# Hybrid Architecture for Real-Time Video Anomaly Detection: Integrating Spatial and Temporal Analysis

Fabien Poirier

# Hybrid Architecture for Real-Time Video Anomaly Detection: Integrating Spatial and Temporal Analysis

Fabien Poirier

Paris 8 University, LIASD, France

November 29, 2024

## 1 Abstract

In this paper, we propose a new architecture for real-time anomaly detection in video data, inspired by human behavior combining spatial and temporal analyses. This approach uses two distinct models: (i) for temporal analysis, a recurrent convolutional network (CNN + RNN) is employed, associating VGG19 and a GRU to process video sequences; (ii) regarding spatial analysis, it is performed using YOLOv7 to analyze individual images. These two analyses can be carried out either in parallel, with a final prediction that combines the results of both analysis, or in series, where the spatial analysis enriches the data before the temporal analysis. Some experimentations are been made to compare these two architectural configurations with each other, and evaluate the effectiveness of our hybrid approach in video anomaly detection.

**Keywords:** Anomaly Detection, Real-Time, YOLOv7, VGG-GRU, Video Analysis, Modular Architecture

## 2 Introduction

Anomaly detection in videos is essential for various applications, ranging from security surveillance to disaster management, and the monitoring of large-scale events such as the Olympic Games. In such situations, the ability to quickly identify anomalies can have significant consequences, whether it is ensuring the safety of people or responding effectively to critical situations. However, traditional detection systems often rely on an isolated analysis of the temporal aspects of videos, limiting their ability to detect anomalies efficiently in complex environments (cf. Section 3). Since videos are multimodal, incorporating both static visual information (images) and dynamic information (sequences of

images), it is necessary to adopt an approach capable of capturing the full richness of potential anomalies as efficiently and quickly as possible. To this end, we propose a hybrid architecture that combines spatial analysis, to detect objects and visual patterns in each image, with temporal analysis to model the dynamics of video sequences. For this purpose, we use YOLOv7 [17], a state-of-the-art object detection model, coupled with a combination of the supervised learning model VGG19 and GRU to model temporal sequences. This approach enables not only the detection of anomalies based on the presence of suspicious objects but also the identification of suspicious behaviors over time. The originality of our approach lies in the integration of these two types of analysis. We explore two main configurations: a parallel approach (where spatial and temporal analyses are performed simultaneously and combined for a final prediction) against a sequential approach (where spatial analysis enriches the temporal analysis). This integration improves the precision and reliability of anomaly detection by leveraging the strengths of each type of analysis. This paper aims to evaluate the effectiveness of these hybrid configurations by comparing the parallel and sequential approaches, while also analyzing the impact of spatial analysis integration into our architecture. We trained and tested both approaches on a proprietary dataset. The results allow us to assess the specific contribution of spatial analysis, highlighting its strengths and limitations within the context of anomaly detection.

# 3 Related works

Anomaly detection in videos has made significant advances in recent years [13], driven by progress in deep learning technologies and data analysis techniques. In this study, Samaila et al. [13] highlights that deep learning has now emerged as a dominant approach compared to traditional machine learning methods. Among the various learning strategies used for video anomaly detection, reinforcement learning remains significantly less explored than supervised and unsupervised learning. Consequently, the authors focus primarily on deep learning techniques and these two prevalent learning paradigms. Therefore, this section examines recent approaches in object and anomaly detection, as well as hybrid models that combine spatial and temporal analyses.

## 3.1 Object Detection

The evolution of object detection models has played a crucial role in improving anomaly detection systems. Zou et al. (2019) in [21] published a detailed review of the evolution of object detection techniques over the past two decades. Initial methods, such as the Recurrent Convolutional Neural Network (RCNN) [6], paved the way for modern object detection but suffered from significant limitations. RCNN required generating a large number of region proposals, which were then processed individually by a convolutional network, resulting in extremely high computational costs and slow inference times. To address these ineffi-

ciencies, Fast RCNN [5] introduced a more integrated approach, while Faster RCNN [12] replaced region proposal mechanisms with a Region Proposal Network (RPN), significantly accelerating the process. Despite these advances, the RCNN family still struggled to meet the demands of real-time detection, especially in scenarios requiring high-speed analysis. This limitation led to the development of the You Only Look Once (YOLO) series [9–11], which redefined object detection by formulating it as a single regression problem. YOLO processes an image in a single pass, dividing it into a grid and simultaneously predicting bounding boxes and class probabilities, achieving unparalleled speed and efficiency. Successive versions of YOLO have brought significant improvements: YOLOv4 introduced advanced data augmentation techniques [2]. YOLOv7, currently the latest version available at the time of this publication, incorporates the YOLOR (You Only Learn One Representation) architecture [18] and eliminates anchor boxes, enabling ultra-fast image analysis [17].

