Detection and Classification of Abnormal Human Actions for Video Surveillance on Edge Devices

1st Jonathan Flores Monroy
Sección de Estudios de Posgrado e Investigación
Instituto Politécnico Nacional
Mexico City, Mexico
jfloresm1510@alumno.ipn.mx

2nd Gibran Benitez Garcia
Graduate School of Informatics and Engineering
The University of Electro-Communications
Tokyo, Japan
gibran@ieee.org

3rd Mariko Nakano Miyatake
Sección de Estudios de Posgrado e Investigación
Instituto Politécnico Nacional
Mexico City, Mexico
mnakano@ipn.mx

4th Hiroki Takahashi
Artificial Intelligence eXploration Research Center
The University of Electro-Communications
Tokyo, Japan
rocky@inf.uec.ac.jp

***Abstract*—The increase in violence in urban areas has created an urgent need for systems capable of accurately identifying actions that endanger citizens. In this context, video surveillance systems supported by video anomaly detection approaches have served as a foundation for developing automatic solutions to identify unusual behaviors; however, most of these methods focus exclusively on detection and do not address the detailed classification of anomalies. This work proposes a modular framework for the joint detection and classification of *Abnormal Human Actions*, understood as anomalous actions caused by humans that pose a threat to public safety. The proposed approach, evaluated on an edge device, achieved an accuracy of 59.02% in *Abnormal Human Actions* classification, a detection rate of 80.0% on the UCF-Crime dataset, and a processing speed of 283.49 fps, demonstrating its potential for deployment in real-world, resource-constrained video surveillance environments.**

***Index Terms*—video anomaly detection, edge devices, online detection and classification, abnormal human actions, modular framework.**

I. Introduction

Crowded streets, public transport systems, and recreational spaces have become frequent settings for crimes such as assaults, robberies, and physical aggression (e.g., Mexico [1]). This problem is particularly severe in Latin America, where criminal organizations increase the frequency and severity of violent incidents. According to Gallup's Global Law and Order Report [2], only 42% of Latin Americans feel safe walking alone at night, the lowest rate worldwide. Many of these crimes occur in front of surveillance cameras; however, timely detection is hindered by the need for constant human monitoring, which entails high costs, operator fatigue, and a high risk of error. This creates a demand for systems capable of automatically detecting and classifying abnormal actions in online scenarios, enabling faster and more efficient responses.

Video Anomaly Detection (VAD) [3]–[11] addresses this need by identifying events that deviate from normal behavior in video sequences. However, most approaches focus solely on temporally detecting anomalous behavior [3]–[9], without classifying its type, key for determining the appropriate and timely response, especially if the situation involves injuries or immediate threats. Many methods also rely on offline processing [3], [4], [7], [10], [11], where predictions are generated after analyzing a pre-recorded video. This limits their applicability in scenarios that require online processing, where decisions must be made as events unfold. While some recent works incorporate classification [6], [10], [11], they often operate offline and overlook practical constraints such as computational efficiency for edge deployment.

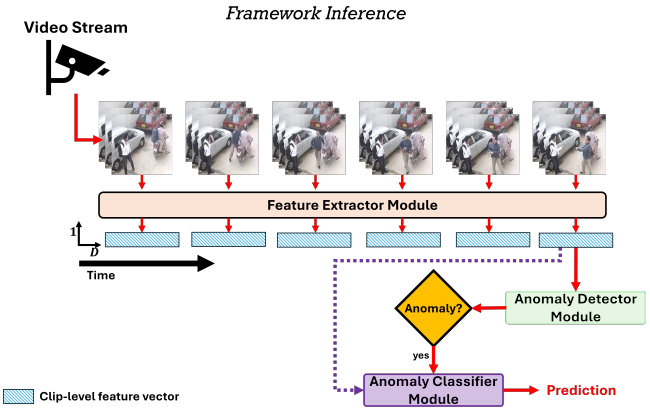

Fig. 1. Overview of the proposed modular framework for clip-level detection and classification of Abnormal Human Actions. The input video stream is divided into consecutive clips, encoded by the feature extractor module, evaluated by the anomaly detector, and (if identified as anomalous) classified into a specific category by the anomaly classifier.

