Optimal Feature Extractor for Video Anomaly Detection in Public Transportation Applications

Jonathan FLORES-MONROY [a,b,1], Gibran BENITEZ-GARCIA [b]
Mariko NAKANO [a] and Hiroki TAKAHASHI [b,c]

[a] *Instituto Politecnico Nacional, Mexico City, Mexico*
[b] *Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo, Japan*
[c] *Artificial Intelligence eXploration Research Center, The University of Electro-Communications, Tokyo, Japan*
ORCiD ID: Jonathan FLORES-MONROY
https://orcid.org/0000-0002-2467-3600

Abstract. Video Anomaly Detection (VAD) is a well-established area of research with significant potential for enhancing video surveillance in urban public transportation. However, current VAD systems often propose powerful methodologies but overlook their use in extreme environments like public transportation, necessitating a balance between performance and computational efficiency. In this paper, we evaluate a key component in many VAD frameworks: feature extractors. We investigate five extractors: Inflated 3D ConvNets (I3D), 3D Convolutional Neural Networks (C3D), Unified Transformer (UniFormer) in Small (UniFormer-S) and Base (UniFormer-B) versions, and Temporal Shift Module (TSM). These are integrated into a VAD architecture employing Bidirectional Encoder Representations from Transformers (BERT) with Multiple Instance Learning (MIL), chosen for its modularity and clear separation between the feature extractor and anomaly detector module. UniFormer-S demonstrated a processing rate of 4.64 clips per second with a computational demand of 28.717 GFLOPs on edge devices like the Jetson Orin NX (8GB RAM, 20W power). On the UCF-Crime dataset, UniFormer-S with BERT + MIL achieves an AUC of 79.74%. These findings highlight the promise of UniFormer-S and the use of edge devices like the Jetson Orin NX in public transportation due to their balance of performance and efficiency.

Keywords. video anomaly detection, edge divice, feature extractor, public transportation security

[1]Corresponding Author: jfloresm1510@alumno.ipn.mx

1. Introduction

In recent years, urban security, particularly in public transportation, has posed a significant challenge worldwide [1–3]. This problem has been exacerbated by rapid urbanization and rising crime rates, with Latin America standing out as a particularly affected region [1,3]. For example, Mexico faces high rates of assaults, as indicated by data from the National Survey of Victimization and Perception of Public Security (ENVIPE), which reports a concerning rate of 5,689 assaults per 100,000 inhabitants [1], highlighting a significant impact on public transport and the urgent need to address these concerns. Despite efforts such as the installation of surveillance systems and increased police presence [4–8], these approaches have proven insufficient to address the situation.

In this context, the use of cutting-edge tools becomes imperative to tackle these challenges. A promising proposal involves employing artificial intelligence methods, particularly in computer vision, to anticipate unusual events. This leads us to the field of Video Anomaly Detection (VAD), where models are developed to identify any abnormal activity in a specific environment, such as public transportation.

As the main objective of VAD is to determine when an anomaly occurs and distinguish it from normal events, two main paradigms have emerged to address this issue: unsupervised learning and weakly supervised learning [9–17, 21, 22, 33]. Typically, this last paradigm is applied when video-level labels are available during training and frame-level annotations are provided during testing. This approach strikes a balance between accuracy and efficiency in the use of annotated data and is widely utilized in most current proposals to develop innovative solutions. One notable approach within of weakly supervised paradigm is Multiple Instance Learning (MIL), proposed by Sultani et al. [9]. This method classifies video segments into positive and negative bags to effectively distinguish between normal and anomalous behaviors using a pre-trained feature extractor. By emphasizing instances with the highest discrepancies in scoring, MIL ensures accurate and reliable anomaly detection.

While MIL is a prominent example, there are various other methodologies that have inspired numerous solutions in the field, demonstrating notable precision. However, despite the effectiveness of these methodologies, their implementation in real-world environments presents significant challenges. Solutions like MIL, although precise, often require stringent hardware and high energy consumption, complicating their integration into mobile environments necessary for deployment in extreme conditions such as public transport. We are convinced that the ability to detect anomalies efficiently and accurately is as crucial as minimizing the computational resources required. Additionally, we believe that the feature extractor is one of the most critical components to consider during the integration of anomaly detection systems in real-world applications. This phase of the process not only directly affects the precision and effectiveness of the system but also determines the level of computational resources consumed during continuous operation.

