveillance Analyzed in America's Largest Cities," *Ventas de Seguridad*, 2023. https://www.ventasdeseguridad.com/en/news/latest-news/431-enterprises/24125-cctv-surveillance-analyzed-in-america-s-largest-cities.html

near zero. This design choice improves learning stability, particularly in weakly labeled settings. The second layer computes a weighted projection via $w \in \mathbb{R}^{512}$, yielding a scalar anomaly score $s_i \in \mathbb{R}$ for each segment.
At the video level, the final prediction is not computed by naïvely aggregating all segment scores. Instead, to account for the fact that most segments are normal and only a few exhibit anomalous behavior, the model selects the top-k segments with the highest anomaly scores $\{s_{i_1}, \ldots, s_{i_k}\}$. Their average forms a global video-level logit $\ell(X)$, which is passed through a sigmoid function to obtain a probability score $\hat{y} \in [0, 1]$. This prediction reflects the model's confidence that the input video contains an anomalous event. Training is conducted in a weakly supervised setting using binary video-level labels $y \in \{0, 1\}$, where 1 denotes the presence of an anomaly somewhere in the video. Crucially, no frame-level or segment-level annotations are available. The training objective is therefore based on the binary cross-entropy loss between the predicted score $\hat{y}$ and the ground truth label y. This formulation allows the model to learn discriminative segment-level scores indirectly by optimizing a video-level objective. This approach is particularly well-suited for large-scale anomaly detection benchmarks such as UCF-Crime, where each video may contain tens of thousands of frames, but annotations are limited to a single binary label. The combination of dense segment scoring, top-k pooling, and dual-pathway feature fusion enables the model to isolate anomalous segments and suppress noise, even in the absence of fine-grained supervision. Moreover, the architecture remains computationally lightweight due to the shallow MLP head and avoids overfitting by leveraging complementary priors from the transformer and convolutional pathways.

2 Related Work

A recent paradigm leverages weak video-level labels to train a deep multiple-instance ranking model that both isolates true anomalies and suppresses spurious activations through temporal regularization. In this framework [1], each video is split into fixed-length segments, C3D features are extracted, and a small MLP assigns an anomaly score to each segment. A hinge-based ranking loss then enforces that the highest-scoring segment in an anomalous bag exceeds that in a normal bag, while two regularization terms promote temporal smoothness—by penalizing large score differences between adjacent segments—and sparsity—by penalizing the sum of all segment scores—thus ensuring that only a few, temporally coherent segments light up in truly abnormal clips. This yields robust localization without any frame-level supervision, even when anomalies are brief or intermittently occluded. However, relying solely on a 3-D CNN backbone can limit the model's ability to capture long-range dependencies, and the non-differentiable hinge introduces gradient stagnation around the margin. To address these issues, our method fuses global, self-attention–based embeddings from a TimesFormer transformer with local, fine-grained motion cues from an I3D network into a single ℓ_2-normalized descriptor per segment. Empirically, this dual-backbone fusion tightens the separation between normal and abnormal bags and enhances frame-level detection accuracy in real-world surveillance benchmarks. A complementary effort exploits multi-scale temporal modeling and learnable "agent" tokens to mine context at varying durations and sharpen the distinction between normal and abnormal segments [2]. In this scheme, each video is processed at several temporal granularities: a multi-scale temporal attention module first captures context from short to long windows, then deformable convolutions generate a set of "time

agent" tokens that dynamically sample and emphasize salient temporal regions. These tokens are fed into a lightweight transformer-style agent network, which adaptively refines segment representations to pull apart incomplete or blurred anomalies from background activity. Building atop this, a robust temporal feature magnitude (RTFM) learning component ranks segments by the magnitude of their feature responses: by selecting the top-k largest magnitudes within an abnormal bag and enforcing a contrastive ranking loss against normal counterparts, it ensures discrete and intermittent anomalies are reliably isolated. While this architecture achieves state-of-the-art AUCs on UCSD-Ped2, CUHK Avenue, ShanghaiTech, and UCF-Crime, its multi-scale sampling and agent token mechanisms incur notable computational overhead. By contrast, our approach preserving both long-range self-attention and fine-grained motion priors without the added cost of dynamic token generation. Deep spatio-temporal autoencoders with LSTM combine motion and appearance[3], using reconstruction errors and active learning to refine anomaly detection. Optical flow–based autoencoders enhance motion sensitivity but often sacrifice pixel-level accuracy. Crowd-focused methods rely on low-dimensional optical flow descriptors and unsupervised classifiers for dense scenes, yet face high localization error when anomalies are sparse. Joint motion-based and dictionary learning models capture both local and global dynamics but struggle with computational overhead and scaling to high-resolution data. Spectral mapping improves efficiency through dimensionality reduction but performs poorly on noisy, high-res inputs. Generation error–based methods use unified anomaly thresholds for robustness but lack temporal precision. Overall, these approaches tend to prioritize either local cues or global temporal coherence, but not both, and all face challenges such as limited data, environmental noise, and efficiency—underscoring the need for models that better delineate normal from abnormal video behavior.