To address current limitations in video anomaly detection, we propose a modular framework for the online detection and classification of Abnormal Human Actions in video. The system is composed of three main components: (i) a feature extractor based on Expanded 3D ConvNet (X3D) in its Small version (X3D-S) [12], which processes video clips and

generates compact spatiotemporal feature representations; (ii) an anomaly detector adapted from the Coarse-to-Fine Pseudo-Labeling method (C2FPL) [5], which assigns an anomaly score to each feature vector; and (iii) a classifier derived from Flores et al. [6], responsible for identifying the specific type of abnormal behavior. Only features with scores above a predefined threshold are classified, reducing unnecessary computation and focusing on truly anomalous content. This design supports online operation and enables fine-grained recognition of human-related anomalies. An overview of the proposed system is shown in Figure 1.

We evaluate the proposed framework on the UCF-Crime dataset [3], which includes 13 different classes of abnormals. The complete system achieves an AUC of 80%, remaining competitive with state-of-the-art methods while enabling online detection. For classification, we compare two settings: one using all 13 classes, and another restricted to 9 human-centered anomalies (e.g., assault, robbery, fighting, etc.), excluding classes such as arrest, explosion, arson and road accidents. The latter setting yields higher accuracy (59.02% vs. 54.14%), showing the benefit of focusing on human-related events. Designed for edge devices, the framework operates online at 283.49 FPS on a Jetson Orin NX (8 GB RAM) [13] when excluding I/O overhead, and at 27.42 FPS when including it.

The main contributions of this work are:

- A lightweight and modular framework for the online detection and classification of Abnormal Human Actions in streaming video, specifically designed for deployment on edge devices.
- A full evaluation of the proposed framework, demonstrating low computational cost and high efficiency, supporting online processing on edge devices.
- Empirical validation using both a focused (9-class) and complete (13-class) setting on UCF-Crime, highlighting improved classification performance in human-centered scenarios and high processing efficiency on edge devices.

II. Related Work

1) Anomaly Detectors: Weakly Supervised Learning [3], [4], [6]–[11] is a dominant paradigm in VAD, requiring only video-level labels and no frame-level annotations. A widely used approach in this setting is Multiple Instance Learning (MIL) [3], where each video is divided into fixed-length *segments*, grouped into positive bags (anomalous videos) and negative bags (normal videos). The objective is to identify the most relevant abnormal segments using ranking-based loss functions. To integrate long-range temporal context, Tan et al. [4] proposed augmenting MIL with Bidirectional Encoder Representations from Transformers (BERT) [14], enabling joint modeling of local anomaly scores and global video context. Al-Lahham et al. [5] introduced the Coarse-to-Fine Pseudo-Labeling (C2FPL) approach, which generates initial pseudo-labels via hierarchical clustering and refines them statistically at the clip level before training a clip-wise detector. Gao et al. [7] developed a compact detection pipeline using reduced variants of X3D [12] with (2+1)D convolutions and the Performer attention mechanism [15], classifying whole videos as normal or abnormal while keeping computational cost low. Most existing methods aim to infer, over time, where evidence of an abnormal event is present within the video. However, many still require processing the whole video before producing results, which limits their applicability in online or progressive-analysis scenarios.

2) Multi-category Classifiers: The recognition of anomaly classes enables models to not only identify abnormal events but also specify their nature. Majhi et al. [10] proposed a dual-branch architecture: one branch detects anomalies using MIL [3], while the other classifies them with attention-based processing and a softmax output. Wu et al. [11] refined Contrastive Language–Image Pretraining (CLIP) [16] features via temporal and semantic alignment, mapping them to text descriptions of event categories. Flores et al. [6] presented a sequential framework with Unified Transformer in its S version (UniFormer-S) for feature extraction, a C2FPL detector [5], and a fully connected classifier. Classifier training was improved by offline pre-selection of high-scoring abnormal clip features using MIL-BERT [4]. Despite their effectiveness, most classification-capable approaches still incur high computational costs and, to our knowledge, have not been evaluated exclusively on human-centric abnormal action classes.

III. Proposed Method

To effectively identify and classify anomalous events in video streams, it is essential to adopt a processing strategy that operates as quickly as possible while using minimal computational resources, without reducing detection accuracy. We build upon the online inference approach proposed in [6], which offers a modular and efficient structure suitable for continuous monitoring scenarios.