In this paper, we evaluate the performance and computational efficiency of five feature extractors commonly used in VAD methodologies [9, 12, 20, 34]. We

first select Inflated 3D ConvNets (I3D) [18, 19], 3D Convolutional Neural Networks (C3D) [9], Unified Transformer in both Small (UniFormer-S) and Base (UniFormer-B) versions [20], and Temporal Shift Module (TSM) [21]. Subsequently, these models were integrated into a framework combining Bidirectional Encoder Representations from Transformers (BERT) [23] with Multiple Instance Learning (MIL) [10]. This combination was chosen for its ability to work with various feature extractors and its well-defined structure, which clearly separates the feature extraction component from the anomaly detection module. It is important to note that this combination improves anomaly detection by better capturing contextual and temporal dependencies in video data, significantly enhancing the detection accuracy, which is crucial for security applications in public transportation.

Secondly, we evaluated their integration and processing rate for each feature extractor, on edge devices like the Jetson Orin Nano [35] and Jetson Orin NX [35], simulating real-time conditions typical of public transportation. Using the UCF-Crime database [9], the most widely used large-scale database in this field, we demonstrate the applicability and efficiency of the selected devices and feature extractors.

Our results demonstrate that UniFormer-S significantly outperforms other feature extractors in terms of accuracy and robustness. UniFormer-S achieves an anomaly detection accuracy of 79.74% AUC, which is higher compared to the other models tested (about 1.5% on average). Additionally, UniFormer-S proves to be computationally efficient, with significantly lower processing times and resource utilization compared to models like C3D, UniFormer-B, and TSM. During our tests, we evaluated UniFormer-S on cutting-edge devices such as the Jetson Orin NX 8GB, simulating real-time conditions typical of public transportation, where it processed 4.64 clips per second. This balance between efficiency and accuracy makes UniFormer-S an ideal candidate for implementation in public transportation systems to enhance urban security.

2. Video Anomaly Detection

Currently, the study of Video Anomaly Detection (VAD) has become popular due to its extensive application in various security settings. The ability to efficiently detect such events at the precise moment has increasingly become a common practice. Therefore, recent research in this field has focused on two main paradigms (unsupervised learning and weakly supervised learning).

2.1. Unsupervised Anomaly Detection.

To address the problem of determining when an anomaly occurs, VAD paradigms such as unsupervised methods [25–31] and one-class classification (OCC) [12, 24] are used. In the unsupervised approach, the system learns normal behavior patterns from unlabeled data and detects anomalies as significant deviations from these patterns. On the other hand, OCC trains a model using only normal behavior data, so any behavior that significantly differs from this is classified as an anomaly.

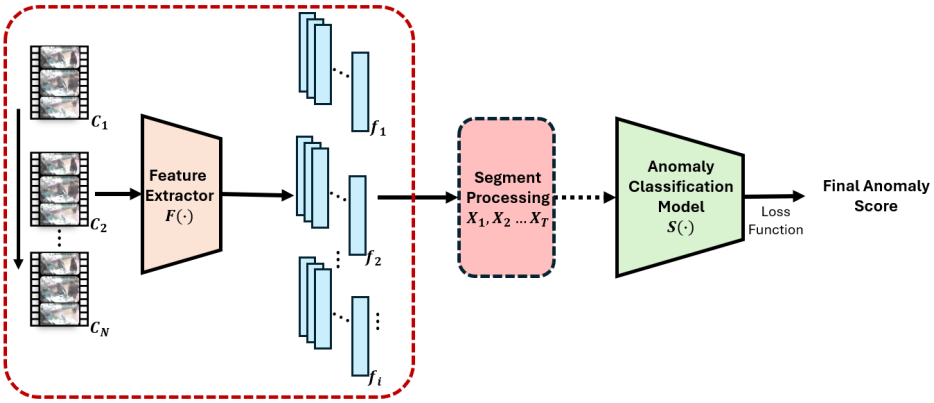

Figure 1. Generic pipeline for VAD using weakly supervised learning. It includes feature extraction, anomaly classifier and loss functions, specific to each proposal. The area outlined in red highlights the primary focus of this paper, while the pink block with the area outlined represents the segment processing stage, which may or may not be included depending on the specific architecture being analyzed.

However, these methods face challenges such as overfitting and difficulties in distinguishing between normal and anomalous due to the lack of prior knowledge about anomalies. The practical implementation of these models can also be problematic, as it requires high-performance devices [25–31]. Additionally, the effectiveness of the OCC approach can be limited by the diversity of video data, where new normal scenes might be mistakenly classified as anomalous [9, 11, 24].