A hybrid approach[4] that combines transfer learning for efficient feature extraction (using pre-trained YOLO and Flownet2 to encode appearance, location, and motion) with a statistical kNN anomaly scoring module allows for rapid online detection, using only present and past information, without needing access to future frames or exhaustive retraining as required in most deep neural models. Unlike previous methods that privilege either local motion cues or global temporal coherence—but rarely both—the proposed system unites both by constructing composite feature vectors per detected object, capturing appearance, motion, and trajectory. Where crowd-focused and dictionary-based methods succeed mostly in dense, crowded scenes and struggle with sparsity, this framework handles both rare and frequent patterns, smoothly transitioning between few-shot and many-shot scenarios. This hybrid approach operates sequentially, with automatic threshold selection and robust statistical decision-making, making it inherently suited for real-time, online deployment. Despite these advances, the system still depends on the diversity and representativeness of the nominal training set, as well as the generalization capacity of the pre-trained neural detectors used for feature extraction. By contrast, our approach explicitly harmonizes local and global cues through joint modeling and enforces tighter decision boundaries, achieving a sharper separation between normal and abnormal patterns, and further improving detection robustness even when labeled data or pre-trained features are limited. A common approach in anomaly detection is to mask part of the input and have the model try to reconstruct what's missing. If the model's prediction is poor—meaning there is a large reconstruction error—it's a sign that the input might be abnormal. This is because the model has only learned to reconstruct normal data, so it struggles with anything unusual.SSPCAB (Self-Supervised Predictive

Convolutional Attentive Block)[5] takes this idea further by adding a special block inside a convolutional neural network (CNN). In CNNs, layers use filters to learn patterns in images or video. SSPCAB works by using masked, dilated convolutions that hide parts of the image in the middle of each filter's view. The network must then predict the missing center area using the surrounding context. This forces the model to understand both small details and the bigger picture in normal examples. Channel attention is also used, so the network learns which features are important, rather than making simple guesses.By penalizing (making the loss bigger) when the network's reconstruction is wrong, the model is trained to be very accurate for normal data. But if something abnormal appears, the error becomes high—this signals an anomaly. SSPCAB can be added to many existing anomaly detection models, and results show it improves detection and localization of anomalies on several datasets.However, if the network becomes too good at reconstructing, it might start to reconstruct even abnormal inputs, making it harder to spot anomalies. The choice of what to mask and how much weight to give the error also matters. Unlike this, our approach puts more focus on balancing local details and global patterns, and sets stricter rules for deciding what's normal, making detection more robust even with limited data.

3 Method

To sidestep dense supervision, we adopt a weakly-supervised Multiple Instance Learning (MIL) framework, and instead of relying solely on uniform segment splits, we perform a more refined preprocessing stage to ensure representative feature construction. Specifically, each video is decoded frame-by-frame, converted from BGR to RGB, and spatially resized to a standard resolution. We divide the entire video into 32 high-level segments, and from each segment, we extract a set of overlapping 16-frame sub-clips using a sliding window with fixed stride. For each sub-clip, we extract deep features using one of two pretrained spatiotemporal models: TimesFormer or I3D. The resulting features are ℓ_2-normalized and averaged across sub-clips within the same segment, producing a robust representation that encodes both short-term motion and long-range semantics. When a segment contains too few frames, we pad or default to a zero-vector of appropriate dimensionality. This preprocessing ensures consistent temporal coverage and allows both backbones to operate within their respective frame constraints, resulting in uniform and dense segment-level embeddings across the dataset.

