A close up of a logo

Description automatically generated

**Enhancing Surveillance Video Anomaly Detection with Rejection Methods**

**Improving Real-World Video Anomaly Detection Through Confidence-Aware Rejection Mechanisms**

IoT project proposal

**Student name: Noam Tsfaty, Liav Cohen  
Supervisor: Dr. Yehudit Aperstein  
Date: 27/04/25**

Contents

[1. Introduction 3](#_Toc196182352)

[2. Literature review 4](#_Toc196182353)

[3. Methodology 5](#_Toc196182354)

[4. Datasets 6](#_Toc196182355)

[5. Work Plan 7](#_Toc196182356)

[6. References 7](#_Toc196182357)

## Introduction

With the increasing reliance on interconnected infrastructures in smart cities and the expansion of Internet of Things (IoT) devices, real-time video surveillance systems are becoming central to maintaining public safety, monitoring human activities, and preventing criminal events. Surveillance cameras installed across intersections, transport hubs, and public spaces continuously generate massive volumes of unstructured video data. Due to human cognitive limitations and the sheer scale of surveillance footage, manual inspection is not feasible, creating an urgent need for automated video anomaly detection systems.

Early methods in video anomaly detection relied heavily on handcrafted features, explicit motion tracking, or fully supervised learning approaches that demand frame-level annotations. These techniques struggle with generalization when confronted with variations in lighting, occlusion, crowd density, and other real-world complexities. A breakthrough was introduced by Sultani et al. [1], who proposed a weakly supervised learning framework using Multiple Instance Learning (MIL). In this method, each video is treated as a bag of segments, and the model learns to assign higher anomaly scores to segments from anomalous videos without requiring explicit segment-level labels.

However, a key limitation of MIL-based models is their inability to gauge the confidence of predictions. In high-stakes security applications, this leads to challenges in distinguishing between strong and uncertain detections, often resulting in false alarms or overlooked anomalies. To address this, we propose integrating a rejection mechanism based on entropy into the MIL framework. This addition allows the model to abstain from making low-confidence predictions, thus improving its precision, robustness, and trustworthiness.

Moreover, traditional models like C3D (Convolutional 3D Networks) used for extracting features have shown limitations in capturing long-range dependencies and complex temporal dynamics. Recent advances in video understanding have introduced Transformer-based architectures that outperform CNNs in many benchmarks. Two leading models—TimeSformer [3] and Video Swin Transformer [4]—demonstrate superior capabilities in capturing spatiotemporal relations.

TimeSformer leverages a convolution-free design and employs divided space-time attention, which allows it to model global dependencies over time and space efficiently. On the other hand, Video Swin Transformer adapts the Swin Transformer to video by incorporating hierarchical local attention through shifted 3D windows, achieving an excellent trade-off between accuracy and efficiency.

## Literature review

Anomaly detection in surveillance videos is a key area of research in computer vision, with direct applications in public safety, crowd monitoring, and infrastructure protection. The foundational work of Sultani et al. [1] proposed an MIL framework using video-level annotations, avoiding the need for expensive frame-level labeling. This framework introduced a ranking loss that pushes anomalous segments higher in score relative to normal ones while applying sparsity and temporal smoothness constraints. Geifman and El-Yaniv [2] addressed the problem of uncertainty in model predictions. They introduced selective classification using entropy-based and margin-based measures, allowing deep neural networks to reject low-confidence predictions. Their methodology, although developed for image classification, lays the theoretical foundation for integrating uncertainty estimation into temporal models like video anomaly detection. Ionescu et al. [5] proposed clustering normal behavior representations and flagging segments that fall outside these clusters as anomalies. While this method performs well for clear outliers, it lacks nuance in handling ambiguous or borderline cases. Our work differs by introducing a rejection mechanism to explicitly handle such uncertain predictions. Kiran et al. [6] conducted a comprehensive survey dividing deep learning approaches for anomaly detection into reconstruction-based (e.g., autoencoders), prediction-based (e.g., LSTMs), and ranking-based systems (e.g., MIL). Reconstruction and prediction methods suffer from poor localization and generalization, while MIL strikes a balance but lacks mechanisms for uncertainty. More recent contributions from Bertasius et al. [3] introduced TimeSformer, which rethinks video modeling using pure Transformers. Their model separates spatial and temporal attention to reduce computational overhead while retaining the ability to learn global dependencies. Similarly, Liu et al. [4] presented the Video Swin Transformer, which introduces 3D shifted windows and hierarchical attention computation. This model retains inductive biases such as locality and hierarchy, enabling efficient spatiotemporal modeling. In summary, the literature supports the integration of advanced Transformer-based architectures with confidence-aware mechanisms like entropy-based rejection, forming a highly interpretable and scalable anomaly detection pipeline.

## Methodology

Our proposed method builds upon the MIL ranking framework and enhances it with both advanced feature extraction and an entropy-based rejection mechanism.  
  