2.2. Weakly Supervised Anomaly Detection.

Weakly supervised learning is defined as a methodology that uses limited or incomplete labels to train deep learning models. In the context of VAD, this approach involves training with video-level labeling. Specifically, it consists of two sets: one with normal videos and another with anomalous videos. Each video in the dataset receives a binary label indicating whether it is normal or anomalous, simplifying the annotation process compared to frame-level labeling. On another hand, in the test set, the videos are labeled with temporal annotations that indicate the exact segments where anomalies occur, providing a more detailed evaluation of the model's performance. Typically, a classic weakly Video Anomaly Detection (wVAD) methodology follows the structure of Figure 1. First, the videos are divided into clips or snippets, which are processed by feature extractors to obtain informative features. These extractors do not directly identify anomalies but prepare the information for subsequent analysis. Subsequently, an anomaly classification model is implemented, which relies on loss functions specifically designed to align with the adopted methodology. Additionally, this approach may include a preprocessing step for the features before analysis by the classification model, although this is not mandatory and depends on the specifics of the study. For example, in multiple instance learning (MIL) [9] which is a typical methodology in wVAD, the snippets are preprocessed to segment the video into a fixed number of segments, grouping them into positive (anomalous) and negative (normal) bags.

MIL uses a Ranking Loss Function [9] to train the model to differentiate between normal and anomalous behaviors. However, this method generates noisy labels due to the difficulty in accurately determining the start and end of an anomalous event.

To overcome MIL's limitations, alternative approaches have been proposed. Zhong et al. [11] reformulate MIL using Graph Convolutional Networks (GCN) to clean label noise. [12] proposes a Robust Temporal Feature Magnitude Learning (RTFM) that focuses on studying feature magnitudes by applying a Multi-scale Temporal Network (MTN), increasing the likelihood of selecting anomalous segments in anomalous videos. In [22] a self-supervised sparse representation (S3R) framework is proposed, which uses self-supervised learning-based dictionaries, combining dictionary-based representation and self-supervised learning, generating pseudo-normal/anomalous samples to train the anomaly detector.

Recent models, such as [13], use transformers-based architectures, fine-tuning a Unified transformer (UniFormer) network [20] where both the extractor and classifier are integrated into an end-to-end method. Others, like [10], enhance models like MIL or RTFM by integrating feature vectors with a Bidirectional Encoder Representations from transformers (BERT) [23], where snippets pre-calculated by a feature extractor are preprocessed into segments and then passed to the BERT architecture to generate a unified feature vector, improving anomaly detection. Nonetheless, these approaches present challenges for real-world integration due to their high computational requirements and complexity.

As observed, while each architecture aims to address and improve different challenges, most of these methodologies share a common denominator that affects their real-world implementation: dependency on feature extractors. These extractors require pre-calculated features before analyzing anomalies, highlighting the importance of consciously selecting the best pre-calculated model to enhance the viability of practical applications.

2.3. Feature Extractors and Their Role in VAD

In VAD, feature extractors are typically derived from Video Action Recognition (VAR). These extractors, designed to identify and classify human actions, prepare data for anomaly detection, which is crucial in environments like shopping malls, busy streets, or public transport. Below are some of the most commonly used feature extractors in VAD:

3D Convolutional Neural Networks (C3D) [32]: Compared to 2D ConvNets, which only capture spatial features from individual frames, 3D ConvNets are more effective at modeling temporal information due to their 3D convolution and 3D pooling operations, which process data both spatially and temporally. While 2D ConvNets treat each frame independently, 3D ConvNets analyze sequences of frames simultaneously, preserving the temporal context between consecutive frames. The 3D convolution operation maintains this temporal information, resulting in an output volume that reflects the spatiotemporal characteristics of the input data. Sultani et al. [9] and RTFM [12] utilized 3D convolutional networks to capture spatiotemporal information in videos. These networks analyze raw RGB video clips and extract features up to the FC6 layer. The 3D ConvNet architec-

ture comprises eight convolutional layers, five pooling layers, two fully connected layers, and a softmax output layer. This configuration makes 3D ConvNets particularly effective for tasks involving motion dynamics and temporal patterns in videos.

Inflated 3D ConvNets (I3D) [19]: I3D improves 2D ConvNets by extending them into 3D, capturing spatiotemporal information through added temporal dimensions to height and width. I3D processes raw RGB video clips by inflating 2D convolutional filters into 3D (e.g., k×k filters become k×k×k), allowing the network to learn actions and patterns across consecutive frames. This architecture, consisting of multiple 3D convolutional layers, pooling layers, and fully connected layers, effectively preserves temporal information and improves motion recognition accuracy. In VAD, recent research [10–12, 14] has replaced C3D with I3D, where spatio-temporal information is typically extracted from the `mixed_5c` layer.

Unified transformer (UniFormer) [20]: UniFormer combines the advantages of 3D convolution and spatiotemporal self-attention in a unified transformer, capturing global and local dependencies in different layers and balancing precision and computational cost. The authors of [13] propose using a UniFormer version that initially extracts features from raw 32-frame clips. These processed snippets go through a process of anomaly selection and classification, comparing Euclidean distances between features of videos labeled as normal and abnormal, along with other methods to maximize differences between normal and abnormal segments. The snippets with the highest probability of being anomalies are used to identify the corresponding raw clips, which may contain anomalies. These raw clips are processed in an end-to-end model that integrates the extractor and classifier in UniFormer, optimizing anomaly detection.