The Multiple-Instance Learning (MIL) framework is particularly effective in video anomaly detection, where precise frame-level labels are often unavailable. In this setup, each video is treated as a "bag" of instances (segments), and only the label of the entire video (bag) is known. This means that if a video contains an anomaly, the model must discover which segments contributed to that label, without explicit segment annotations. This paradigm encourages the model to assign high anomaly scores only to the most suspicious segments and suppress scores for benign ones.

Lightweight MIL head. We implement the segment-scoring function

$$f : \mathbb{R}^d \to \mathbb{R}$$

as a two-layer MLP with LeakyReLU:

$$s_i = f(x_i) = w^\top \phi(W v_i + b), \quad \phi(u) = \text{LeakyReLU}(u; 0.1),$$

where

$$W \in \mathbb{R}^{512 \times d}, \quad b \in \mathbb{R}^{512}, \quad w \in \mathbb{R}^{512}.$$

The segment scorer is a small two-layer neural network that takes the fused feature vector v_i from each segment and outputs a scalar anomaly score s_i. The activation function used here is LeakyReLU with a negative slope of 0.1. This function helps prevent "dying neurons" by allowing a small gradient flow even for negative activations. The weights W, b, and w define the two linear layers and the final projection to a single scalar. The lightweight design ensures computational efficiency and avoids overfitting.

$$\text{video\_logit}(X) = \frac{1}{k} \sum_{i \in \text{TopK}(s)} s_i, \qquad \hat{y} = \sigma(\text{video\_logit}(X)),$$

and we train by minimizing the binary cross-entropy with logits:

$$\mathcal{L} = \text{BCEWithLogits}(\text{video\_logit}(X), y),$$

using only video-level normal/abnormal labels.
Top-k pooling is a core MIL mechanism that selects the k highest segment scores in the video. These are averaged to form the video-level anomaly score. This operator acts as a soft approximation for hidden segment labels, assuming that only a small number of segments in an anomalous video actually contain the anomaly. By using only the most suspicious segments, the model avoids being diluted by background content. The final prediction $\hat{y}$ is computed by applying a sigmoid to the pooled score. Training is performed using BCEWithLogits, which internally combines a sigmoid activation with binary cross-entropy loss, offering numerical stability and a direct mapping from logits to probability-based loss.

Dual-backbone fusion. We propose to fuse complementary spatiotemporal encoders within the MIL framework. In each segment x_i we compute:

$$\underbrace{t_i}_{\in \mathbb{R}^{768}} = \text{TimesFormer}(x_i), \qquad \underbrace{h_i}_{\in \mathbb{R}^{1024}} = \text{I3D}(x_i),$$

then ℓ_2–normalize and concatenate:

$$v_i = \frac{[t_i; h_i]}{\|[t_i; h_i]\|_2} \in \mathbb{R}^{1792}.$$

To capture both fine-grained motion cues and broader contextual semantics, we use a dual-backbone approach. I3D is a 3D convolutional network that excels at capturing local motion patterns by operating on spatiotemporal volumes. It detects sudden changes like running, falling, or waving. In contrast, TimesFormer is a transformer-based model that uses attention mechanisms to model global dependencies across time and space, capturing higher-level activities like loitering or group behavior.
Each backbone processes the segment independently and produces a descriptor: h_i from I3D and t_i from TimesFormer. These are concatenated and ℓ_2-normalized to form the unified segment representation v_i.