### 3.2 Anomaly Detection

Anomaly detection technologies have also evolved and diversified. As for methods focusing solely on temporal analysis, we find technologies such as LSTMs, GRUs, CNNs, and GANs, which have been widely used for anomaly detection in videos, as mentioned by Samaila et al. [13]. Among these techniques, less common models like C3D (3D Convolutional Networks) have also shown promising results for extracting spatio-temporal features. This technology was previously used by Du Tran et al. [14] for action recognition, addressing problems similar to those discussed here. Lin Wang et al. [19] adopted this architecture to propose a weakly-supervised anomaly detection method, using a multi-instance pseudo-label generator and an anomaly detector enhanced by attention. Their goal is to overcome the limitations of traditional anomaly detection methods, particularly the lack of labeled data, through a weakly-supervised approach. The pseudo-label generator produces approximate labels for anomalous videos, transforming anomaly detection into a supervised learning problem. Videos are first processed to extract spatio-temporal features using the C3D network, which serves as a feature encoder. The model also integrates attention modules to focus on the anomalous regions of the videos. Finally, it is trained using the generated pseudo-labels and normal videos, and the C3D network parameters are fine-tuned to adapt to the task-specific features. Futhermore, Vision Transformers (ViT), although more recent, are being explored for their ability to capture complex relationships in video sequences. In their paper, Waseem Ullah et al. [15] present the ViT-ARN model, which consists of two distinct models, each playing a specific role in anomaly detection and recognition in videos. The first model is an anomaly detection model based on one-class classification (OCC), aimed at predicting whether an anomaly is present in a video or not. This model relies on a VGG-type network to extract features from images, followed by a fully connected network to classify events as normal or anomalous. The second model is dedicated to recognizing the types of anomalies, using a Vision Transformer (ViT) to extract spatio-temporal features from

videos. The images are divided into patches processed by a transformer encoder, and the process is refined by a Multi-Reservoir Echo State Network (MrESN). The final prediction is then made by a fully connected model. Regarding hybrid models combining spatial and temporal analysis, Doshi and Yilmaz [3] proposed an approach combining YOLOv3 (non-retrained) for object detection and FlowNet2 for optical flow feature extraction. These features are then processed by a KNN algorithm applied to surveillance camera images, as well as datasets such as CUHK Avenue, UCSD Pedestrian, and ShanghaiTech. However, it is worth noting that YOLOv3, although effective at its release, is now considered relatively outdated compared to newer versions like YOLOv7, which offer better performance and efficiency. Subsequently, they proposed MONAD (Multi-Objective Neural Anomaly Detector), an architecture composed of two main modules:

- The first module is a feature extraction module based on deep learning, using a generative adversarial network (GAN) to predict future video frames and compute the prediction error (MSE). This module also uses a lightweight object detector, YOLOv3, to extract localization information (center and area of the bounding box) and appearance information (class probabilities) of the objects detected in each frame. For each object, a feature vector is constructed by combining the prediction error, localization information, and class probabilities.

- The second module, dedicated to anomaly detection, uses a non-parametric sequential algorithm to analyze feature vectors in real time. It compares new observations with normal training data using a k-nearest neighbors (KNN) approach [4].

However, using GANs to predict future frames is not ideal for real-time applications, as these models are often computationally expensive and introduce significant latency. More recently, Mostafa [1] introduced the AVAD (Autoencoder-based Video Anomaly Detection) method, which uses a convolutional autoencoder to detect abnormal frames and YOLOv5 (non-retrained) to identify objects responsible for anomalies. This method was applied to the same datasets used by Doshi and Yilmaz, including UCF Crime. However, the autoencoder-based approach has a significant limitation: the model must reconstruct each image in the input sequence, making it suboptimal for real-time applications due to the computational overhead associated with this process.