Temporal Shift Module (TSM) [34]: On the other hand, there are proposals that offer alternatives of 3D convolutional networks (like I3D or C3D) and transformers-base (such as UniFormer), for example, with a Temporal Shift Module (TSM). TSM's central idea is to improve temporal modeling by employing 2D convolutional networks. This is achieved by shifting a portion of the feature map channels forward and another portion backward in time, while the rest of the channels remain static. This approach allows 2D networks to capture temporal dependencies, converting them into pseudo-3D, without significantly increasing computational complexity. In this case, [33] proposed a model for detecting anomalies in videos using temporal convolutions called Anomaly Detection Network (ADNet). ADNet processes video clips in sliding windows, using features extracted by pre-trained models such as I3D and TSM. The AD loss function, designed to maximize the distance between hard pairs of opposing classes, improves the accuracy in detecting anomalous segments. Experiments showed that I3D features are more useful for segmenting normal events, while TSM features are better for segmenting anomalous events.

Knowing a bit about feature extractors and how they are implemented in various VAD proposals such as [9, 10, 13, 33], the choice of the optimal model for anomaly detection in real-world environments, like public transportation, remains an open question. Therefore, in this study, we set out to evaluate several extractors to determine which offers the best performance in practical applications.

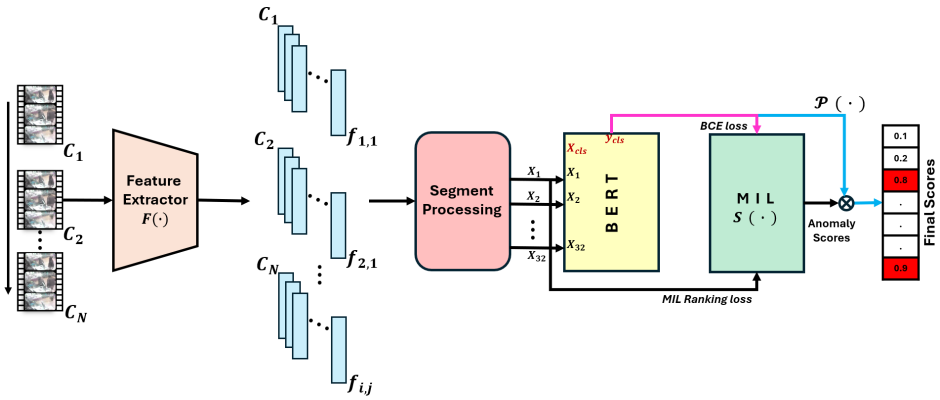

Figure 2. Block diagram of the BERT + MIL architecture. Pink arrow indicate the training phase, blue arrows represent the inference or testing phase, and black arrows show the common process.

3. BERT + MIL

Weijun et al. [10] proposed an innovative methodology for wVAD, presenting a novel perspective. They noticed that most previous research focused on how to select the correct snippets (both abnormal and normal snippets, $f_i, i = 1, 2, ..., N$) to train a wVAD model [9, 11, 12, 14, 22, 33]. However, these techniques often overlooked the power of full video classification, particularly in anomalous videos.

To address this limitation, [10] proposed incorporating explicit full video classification supervision into the MIL [9] and RTFM [12] frameworks. Full video classification refers to the task of categorizing a whole video as normal or abnormal using the aggregated features of all its snippets. This process involves aggregating the features extracted from individual video snippets into a unified classification embedding ($\mathbf{y}_{cls}$) that captures the global information of the video.

The authors suggest using BERT [23], a bidirectional self-attention model that can capture contextual relationships in both directions within a data sequence. This capability makes it particularly suitable for tasks that require an understanding of the global context, such as full video classification. In their proposal, the features of the snippets ($\mathbf{f}_i$) are combined into a classification embedding ($\mathbf{y}_{cls}$), which can then be used to significantly improve the performance of anomaly detection.

To train this model, the authors used different types of losses for MIL and RTFM. In the case of MIL, the binary cross-entropy (BCE) loss is applied to the BERT classification embedding ($\mathbf{y}_{cls}$), which helps to improve the accuracy of full video classification. Additionally, the MIL ranking loss, which includes smoothness and sparsity terms [9], is used to effectively separate the features of normal and anomalous snippets.

On the other hand, for RTFM, the methodology includes a robust feature magnitude loss, which selects the most representative snippets in terms of feature temporal magnitude. The combination of this loss with the BCE loss applied to the BERT classification embedding ($\mathbf{y}_{cls}$) enhances the model's ability to distinguish between normal and anomalous videos. In RTFM, the feature temporal

magnitude loss is used to ensure that the most informative snippets are selected, optimizing the separation between normal and anomalous features.