1) Framework Inference: The proposed framework, illustrated in Figure 2, begins with the Raw Video Preprocessing stage. The input video stream V is divided into consecutive, non-overlapping clips of T' frames, obtained by selecting one frame every G frames, each resized to a fixed spatial resolution $H \times W$ and center-cropped to reduce computational cost. Each resulting clip is represented as a tensor $\mathbb{R}^{h \times w \times T' \times Ch}$, where $h \times w$ is the cropped resolution. Subsequently, each $clip_i$ is processed by the feature extractor $F(\cdot)$. This stage outputs a single feature vector $f_i \in \mathbb{R}^{1 \times D}$, where D corresponds to the dimensionality of the spatiotemporal features extracted for each segment of the clip. After feature extraction, the anomaly detector $S(\cdot)$ evaluates each f_i and assigns an anomaly score $score_i \in [0, 1]$. If $score_i < thr$, the clip is labeled as normal; otherwise, f_i is passed to the multi-category classifier $\mathcal{C}(\cdot)$, which assigns it to one of the predefined anomaly classes.

2) Feature Extractor: X3D [12] is an architecture designed for efficient video action recognition, built upon a minimal 2D backbone that is progressively expanded along six key axes: clip temporal duration, frame sampling rate, spatial resolution, channel width, network depth, and bottleneck width. It processes short video clips of T' frames obtained through uniform

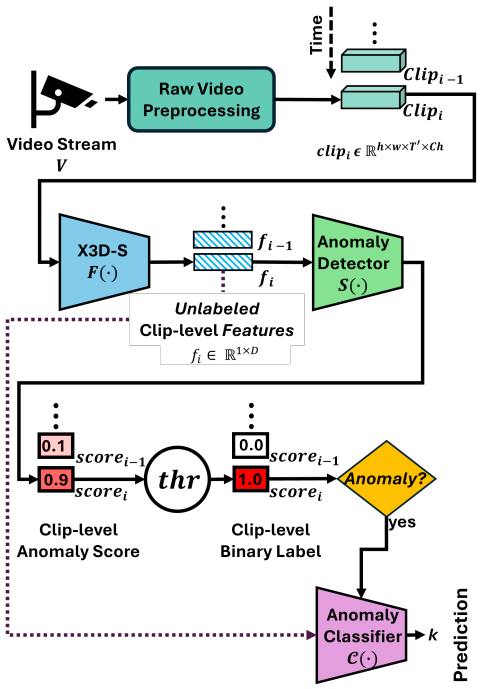

Fig. 2. General flowchart of the proposed modular framework for clip-level anomaly detection and classification. The video stream V is divided into consecutive clips, resized, and center-cropped. Each clip is encoded by the pre-trained X3D feature extractor $F(\cdot)$ [12], evaluated by the anomaly detector $S(\cdot)$, and, if deemed anomalous, classified by $\mathcal{C}(\cdot)$ into a specific category. The purple arrow indicates the feature vector sent to the anomaly classifier when the anomaly score exceeds the predefined threshold. The design enables continuous and progressive inference.

sampling and employs factorized (2+1)D convolutions, separating spatial and temporal processing to reduce computational cost while improving motion modeling. During training, each axis is expanded independently and its impact on accuracy and GFLOPs is evaluated, retaining only the expansion with the best performance–efficiency trade-off, and repeating this process until a target computational budget is reached. This results in a family of models: X3D-L [12], deeper and more accurate; X3D-M [12], a balanced compromise; and X3D-S, the most lightweight version with only 2.84 GFLOPs and 2.3 million parameters, making it an ideal choice for *edge devices* due to its optimal balance between accuracy and efficiency for continuous video analysis. In our work, X3D-S [12] is adopted as the feature extractor within the proposed modular framework (Fig. 1), leveraging its low computational cost and ability to process clips sequentially, thus enabling efficient detection and classification of Abnormal Human Actions without compromising performance in continuous monitoring scenarios.

3) Anomaly Detector S(·): We adopt the C2FPL architecture [5] (Figure 3), composed of two fully connected (FC) layers with ReLU activation and dropout, followed by two self-attention layers and a final sigmoid for binary classification. C2FPL operates under a weakly supervised paradigm, requiring video-level labels to form normal and anomalous subsets,

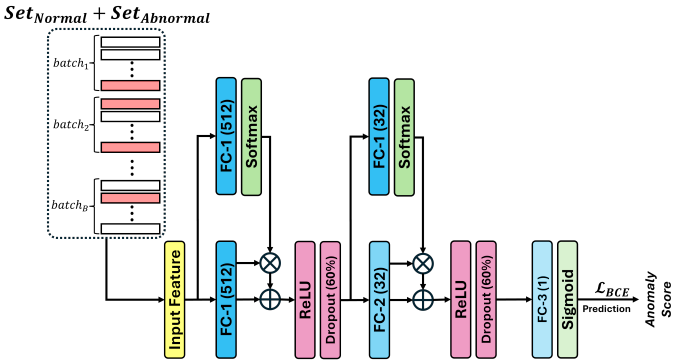

Fig. 3. Detailed architecture of Anomaly Detector $S(\cdot)$.