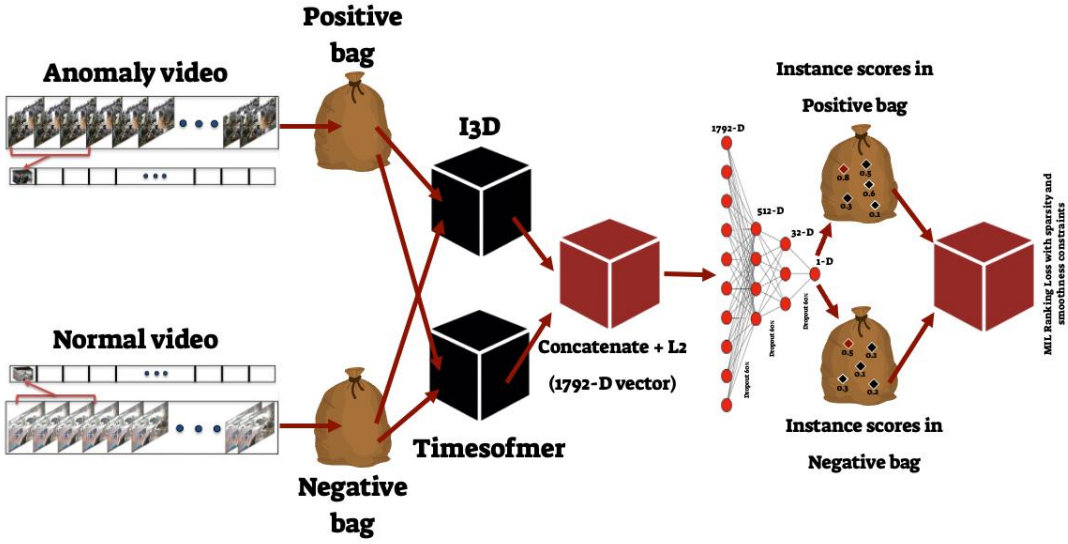


Figure 1: Architecture overview. Each video is divided into segments, encoded by two distinct spatiotemporal networks (I3D and TimesFormer), and scored by a shared MIL head. Top-k scores are aggregated into the video-level anomaly prediction.

This single descriptor unifies global self-attention with local 3D convolutional motion cues in a compact vector.
This design allows the model to address the central challenge in anomaly detection: the fact that abnormal events are rare, diverse, and context-dependent. By leveraging two encoders—one based on I3D and the other on TimesFormer—the model achieves both short-term temporal precision and long-range semantic reasoning. The I3D pathway captures sudden, motion-heavy deviations like running or falling, while TimesFormer contributes interpretive strength over prolonged and less kinetic activities like loitering or trespassing. The feature normalization ensures that neither stream dominates, and the concatenation creates a high-dimensional descriptor rich in multiscale semantics.
The top-k pooling operator serves as a weak surrogate for segment-level labels. Instead of averaging over all segment scores, it emphasizes the most salient segments—those most likely to exhibit anomalous traits. This mechanism improves discriminability under the weak supervision regime, allowing the model to focus on informative cues even without explicit temporal annotations.

4 Datasets

The dataset contains 1,900 untrimmed surveillance videos divided into normal and anomalous events. Anomalies span a wide spectrum, including categories such as "Assault", "Arson", "Robbery", and "Road Accidents". Each video is only weakly labeled, indicating whether an anomaly occurs somewhere in the video, without frame-level annotations. This reflects the practical difficulty of obtaining detailed temporal labels in real surveillance settings, motivating our weakly-supervised approach.

The distribution of anomaly types is summarized in Table 1. As seen, most anomaly classes include 50 labeled instances, with the exception of "RoadAccidents" and "Robbery", which each contain 150 videos. In contrast, the number of normal videos reaches 950, with 800 used for training and 150 for evaluation, highlighting the inherent imbalance and the necessity for robust learning techniques.

Table 1: Distribution of anomaly types and normal events in UCF-Crime dataset.

Anomaly Type	**Total Videos (Training)**
Abuse	50 (48)
Arrest	50 (45)
Arson	50 (41)
Assault	50 (47)
Burglary	100 (87)
Explosion	50 (29)
Fighting	50 (45)
RoadAccidents	150 (127)
Robbery	150 (145)
Shooting	50 (27)
Shoplifting	50 (29)
Stealing	100 (95)
Vandalism	50 (45)
Normal events	950 (800)

Figure 2 shows the distribution of video lengths in minutes. The majority of videos are shorter than 2 minutes, with a long tail of longer clips, some exceeding 10 minutes. This variability in temporal extent adds to the complexity of learning temporal dependencies under weak supervision.