This approach demonstrated how full video classification can enrich and improve classical methodologies in wVAD. By integrating this technique into various anomaly detection frameworks, it not only achieves great versatility, but also significantly enhances model accuracy and performance. The authors highlight that their methodology can easily incorporate any feature extractor, providing considerable flexibility in its application. Therefore, selecting this model was ideal for our research, as it allowed us to evaluate how various feature extractors perform under this approach. This helped us identify the most effective feature extractor for weakly supervised video anomaly detection.

4. Experimental Setup

As stipulated in [10], the methodology is divided into two processes: training and inference. For training, a similar methodology was proposed, with slight variations due to the specific properties that each feature extractor requires. First, N raw video clips are extracted. These clips have a temporal length of 16 consecutive frames, non-overlapping, with a central crop of 224x224 pixels in RGB format. Subsequently, the features of each raw clip are extracted, resulting in N snippets of features denoted as $f_i, i = 1, 2, 3, \ldots, N$, with variable features dimensionality D depending on the extractor used.

For our experiments, we used the most common pre-trained feature extractors in wVAD:

- **C3D**: We followed the methodology of [9, 32], using features from the *fc6* layer with a dimensionality D of 4096 per snippet.
- **I3D**: Based on [10, 12], we used features from the *mix_5c* layer, resulting in a dimensionality D of 1024 per snippet.
- **TSM**: Following [33], we used TSM adjusted to 16-frame clips, with ResNet50 as the backbone [34, 37], obtaining a dimensionality D of 2048 per snippet.
- **UniFormer**: Following the approach proposed by [13], we also used UniFormer. However, unlike the authors where they used a version of UniFormer-B with a temporal length of 32 frames, we used UniFormer-B and UniFormer-S, both versions specifically designed to study 16 frames per clip. Both versions have a dimensionality D of 512 per snippet.

Once all the features f_i from each video are obtained, these features or snippets are segmented and normalized into T static segments denoted as $x_i = 1, 2, 3, \ldots, T$, following [10]. This is done because MIL requires an input of T fixed segments. That is, each video has N snippets that are then divided into T segments for processing. These segmented features x_i are sent to a BERT module, which subsequently generates the classification feature y_{cls}.

The next step, as stipulated in [10], is to provide MIL with the features x_i so that it can perform the normal anomaly detection process using MIL's specific loss functions, such as the ranking loss and the smoothness and sparsity terms [9, 10].

Additionally, the BCE loss is applied along with the classification feature y_{cls} during training, enriching the process. These loss functions allow MIL to evaluate the segments and assign scores based on their degree of anomaly, incorporating the y_{cls} feature. This process is repeated for each video until the training is complete.

During inference, the input features f_i are segmented into T segments. The video classification score $p(\hat{y}_{cls})$, where $\hat{y}_{cls}$ represents the class label predicted by BERT, is combined with the MIL segment scores $s(x_i)$, where x_i represents the i-th segment of the video, resulting in the final segment anomaly score. The equation to calculate the segment anomaly score is:

$$\text{score}(x_i) = s(x_i) \cdot p(\hat{y}_{cls}) \tag{1}$$

When only the MIL model is implemented, the video classification score $p(\hat{y}_{cls})$ is not used, and the final score is based solely on $s(x_i)$, as mentioned in [9, 10]. This process will allow us to evaluate the effectiveness of the resulting features and determine which feature extractor is the most suitable for our experimental setting.

5. Experiments

Table 1. Comparative Performance of Feature Extractors

Model Information			**Clips per Second**			
Model	**Params**	**GFLOPs**	**RTX 3090**	**Jetson Orin NX 16GB**	**Jetson Orin NX 8GB**	**Jetson Orin Nano 8GB**
UniFormer-B [32]	49.608M	64.202	31.50	2.42	2.09	1.63
C3D [20]	61.214M	154.120	34.22	3.68	3.17	2.67
UniFormer-S [32]	21.195M	28.717	66.08	5.40	4.64	3.54
TSM [21]	23.715M	66.110	85.42	8.64	7.59	6.04
I3D [18, 19]	**12.697M**	**27.877**	**101.23**	**16.92**	**15.27**	**11.34**

5.1. Dataset

We have selected the UCF-Crime dataset, proposed by Sultani et al. [9], which closely aligns with our objectives.

UCF-Crime is a comprehensive dataset for anomaly detection, consisting of 1,610 training videos (810 labeled as abnormal and 800 as normal) and 290 test videos (140 labeled as abnormal and 150 as normal), totaling 128 hours of real-world surveillance footage, both indoors and outdoors. This dataset includes 13 categories of anomalies, such as abuse, arrest, arson, assault, accident, burglary, explosion, fighting, shooting, stealing, shoplifting, and vandalism. Each category is labeled at the video level in the training set and at the frame level in the 290 test videos, allowing for detailed analysis.