3.1 Feature Extraction with Transformer Backbones

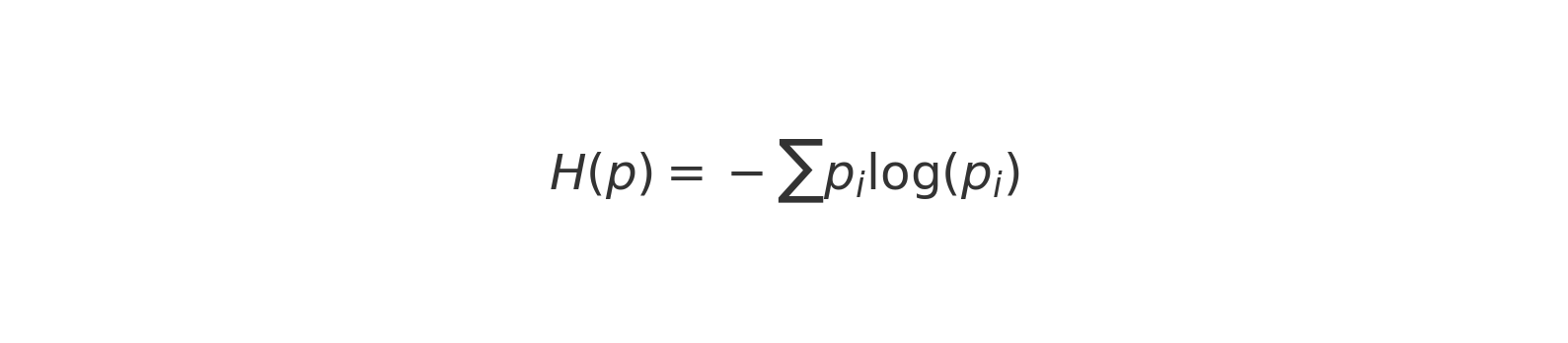
Each video is divided into non-overlapping segments (e.g., 16-frame clips). For each segment, we extract feature representations using one of the following:

* **TimeSformer [3]**: Applies separate self-attention across spatial and temporal dimensions. This factorized attention mechanism allows the model to capture long-term temporal dependencies effectively.
* **Video Swin Transformer [4]**: Divides the input into 3D windows and applies self-attention locally, while using shifted windows to ensure inter-window connectivity. The model processes videos hierarchically and scales well with input size.

**Visual Placement:**

* **Figure 1 (TimeSformer Architecture)** – placed at the beginning of this section.
* **Figure 1 & 2 (Video Swin Transformer)** – inserted immediately after to contrast the two architectural designs.

### 3.2 MIL Ranking Framework Let a video V be composed of n non-overlapping segments {v1, v2, ..., vn}. Each video is labeled either as normal (0) or anomalous (1), with no per-segment annotations. The training objective is to assign higher anomaly scores to segments from anomalous videos compared to those from normal ones. A neural network computes an anomaly score S(vi) ∈ ℝ for each segment. The loss function is defined as: Where: - S(a) and S(n) are scores for anomalous and normal segments, respectively. - The first term is a ranking hinge loss. - The second term enforces sparsity—only a few segments should have high scores. - The third term enforces temporal smoothness. 3.2 Rejection Module After computing segment-wise anomaly scores, we apply a softmax layer:

Then we compute the entropy of this distribution:  
  
  
We introduce a threshold τ, and any segment with entropy above this value is rejected. This allows the system to filter out segments where it is unsure of the anomaly classification, enhancing both reliability and interpretability.  
  
We will evaluate the improvement brought by this module in terms of standard classification metrics as well as new metrics like rejection rate, reduced false positives, and coverage-accuracy trade-offs.  
  
Visual diagrams will be created for both the base MIL framework and the enhanced model with rejection, showing the flow of data and the decision pipeline.

## Datasets

We will use the UCF-Crime dataset, a large-scale video anomaly dataset introduced by Sultani et al. [1]. It consists of 1900 long, untrimmed surveillance videos totaling over 128 hours. The dataset covers 13 different anomaly categories such as fighting, robbery, road accidents, and burglary, alongside a wide range of normal behaviors.

Each video is labeled at the video-level as either normal or anomalous. No frame-level or segment-level labels are provided for training, aligning perfectly with the MIL training paradigm. For evaluation, segment-level ground truth annotations are available, allowing precise calculation of ROC-AUC and other detection metrics.

We will preprocess the dataset by splitting videos into fixed-length segments (e.g., 16-frame clips) and extracting C3D or I3D features. These features will serve as inputs to the MIL model. The dataset will be split into 70% training and 30% testing, ensuring consistency with the baseline study.

## Work Plan

|  |  |  |  |
| --- | --- | --- | --- |
| **Week** |  |  |  |
| **1** |  |  |  |
| **2** |  |  |  |
| **3** |  |  |  |
| **4** |  |  |  |

## 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 (CVPR), pp. 6479–6488.

[2] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In Advances in Neural Information Processing Systems (NeurIPS).

[3] Bertasius, G., Wang, H., & Torresani, L. (2021). Is Space-Time Attention All You Need for Video Understanding? arXiv preprint arXiv:2102.05095.

[4] Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2021). Video Swin Transformer. arXiv preprint arXiv:2106.13230.

[5] Ionescu, R. T., Smeureanu, S., Alexe, B., & Popescu, M. (2019). Detecting abnormal events in video using narrowed normality clusters. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7864–7872).

[6] Kiran, B. R., Thomas, D. M., & Parakkal, R. (2018). An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. Journal of Imaging, 4(2), 36.