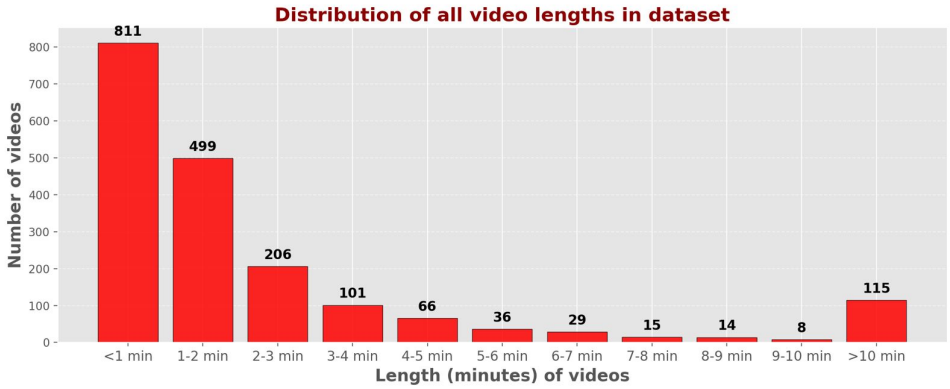


Figure 2: Histogram showing distribution of video durations across the dataset. Most videos are shorter than 2 minutes, with significant class imbalance in clip duration.

In Figure 3, we analyze the number of frames per testing video, revealing significant variation. Some videos contain just a few hundred frames, while others exceed 100,000. This heterogeneity demands scalable and efficient inference pipelines.

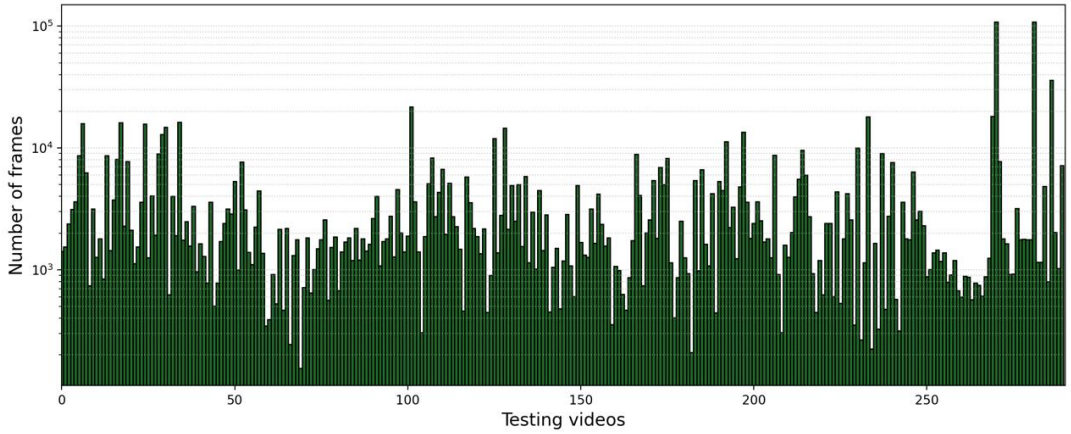


Figure 3: Log-scale distribution of the number of frames per video in the test set. Note the wide variability, with some videos exceeding 10^5 frames.

To further investigate the annotation quality, Figure 4 presents the fraction of anomalous frames in each test video that was labeled as containing an anomaly. The degree of sparsity is evident: many anomalous videos contain less than 30% of their frames depicting actual anomalies. This sparsity reinforces the importance of our model's design, which relies on top-k pooling to surface the most informative segments for classification.

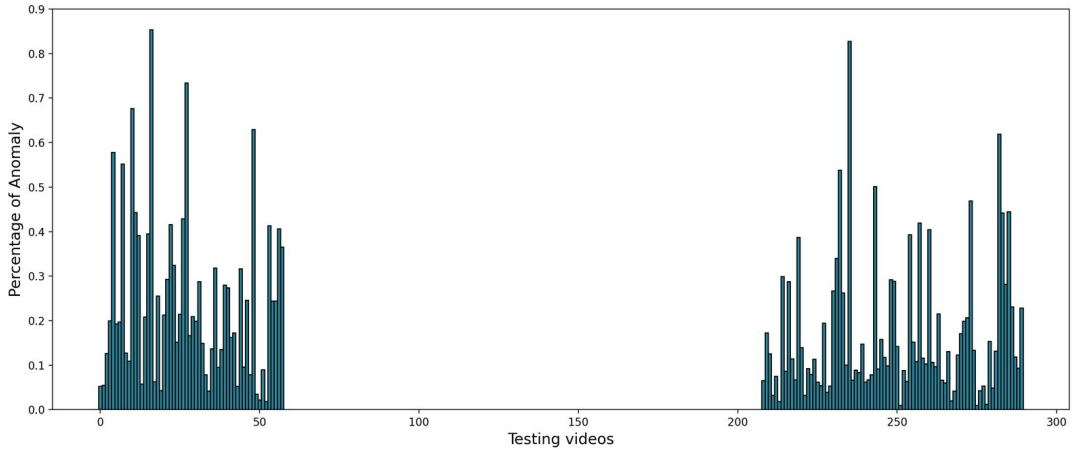


Figure 4: Proportion of anomalous frames within each anomalous testing video. Most anomalies are temporally sparse, often under 30% of the video duration.