For the evaluation of the UCF-Crime dataset, we follow previous works [9–14], using the frame-level Area Under the ROC Curve (AUC) as the evaluation metric. In our case, when obtaining the final scores from the inference, we assign the score of each segment X_i to all the snippets f_i that make up that segment. Subsequently, since each snippet consists of 16 frames, we repeat the score for each snippet 16 times. This means that all snippets that make up segment X_i will have the same score, and each frame within these snippets will repeat this score.

5.2. Analysis of Results

5.2.1. Feature Extractor Efficiency Analysis

Following the guidelines established in section 4, initial tests were conducted to determine which feature extractor offers the best performance when processing raw video clips (Table 1). All feature extractors were evaluated under the same conditions: the maximum available power capacity of the devices was used, and it was ensured that no other processes were active during the tests.

Table 2. Overview of Edge Devices

	Jetson Orin NX 16GB	**Jetson Orin NX 8GB**	**Jetson Orin Nano 8GB**
AI Performance	100 TOPS	70 TOPS	40 TOPS
GPU	1024-core NVIDIA Ampere architecture GPU with 32 Tensor Cores		
CPU	8-core Arm Cortex	6-core Arm Cortex	
Power	25W	20W	15W

For this test, each extractor processed 500 clips sized 16x3x224x224 (16 frames of 224x224 pixels in RGB format). The feature extractors were implemented on two types of devices: edge devices (Table 2), and a server with an RTX 3090 GPU with 24 GB of VRAM. Among the edge devices, the NVIDIA Jetson Orin Nano [35] offers up to 40 TOPS of AI performance with power options between 5W and 15W, featuring a 1024-core Ampere architecture GPU with 32 Tensor cores, making it ideal for real-world integration due to its low energy consumption and high video analysis capability. Similarly, the NVIDIA Jetson Orin NX [35], available in 8GB and 16GB versions, achieves 70 TOPS and 100 TOPS of AI performance respectively, with power options between 10W and 25W. Both versions also feature a 1024-core Ampere architecture GPU with 32 Tensor cores, making them suitable for VAD due to their high real-time processing capacity and energy efficiency.

As shown in Table 1, on the server with the RTX 3090, the feature extractor that achieved the best results was I3D, processing 101.23 clips per second with a computational complexity of 27.877 GFLOPs. This is significantly superior to the UniFormer-B, C3D, and TSM extractors. UniFormer-S, the closest competitor to I3D, processed 35.92% fewer clips (65.31 clips per second) with a similar computational complexity of 28.707 GFLOPs.

For the edge devices, the tests were replicated under the same conditions described above. On the Jetson Orin NX 16GB, I3D processed 16.92 clips per

second. On the Jetson Orin Nano, I3D achieved a rate of 11.34 clips per second, demonstrating to be a viable option for real-time implementations.

Although I3D is the most efficient feature extractor, we consider UniFormer-S to be a viable option as well. UniFormer-S processed clips efficiently on all devices, both on the server and peripheral devices, with a computational complexity similar to I3D. Considering that a traditional monitoring video is processed at 30 fps (1.87 clips per second in our conditions), all models, except for UniFormer-B, are optimal for real-time work and real-world environments. However, UniFormer-S and I3D may have an advantage as they do not require as much computational power for implementation on mobile devices.

Therefore, the I3D and UniFormer-S feature extractors are the most efficient options for feature extraction in real-world environments. Additionally, we consider that the most suitable edge device for integration in public transportation is the Jetson Orin NX in its 8GB version, which consumes 20W, offering a balance between the Jetson Orin Nano and the Jetson Orin NX 16GB.

Table 3. Performance Comparison of Different Feature Extractors in the Proposed Architecture by [10]

	32 segments		**48 segments**	
Model	**AUC-Bert+MIL**	**AUC-MIL**	**AUC-Bert+MIL**	**AUC-MIL**
C3D [32]	75.93	75.91	75.83	75.82
I3D [18, 19]	77.15	74.35	76.25	68.99
TSM [21]	77.36	70.60	78.10	71.07
UniFormer-S [32]	79.74	76.47	**79.68**	**77.95**
UniFormer-B [20]	**80.52**	**77.94**	79.45	77.88

5.2.2. Classification Efficiency of Feature Extractors

In this section, we select the best feature extractor by seeking a balance between processing speed, computational complexity and the quality of the resulting features. For these tests, feature extractors were integrated into the architecture proposed by [10], following the training and inference process described in section 4. To enhance our analysis, we proposed using varying segment lengths to evaluate the model and feature extractors' performance under different temporal conditions. Therefore, we conducted tests with both 32 and 48 fixed segments per video, as shown in Table 3. This approach allowed us to ensure a more robust assessment of their capabilities.