### 3.3 Dataset

Regarding existing datasets, Zhu et al. [20] lists several datasets suitable for anomaly detection in videos. These include the Dashcam Accident Dataset (DAD), the Car Accident Dataset (CADP), A3D, DOTA Detection of Traffic Anomaly (DADA), UCSD, ShanghaiTech, and UCF Crime. These datasets can be categorized into three groups. On one hand, there are datasets that focus exclusively on a specific type of anomaly, such as DAD, CADP, A3D,

and DADA, which are dedicated to traffic-related issues. On the other hand, datasets like UCSD and ShanghaiTech address low-impact anomalies related to safety, such as bicycles on sidewalks. Finally, the third category includes UCF Crime, which is the only dataset relevant to our study. It contains 1,900 raw videos divided into 13 types of anomalies: abuse, arrests, arson, assault, road accidents, burglary, explosions, fighting, armed robbery, shootings, shoplifting, theft, and vandalism, as well as videos without anomalies. However, despite its diversity, UCF Crime is insufficient for effectively training artificial intelligence models for video anomaly detection, as noted by Vrskova et al. [16]. In addition to being poorly cleaned and unbalanced, UCF Crime lacks a sufficient amount of data to meet the needs of anomaly detection models. Jacob [7] further points out that no dataset is currently rich enough to properly train a deep learning model for video anomaly detection. This view is reinforced by Samaila et al. [13], who indicate that public datasets are limited in size and diversity, and most data is too unrealistic.

# 4 Spatio-temporal Video Analysis Architecture

In this article, we propose an architecture composed of two complementary analyses: a spatial analysis and a temporal analysis. Figure 1 illustrates this dual approach and its overall structure.

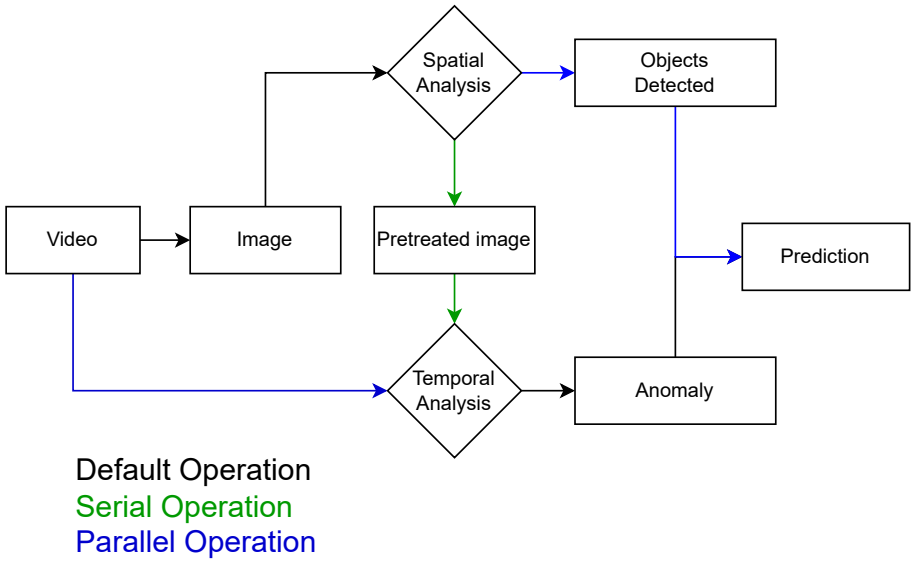


Figure 1: Spatio-temporal Video Analysis Architecture

On one hand, for spatial analysis, we have chosen to use YOLOv7 for its remarkable performance in terms of execution speed and detection accuracy. The

temporal analysis, on the other hand, relies on a neural network combining VGG19 (Visual Geometry Group) and GRU (Gated Recurrent Unit), also including an Multi-Layer Perceptron (MLP). This model is designed to process sequences of 15 images, each with a size of 112x112 pixels. VGG-GRU has already been used to achieve high performance in the detection of anomalies in videos [8]. Figure 2 presents in detail the structure of this temporal analysis model.