5 Experiments

To empirically validate the effectiveness of our proposed dual-stream MIL-based anomaly detection framework, we conducted extensive evaluations on a benchmark dataset. The model was trained for 60 epochs using the Adam optimizer with a learning rate of 1×10^{-4}, dropout probability of 0.5, and batch size of one complete video sample. These hyperparameters were selected based on empirical stability and alignment with prior work in weakly-supervised anomaly detection.

A key parameter in the MIL setting is the number of top segments, k, used to aggregate instance-level anomaly scores into a global video-level prediction. Since the optimal value of k is data-dependent, we employed a greedy search strategy over $k \in \{1, 5, 10, 20, 30\}$ and chose the value that maximized validation ROC-AUC. Furthermore, we adopted a weighted loss formulation by assigning decaying weights to the top-k instance scores, i.e., $w = [1.0, 0.5, 0.25, 0.125, 0.125]$, to emphasize stronger anomalies while maintaining sensitivity to less prominent events.

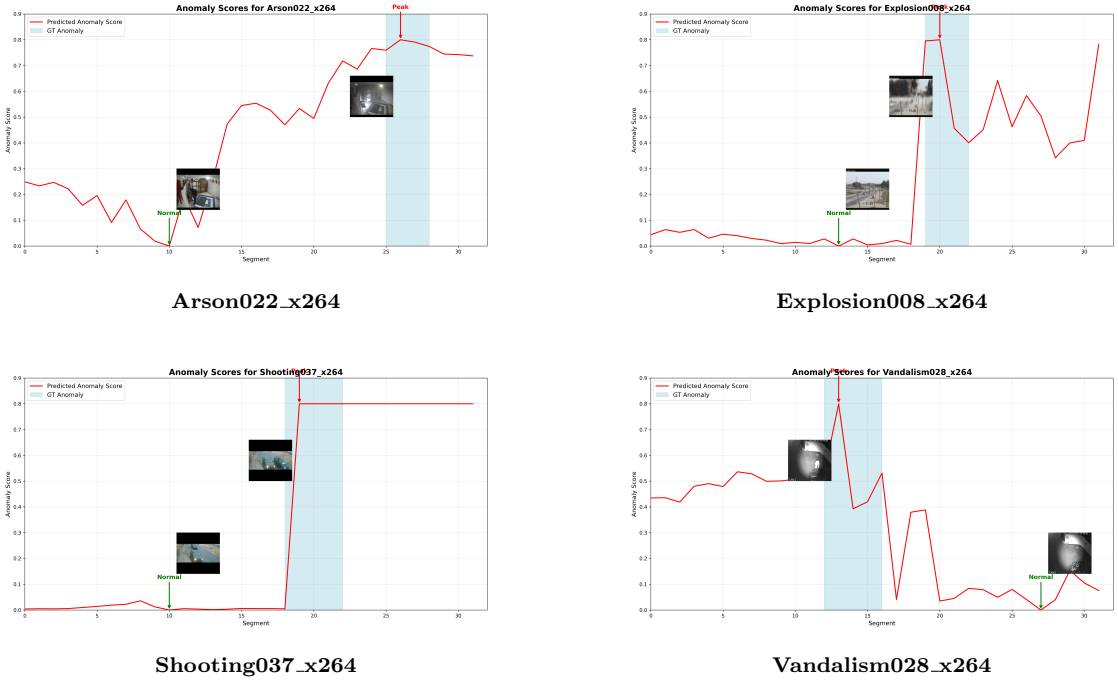


Figure 5: Predicted anomaly score curves with embedded video frames for four representative videos from UCF-Crime: Arson, Explosion, Shooting, and Vandalism. The red line denotes the predicted anomaly score over time (segment index), and the blue shaded region marks the ground truth anomaly segment. Each frame illustrates a critical peak or normal prediction location. All examples show strong correlation between prediction and GT labels, highlighting temporal precision and interpretability.