As shown in Table 3, the feature extractor that provides the most rich features is UniFormer-B, achieving an AUC of 80.52 at the frame level when processing 32 segments using the BERT + MIL modality. This makes it an optimal choice if precision is sought in controlled environments or laboratory analysis where computational consumption and analysis speed are not an issue, as this architecture proved to be the slowest across all devices (Table 1), making it impractical for real-world environments such as public transportation.

While I3D was the model that could process the most clips in the shortest time and with the least resources, the quality of its features is good but not optimal.

We observed that increasing the number of segments from 32 to 48 resulted in a decrease in I3D's AUC from 77.15 to 76.25. This drop in performance suggests that I3D is sensitive to segment length. Additionally, it was only superior to C3D, which has a relatively less complex architecture and consequently produces slightly lower quality features. While it is not a deficient model, we consider that I3D may be more suitable for real-world environments that do not require immediate action, such as highway accidents, fires, police abuse, etc.

Therefore, we consider that UniFormer-S might be the best option based on the existing evidence. According to the data shown in Tables 3 and 1, this model achieves a good balance among computational complexity, analysis speed, and performance in predicting anomalous events. It achieved the second-highest AUC (79.74%) in its BERT+MIL mode with 32 fixed segments and demonstrated notable robustness to temporal changes, maintaining consistent performance across all anomaly detector modes evaluated in this study. This makes it the best option for integration in real-world environments, such as public transportation, as well as a viable option for implementation in mobile devices.

6. Conclusion

This study evaluated feature extractors in VAD to determine their efficiency and effectiveness in terms of accuracy and computational performance, with a practical focus on their applicability for public transportation security. Using the UCF-Crime database, tests were conducted on a server with an RTX 3090 GPU and edge devices like the Jetson Orin Nano and Jetson Orin NX. I3D proved to be the fastest, processing 101.23 clips per second with a complexity of 27.877 GFLOPs, but its ability to handle temporal variations was limited. UniFormer-B achieved the highest AUC (80.52%) with BERT+MIL using 32 segments, although its high computational complexity makes it more suitable for controlled environments. UniFormer-S balanced computational complexity, analysis speed, and anomaly prediction accuracy, achieving an AUC of 79.74% with BERT+MIL using 32 segments. Given its balanced performance, we confirm that the Jetson Orin NX 8GB at 20W would be an excellent pairing with UniFormer-S for mobile integration. Considering its robustness and consistency, this feature extractor is the best option for real-world environments and mobile devices. Future work will focus on developing and optimizing a VAD model applicable in real-time conditions, detecting anomalies in real-time using both UniFormer-S and the Jetson Orin NX 8GB, which have proven to be ideal for these purposes. Additionally, further studies will address more comprehensive evaluations and additional testing to enhance the robustness and accuracy of the models.

References

[1] Instituto Nacional de Estadística y Geografía (INEGI). Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública (ENVIPE) 2023. INEGI. https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2023/ENVIPE/ENVIPE23.pdf. Published 2023. Accessed November 19, 2023.

[2] Newton A. Crime on Public Transport. In: Crime Prevention and Security Management, edited by Tilley N, Bruinsma GJN, Sidebottom A. Springer; 2014:709-720. doi:10.1007/978-1-4614-5690-2_301.

[3] Medina S. Towards Safe and Empowering Streets and Public Transport Systems for Women and Girls in Latin America. [Online]. Available at: https://womenmobilize.org/towards-safe-and-empowering-streets-and-public-transport-systems-for-women-and-girls-in-latin-america/. Published Dec, 2022. Accessed [May 2023].

[4] Instituto Mexicano del Transporte. Sistemas inteligentes de transporte. Instituto Mexicano del Transporte. https://imt.mx/resumen-boletines.html?IdArticulo=127&IdBoletin=41. Published May, 2023. Accessed [Dec 2023].

[5] Palma Montes M. ¿Cómo disminuir los asaltos en el transporte público? Alcaldes de México. https://www.alcaldesdemexico.com/notas-principales/como-disminuir-los-asaltos-en-el-transporte-publico/. Published September 23, 2021. Accessed [May 2024].

[6] Gobierno de la Ciudad de México. CON ESTRATEGIAS DE PREVENCIÓN Y VIDEOVIGILANCIA, METROBÚS BRINDA MAYOR SEGURIDAD A LA CIUDADANÍA. Metrobús CDMX. https://www.metrobus.cdmx.gob.mx/comunicacion/nota/BMB160622. Published June 16, 2022. Accessed [May 2024].

[7] Sol de Toluca. Proyectan nuevo sistema de vigilancia con el C-5 en el transporte del Edomex. Sol de Toluca. https://www.elsoldetoluca.com.mx/local/proyectan-nuevo-sistema-de-vigilancia-con-el-c-5-en-el-transporte-del-edomex-10839990.html. Published October 13, 2023. Accessed [May 2023].