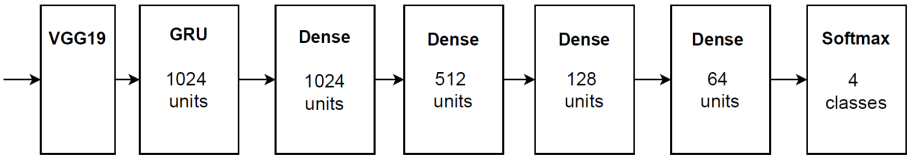


Figure 2: Structure of the Temporal Analysis Component VGG-GRU.
The GRU layer has a dropout rate set to 50%, as do all Dense layers, which also have L2 regularization fixed at 0.01.

These two models have been trained on proprietary data sets[1]. The data set for YOLOv7 consists of 10,000 images of firearms, 2,000 images of fires containing flames or smoke, people images, and images unrelated to these objects to ensure robust and adaptable detection. All images are in JPEG format. The VGG-GRU model, meanwhile, was trained on a corpus of MP4 videos representing three classes of anomalies: fights (978 videos), gunshots (311 videos), and fires (298 videos). These videos were carefully edited to retain only the segments that contained anomalies. Videos that did not contain anomalies were added to represent the"normal" class. One of the main advantages of this architecture lies in its flexibility. The spatial analysis model can be positioned either in parallel or in series, depending on the specific needs of the application. In addition, by configuring the frame interval during the video preparation process, it is possible to handle video streams (continuous video) or downloaded videos (finished video). It is also possible to completely deactivate one of the two analysis modules or replace them with more recent versions without requiring the entire architecture to be retrained. For instance, a more advanced version of YOLO can be seamlessly integrated. Furthermore, the architecture can run on a CPU, although GPU execution is recommended for optimal performance.

# 5 Experiments

In this section, we present various experiments that aim to compare the performance of our different architectures. We will begin by evaluating the architecture in which our components are arranged in parallel.

[1]The data sets used are confidential and cannot be shared publicly.

### 5.1 Parallel Architecture

This architecture uses YOLOv7 to detect various key objects related to our anomalies, based on the assumption that the absence of these objects makes the occurrence of the anomaly highly unlikely. We focus on detecting people for the "fight" anomaly, the presence of flames and smoke for "fire", and the Intersection over Union (IoU) between a person and a firearm for the "gunshot" anomaly. Once predictions are made by our two analysis modules, they are combined using the following logical rule:

1. If our model detects an anomaly, the predicted class will correspond to this anomaly.
2. Otherwise, if a key object is detected, we will predict the anomaly associated with that object. However, in the specific case of a weapon, we will first check if the IoU between the weapon and a person is greater than 0 before predicting the "gunshot" anomaly.[2]

For the "gunshot" class, given that it involves two distinct objects and that a firearm only represents a risk when it is within reach of a person, we use the IoU between these two objects when combining the results of our detections. This approach allows for a more precise evaluation of the spatial context of the detected objects. Although the architectures presented in section 3 are diverse, they are not systematically comparable with one another. Most of them focus on unsupervised learning techniques or rely on models that do not necessarily meet the constraints we have set, such as real-time processing. Furthermore, while many architectures claim to be suitable for real-time anomaly detection, it is important to highlight that none of these studies provide specific timing metrics or performance benchmarks to support these claims. For this reason, we chose to compare our model with C3D, a well-known architecture designed for a related task, namely action recognition, and for which pretrained weights are available. Our initial tests on our dataset demonstrate that our architecture, based on a combination of CNN + RNN + GRU, delivers slightly better performance than C3D, achieving an F1-score of 36.5% compared to 35.7% [8]. The results of our model are presented in Table 1.

[2]Our model was designed to detect real weapons. In this context, no tests have been conducted on images containing toys, drawings, or other types of representations.

Table 1: VGG-GRU + YOLO Performance

| Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|
| 78.42% | 85.60% | 78.42% | 81.16% |

Confusion matrix (in percent)

| Truth \ Predicted | Fight | Gunshot | Fire | Normal |
|---|---|---|---|---|
| Fight | **63.66%** | 6.58% | 1.93% | 27.83% |
| Gunshot | 9.94% | **66.06%** | 9.33% | 14.67% |
| Fire | 13.66% | 15.73% | **57.71%** | 12.9% |
| Normal | 7.43% | 5.96% | 3.98% | **82.63%** |

Although the precision score presented in Table 1 is 85.6%, the recall (78.42%) could be improved. These results demonstrate good anomaly detection but the confusion matrix reveal weaknesses for specific classes, such as the "fire" class, where confusion with other types of anomalies is high.