5.1 ROC and PR Curves

The ROC and PR curves in Figure 6 provide a continuous view of the model's discriminative ability across varying thresholds. The Area Under the ROC Curve (ROC-AUC) is 0.907, indicating excellent separation between normal and anomalous samples. The Precision-Recall AUC (PR-AUC) is 0.905, further emphasizing that the model maintains high precision even as recall increases—a particularly meaningful property in class-imbalanced regimes such as anomaly detection. The monotonic shape of the ROC curve and the steep drop in the PR curve at high recall values confirm the model's ability to identify highly anomalous instances with strong confidence.

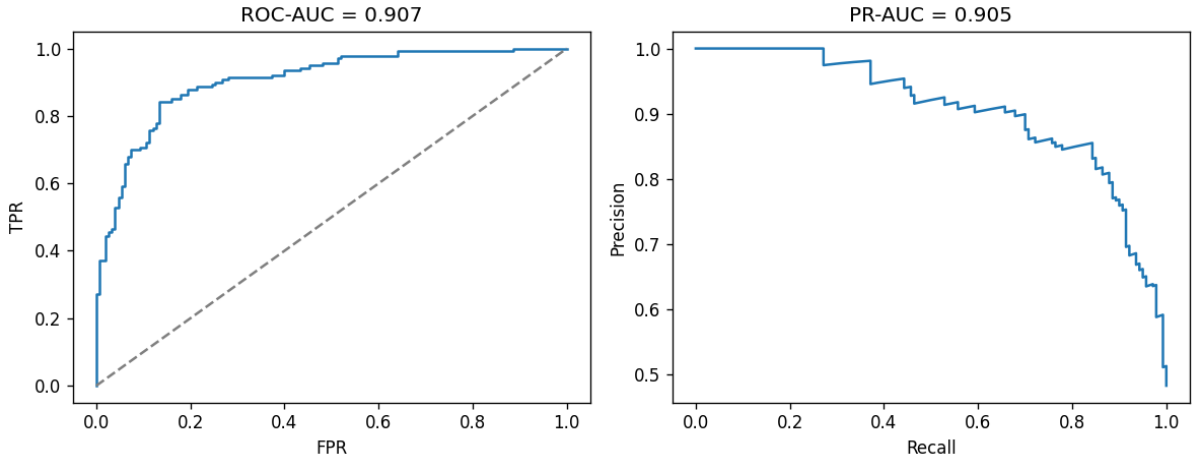


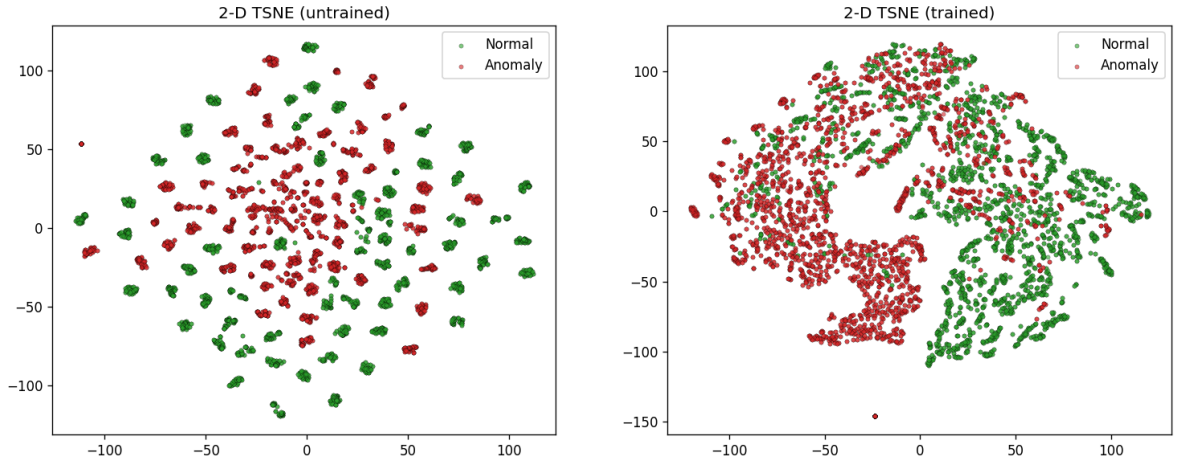
Figure 6: Left: ROC curve with AUC = 0.907. Right: PR curve with AUC = 0.905. The curves validate the model's robustness under class imbalance and confirm generalization.