[8] El Universal. "Cámaras en el transporte público no tienen internet": Movilidad del Edomex. El Universal. https://www.eluniversal.com.mx/edomex/camaras-en-el-transporte-publico-no-tienen-internet-movilidad-del-edomex/. Published March 11, 2024. Accessed [May 2024].

[9] Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 6479-6488

[10] Tan W, Yao Q, Liu J. Overlooked video classification in weakly supervised video anomaly detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2024. p. 202-210.

[11] Zhong JX, Li N, Kong W, Liu S, Li TH, Li G. 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), 2019, pp. 1237-1246.

[12] Tian Y, Pang G, Chen Y, Singh R, Verjans JW, Carneiro G. Weakly-Supervised Video Anomaly Detection With Robust Temporal Feature Magnitude Learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4975-4986.

[13] Karim H, Doshi K, Yilmaz Y. Real-Time Weakly Supervised Video Anomaly Detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024. p. 6848-6856.

[14] Al-Lahham N, Tastan N, Zaheer MZZ, Nandakumar K. A Coarse-to-Fine Pseudo-Labeling (C2FPL) Framework for Unsupervised Video Anomaly Detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024. p. 6793-6802.

[15] Luo W, Liu W, Gao S. A revisit of sparse coding based anomaly detection in stacked RNN framework. In: Proceedings of the IEEE International Conference on Computer Vision. 2017. p. 341-349.

[16] Zhang Y, Lu H, Zhang L, Ruan X, Sakai S. Video anomaly detection based on locality sensitive hashing filters. Pattern Recognition. 2016;59:302-311.

[17] Chen C, Xie Y, Lin S, Yao A, Jiang G, Zhang W, Qu Y, Qiao R, Ren B, Ma L. Comprehensive Regularization in a Bi-Directional Predictive Network for Video Anomaly Detection. Proceedings of the AAAI Conference on Artificial Intelligence. 2022;36(1):230-8. doi:10.1609/aaai.v36i1.19898.

[18] Park J, Kim J, Han B. Learning to Adapt to Unseen Abnormal Activities under Weak Supervision. Proceedings of the Asian Conference on Computer Vision (ACCV). 2020.
[19] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. p. 6299-6308.
[20] Li K, Wang Y, Gao P, Song G, Liu Y, Li H, Qiao Y. UniFormer: Unified transformer for Efficient Spatiotemporal Representation Learning. CoRR. 2022;abs/2201.04676. Available from: [https://arxiv.org/abs/2201.04676].
[21] Lin J, Gan C, Han S. Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019. p. 7083-7093.
[22] Wu JC, Hsieh HY, Chen DJ, Fuh CS, Liu TL. Self-supervised sparse representation for video anomaly detection. In: European Conference on Computer Vision; 2022. pp. 729-745.
[23] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018.
[24] Al-lahham A, Tastan N, Zaheer MZZ, Nandakumar K. A Coarse-to-Fine Pseudo-Labeling (C2FPL) Framework for Unsupervised Video Anomaly Detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024. p. 6793-6802.
[25] Basharat A, Gritai A, Shah M. Learning object motion patterns for anomaly detection and improved object detection. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2008. p. 1–8.
[26] Medioni G, Cohen I, Br´emond F, Hongeng S, Nevatia R. Event detection and analysis from video streams. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2001;23(8):873–889.
[27] Fang Z, Liang J, Zhou JT, Xiao Y, Yang F. Anomaly detection with bidirectional consistency in videos. IEEE Transactions on Neural Networks and Learning Systems. 2020.
[28] Ionescu RT, Smeureanu S, Alexe B, Popescu M. Unmasking the abnormal events in video. In: Proceedings of the IEEE International Conference on Computer Vision. 2017. p. 2895–2903.
[29] Abati D, Porrello A, Calderara S, Cucchiara R. Latent space autoregression for novelty detection. IEEE Transactions on Neural Networks and Learning Systems. 2020.
[30] Bergman L, Hoshen Y. Classification-based anomaly detection for general data. arXiv preprint arXiv:2005.02359. 2020.
[31] Bergmann P, Fauser M, Sattlegger D, Steger C. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020.
[32] Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE international conference on computer vision. 2015; p. 4489-4497.
[33] Öztürk HI, Can AB. Adnet: Temporal anomaly detection in surveillance videos. In: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV. Springer; 2021. p. 88-101.
[34] Lin J, Gan C, Han S. Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019; p. 7083-7093.
[35] NVIDIA. Jetson Orin Nano Series. [Online]. Available at: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/. [May 2024].
[36] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; p. 770-778.
[37] Carreira J, Noland E, Banki-Horvath A, Hillier C, Zisserman A. A short note about kinetics-600. arXiv preprint arXiv:1808.01340. 2018.