## 5.2 Serial Architecture

As part of our serial analysis, we leveraged YOLOV7 to enrich our input data by applying various preprocessing techniques. Our main approach consisted of removing the background from images, leaving only the key objects detected by YOLO. This method is illustrated in Figure 3.

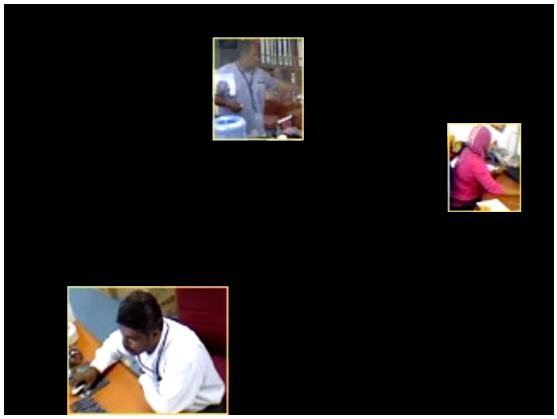

Figure 3: Example of Mask Generated with YOLO

The objective of this technique is to focus our model's attention on relevant elements, thus avoiding it from concentrating on parasitic background movements. During this experimentation, we explored two options for cases where no key objects were detected by YOLO:

- Retaining the original image without background: The results of this approach are presented in Table 2.
- Introducing an entirely black image: This alternative is presented in Table 3.

Table 2: Performance for Mask without Black Background

| Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|
| 72.22% | 82.37% | 72.22% | 75.68% |

Confusion matrix (in percent) for mask without black background

| Truth \ Predicted | Fight | Gunshot | Fire | Normal |
|---|---|---|---|---|
| Fight | **63.06%** | 9.33% | 1.96% | 25.65% |
| Gunshot | 25.99% | **41.45%** | 2.29% | 30.27% |
| Fire | 19.23% | 15.98% | **32.44%** | 32.35% |
| Normal | 15.33% | 4.41% | 2.07% | **78.19%** |

Table 3: Performance for Mask with Black Background

| Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|
| 75.58% | 79.34% | 75.58% | 76.50% |

Confusion matrix (in percent) for mask with black background

| Truth \ Predicted | Fight | Gunshot | Fire | Normal |
|---|---|---|---|---|
| Fight | **55.52%** | 4.43% | 1.76% | 38.29% |
| Gunshot | 16.20% | **41.44%** | 1.53% | 40.83% |
| Fire | 14.09% | 12.49% | **20.59%** | 52.83% |
| Normal | 10.55% | 2.98% | 2.09% | **84.38%** |

Comparing these two approaches, we noted that using a black image in the absence of detected objects improved the detection of the "normal" class, increasing from 78% to 84%. However, this improvement came at the expense of precision for other classes, particularly "fight" and "fire." For the "fight" class, precision decreased from 63% without a black image to 55% (-8%), while for the "fire" class, it dropped from 32% to 20% (-12%). Given the behavioral nature of certain anomalies, particularly those involving human interactions, we leveraged the flexibility of our architecture to integrate YOLOv7-pose as a replacement for standard YOLOv7. This adaptation allows us to trace the skeleton of each person present on the screen, thus offering a more refined analysis of movements and postures. We experimented with two preprocessing approaches using YOLOv7-pose:

- Preservation of the background with superimposition of detected skeletons: Visual illustration (see Figure 4) and Detailed results (see Table 5);
- Removal of the background, presenting only the skeletons on a black background: Visual illustration (see Figure 5) and Detailed results (see Table 4).

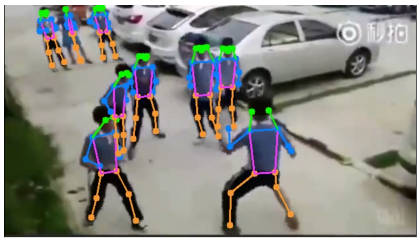

Figure 4: Pose Estimation by YOLOv7 with background

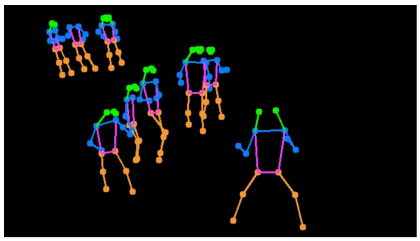

Figure 5: Pose Estimation by YOLOv7 without background

Table 4: YOLOv7-Pose + VGG-GRU without Background (3 Classes)

| Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|
| 87.2% | 90.9% | 87.8% | 89% |