5.2 Representation Quality via t-SNE Visualization

To assess the representational structure learned by the model, we visualize the segment-level embeddings using 2D t-SNE projections. t-distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear dimensionality reduction algorithm that maps high-dimensional data into a low-dimensional space (typically 2D or 3D) by preserving local structures. It converts pairwise similarities in the high-dimensional space into joint probability distributions and minimizes the Kullback–Leibler divergence between these distributions in the low-dimensional embedding. Unlike linear techniques such as PCA, t-SNE excels at revealing subtle cluster boundaries and is particularly well-suited for analyzing the internal structure of learned neural representations.

Figure 7 shows the t-SNE projection before training. The embeddings are largely entangled, with no discernible structure between normal and anomalous instances. This qualitative difference illustrates the transformative impact of MIL optimization on the feature space, enabling the network to learn anomaly-sensitive embeddings purely from weak video-level supervision.

Figure 7 presents the embedding space after training. A clear separation emerges between normal (green) and anomalous (red) segments, suggesting that the MIL head has successfully aligned feature spaces to maximize inter-class margins. Notably, anomalous segments form spatially coherent clusters, indicating the model's ability to localize and group semantically similar outliers even in the absence of temporal labels.



(a) Untrained (b) Trained

Figure 7: 2D t-SNE visualizations of segment embeddings. (a) Before training, normal and anomalous segments are entangled. (b) After training, a clear class separation emerges.

Method (Weak Supervision)	**Backbone / Key Idea**	**UCF-Crime AUC (%)**
Ours: 64-seg I3D + TimeSformer, Top-k MIL	Feature fusion + Top-k MIL	**90.70**
MSTAgent-VAD (VideoSwin, multi-scale, RTFM)	Transformer features + RTFM	89.27
MSTAgent-VAD (I3D variant)	I3D features + RTFM	85.52
Chen et al. 2023 (VST-RGB)	Vision Swin-Transformer	86.67
Tian et al. 2021 (I3D + sparsity, smoothness)	I3D MIL-ranking	84.30
Sultani et al. 2018 (original MIL ranking)	I3D	75.41

Table 2: Comparison of weakly-supervised methods on UCF-Crime dataset. Our method outperforms prior work in ROC-AUC, highlighting the benefit of combining motion and semantic features with top-k MIL aggregation.

6 Conclusion

In this work, we proposed a novel weakly-supervised framework for anomaly detection in videos, combining a dual-stream feature extraction backbone with a lightweight Multiple Instance Learning (MIL) head. By fusing the complementary strengths of I3D and TimesFormer encoders, our model captures both short-range motion cues and long-range semantic context. This rich representation is then aggregated using a top-k pooling mechanism, enabling effective training with only coarse video-level labels. Our empirical evaluations demonstrate the model's ability to achieve robust anomaly classification across various performance metrics. The ROC-AUC and PR-AUC scores, alongside the confusion matrix analysis, validate the method's precision and recall characteristics under class imbalance. Furthermore, the t-SNE projections of the learned embedding space highlight the model's capacity to construct a discriminative and structured representation of normal and anomalous events, even without explicit segment-level supervision. The proposed approach balances efficiency, accuracy, and interpretability. It avoids the cost of dense annotations, leverages pretrained spatiotemporal backbones, and maintains a compact inference architecture. This makes it particularly well-suited for real-world deployment in safety-critical domains such as urban surveillance, industrial monitoring, and autonomous security systems. Future work could explore adaptive thresholding strategies, temporal attention mechanisms, and alternative backbone architectures such as VideoMAE.

References

[1] Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6479-6488).

[2] Zhao, S., Zhao, R., Meng, Y., Gu, X., Shi, C., & Li, D. (2025). MSTAgent-VAD: Multi-scale video anomaly detection using time agent mechanism for segments' temporal context mining. *Expert Systems with Applications, 276*, 127154.

[3] Anoopa, S., & Salim, A. J. M. T. P. (2022). Survey on anomaly detection in surveillance videos. *Materials Today: Proceedings, 58*, 162-167.

[4] Doshi, K., & Yilmaz, Y. (2020). Any-shot sequential anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 934-935).

[5] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13576-13586).