Confusion Matrix (in %) for Pose Analysis without background

| Truth \ Predicted | Fight | Gunshot | Normal |
|---|---|---|---|
| Fight | **64.5%** | 1% | 34.5% |
| Gunshot | 11.7% | **70.7%** | 17.6% |
| Normal | 9.5% | 0% | **90.5%** |

Table 5: YOLOv7-Pose + VGG-GRU with Background (4 Classes)

| Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|
| 87.3% | 87.6% | 87.3% | 87.1% |

Confusion Matrix (in %) for Pose Analysis without Black Background

| Truth \ Predicted | Fight | Gunshot | Fire | Normal |
|---|---|---|---|---|
| Fight | **60.5%** | 2.4% | 1.3% | 35.8% |
| Gunshot | 10% | **55.6%** | 14.8% | 19.6% |
| Fire | 15.5% | 10.6% | **48%** | 25.9% |
| Normal | 3.4% | 0.6% | 1% | **95%** |

The results of these two experiments (see Table 4 and Table 5) revealed interesting trade-offs. The analysis of poses on a black background significantly improved the detection of anomalies related to human behaviors, such as fights (+4%) and gunshots (+15%), compared to the normal background. This improvement is explained by the increased focus of the model on people's movements and interactions, without distraction from the background. However, this approach also led to a notable decrease in the system's ability to detect fires. The removal of the background effectively eliminated crucial visual information for identifying flames and smoke, rendering these anomalies undetectable. This experiment highlights the importance of a judicious balance between focusing on human behaviors and preserving contextual information from the environment. It also underscores the need for an adaptive approach in anomaly detection, capable of adjusting based on the specific nature of the anomalies to be detected. Given the potential impact of the anomalies discussed in this article, we aimed to conclude our experiments by measuring various processing times. Each of these measurements was performed on a laptop equipped with 32 GB of RAM, an Intel Core i9 processor with 16 cores clocked at 2.3 GHz, and an Nvidia GeForce RTX2080 GPU with 8 GB of dedicated memory. As expected, the parallel architecture proved to be significantly faster than the sequential one (cf. Table 6 and Table 7).

Table 6: Execution time of YOLO and CGRU in parallel

| Video duration | Average detections | Processing time |
|---|---|---|
| 16s | 601ms | 15s |
| 44s | 533ms | 35s |
| 9s | 994ms | 12s |
| 35s | 1.1s | 57s |
| 23s | 1s06 | 35s |
| 1min 43 | 758ms | 116s (1min 56) |
| 50s | 826ms | 61s |
| 1min | 30 886ms | 83s (1min 23) |
| 2s | 847ms | 847ms |
| 9s | 870ms | 11s |
| 2s | 1s | 1s |

Table 7: Execution time of YOLO and CGRU in serie

| Video duration | Average detections | Processing time |
|---|---|---|
| 16s | 1s | 26s |
| 44s | 1s | 71s (1min 11) |
| 9s | 1.5s | 20s |
| 35s | 1.5s | 81s (1min 21) |
| 23s | 1.5s | 48s |
| 1min 43 | 1.2s | 193s (3min 13) |
| 50s | 1.3s | 102s (1min 42) |
| 1min 05 | 1.4s | 134s (2min 14) |
| 2s | 1.3s | 1.3s |
| 9s | 1.3s | 17s |
| 2s | 1.5s | 1.5s |

This can be explained by the fact that, in the sequential setup, YOLOv7 must pre-process each video frame before being analyzed by our VGG-GRU model. However, the average prediction time for the sequential architecture remains reasonably fast, with an average of 1 to 1.5 seconds per prediction, allowing for a timely response when an anomaly is detected. It is worth noting that the execution speed of the parallel model can be further improved by adjusting the number of frames YOLOv7 needs to analyze. Due to the combination of both models, it is not necessary to analyze every frame in a sequence to make a prediction. Because of the redundant information between successive frames, the model can be configured to analyze a reduced number of frames per sequence, which enhances processing speed.

# 6 Conclusion and Future works

This paper proposes a hybrid architecture combining spatial and temporal analyses for real-time video anomaly detection. This approach leverages the strengths of YOLOv7 for object detection and the VGG-GRU model for temporal sequence analysis, offering flexibility in the arrangement of the modules according to specific needs. Our various experiments have allowed us to identify two optimal configurations for video anomaly detection, each addressing specific requirements:

**For precise anomaly detection** The serial configuration, combining YOLOv7 with VGG-GRU, has proven particularly effective. This approach excels in identifying human behavioral anomalies. The integration of pose estimation preprocessing and background removal has significantly improved results, offering detailed analysis adapted to situations where accuracy is paramount.

**For real-time analysis** The parallel architecture, combining YOLOv7 and our VGG-GRU, offers an optimal balance between reliability and speed. This

configuration ensures prompt detection of anomalies while maintaining adequate accuracy, thus meeting the needs of applications requiring instant processing.

These results highlight the importance of an adaptive approach in video anomaly detection. Our modular architecture effectively responds to various scenarios, paving the way for diverse applications in the fields of security, surveillance, and event management. The flexibility of our system allows for prioritizing either precision or speed, depending on the specific requirements of each application.

Our study has highlighted several challenges and improvement perspectives for our anomaly detection system. Firstly, we observed that not all anomalies are necessarily linked to identifiable key objects, particularly in the case of natural disasters. Moreover, the presence of key objects does not always signify danger, as illustrated by the example of armed military personnel in airports. To overcome these limitations, we are considering directly transmitting the information collected by YOLO and VGG+GRU to our Multi-Layer Perceptron, allowing it to automatically learn the conditions for anomaly detection. We also noticed confusion between certain anomaly classes, suggesting the potential benefit of exploring a binary model to verify this hypothesis.

Another promising approach would be to combine several different processing methods. For example, by associating a mask with the detected objects and representing people by their skeletons, which could improve the reliability of the sequential model, at the cost of reduced detection speed. Finally, to enhance the robustness and versatility of our system, it would be beneficial to enrich our dataset with new anomaly classes and examine the execution speed of our architecture compared to other existing models in the field.

# References


[1] Manal Mostafa Ali. Real-time video anomaly detection for smart surveillance. *IET Image Processing*, 17(5):1375–1388, 2023.

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection, 2020.

[3] Keval Doshi and Y. Yilmaz. Continual learning for anomaly detection in surveillance videos. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1025–1034, 2020.

[4] Keval Doshi and Yasin Yilmaz. Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. *Pattern Recognition*, 114:107865, 2021.

[5] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[7] Seby Jacob. *Anomaly detection from videos: A deep learning approach.* McGill University (Canada), 2019.

[8] Fabien Poirier, Rakia Jaziri, Camille Srour, and Gilles Bernard. Enhancing anomaly detection in videos using a combined yolo and a vgg gru approach. In *2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, pages 1–6. IEEE, 2023.

[9] Joseph Redmon. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[10] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[11] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[13] Yau Alhaji Samaila, Patrick Sebastian, Narinderjit Singh Sawaran Singh, Aliyu Nuhu Shuaibu, Syed Saad Azhar Ali, Temitope Ibrahim Amosa, Ghulam E. Mustafa Abro, and Isiaka Shuaibu. Video anomaly detection: A systematic review of issues and prospects. *Neurocomputing*, 591:127726, 2024.

[14] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[15] Waseem Ullah, Tanveer Hussain, and Sung Wook Baik. Vision transformer attention with multi-reservoir echo state network for anomaly recognition. *Information Processing & Management*, 60(3):103289, 2023.

[16] Roberta Vrskova, Róbert Hudec, Patrik Kamencay, and Peter Sykora. A new approach for abnormal human activities recognition based on convlstm architecture. *Sensors (Basel, Switzerland)*, 22, 2022.

[17] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022.

[18] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. You only learn one representation: Unified network for multiple tasks. *arXiv preprint arXiv:2105.04206*, 2021.

[19] Lin Wang, Xiangjun Wang, Feng Liu, Mingyang Li, Xin Hao, and Nianfu Zhao. Attention-guided mil weakly supervised visual anomaly detection. *Measurement*, 209:112500, 2023.

[20] Sijie Zhu, Chen Chen, and Waqas Sultani. Video anomaly detection for smart surveillance. In *Computer Vision: A Reference Guide*, pages 1315–1322. Springer, 2021.

[21] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019.