as provided in datasets like UCF-Crime [3]. All videos are segmented into non-overlapping clips, resized, center-cropped, and processed with an Inflated 3D ConvNet (I3D) feature extractor [17]. Normal features ($y = 0$) come directly from the normal training set, while anomalous features ($y = 1$) are selected from the anomalous subset using a probabilistic strategy based on hierarchical clustering and statistical refinement. In our aproach, we maintain the same detector architecture but replace the I3D features with those extracted by X3D-S to reduce computational cost. Moreover, instead of applying the original probabilistic selection, we adapt the idea from [6] of using a pre-trained VAD model to identify and select, at the clip level, the feature vectors with the highest anomaly scores within each anomalous video, as detailed in Subsection III-A.

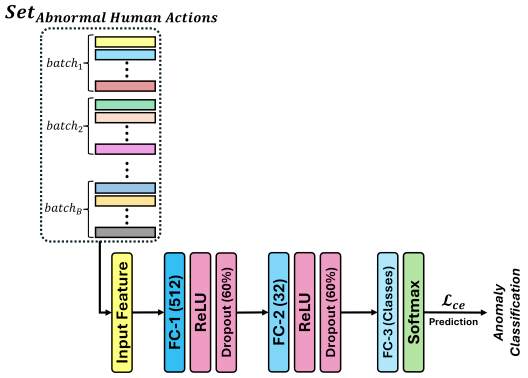

Fig. 4. Detailed architecture of Multi-Category Classifier $C(\cdot)$.

4) Multi-Category Classifier $\mathcal{C}(\cdot)$: For our classifier, we use the anomalous feature vectors selected during the abnormal subset creation process, each labeled according to its predefined human abnormal action class. This is why a dataset is required where anomalous videos are labeled by anomaly type, as is the case with UCF-Crime [3]. The architecture, shown in Figure 4 and originally proposed in [6], is a fully connected network with three linear layers with 512 input units, 32 hidden units, and an output layer matching the number of anomaly classes. ReLU activation is applied between layers, with 60% dropout in the hidden layer, and a Softmax output producing class probabilities.

A. Training and Integration Stage

1) Generation of Abnormal Set: To ensure robust training, it is essential to first generate an anomalous subdataset containing the most representative features of each Abnormal Human Action class. This subdataset serves as the basis for independently training the anomaly detector $S(\cdot)$ and the multi-class classifier $\mathcal{C}(\cdot)$, which are later integrated into the proposed framework. For this process, a training set composed of videos labeled as anomalous and classified by type of abnormality (focused on Abnormal Human Actions) is required, such as the one provided by UCF-Crime [3], which includes classes like *Fighting*, *Robbery*, and *Shooting*. Following the approach of Flores et al. [6], each video V_j^k (where $k \in \{1, \ldots, K\}$ is the class index and j is the video index within that class) is divided into N_j clips $\text{clip}_{j,i}$ (where i is the clip index) of T' frames (sampled every G frames), preprocessed, and represented as vectors $f_{j,i} \in \mathbb{R}^{1\times D}$, where D is the dimensionality of the spatiotemporal representation extracted using X3D-S, unlike the original version, which employs UniFormer-S [18] as the backbone. These vectors are grouped into $\hat{T}$ fixed segments ($X_j^k \in \mathbb{R}^{\hat{T}\times D}$) by dividing N_j into contiguous groups of size $N' = N_j/\hat{T}$ and averaging within each group, applying rewind strategy [6] if N_j is not divisible. Next, an *offline* VAD model, MIL+BERT [4], assigns each segment t (where $t \in \{1, \ldots, \hat{T}\}$ is the fix segment index) an anomaly score $\text{score}(x_{j,t}) \in [0, 1]$, which is then projected back to the clip level as $\mathbf{p} = [\text{score}_{j,1}, \ldots, \text{score}_{j,N_j}]$. Unlike the original strategy that selects the feature vectors with the highest score, we use a sliding window that covers a proportion λ of the sequence to find the region with the greatest score variation, and take its clips as the most representative of the anomaly. This process is repeated for each video of each class k until the anomalous subdataset used for training both models is generated.

2) Training of $S(\cdot)$ and $\mathcal{C}(\cdot)$: Once the anomalous subdataset is generated, it is used to train the anomaly detector $S(\cdot)$ and the multi-class classifier $\mathcal{C}(\cdot)$.

Anomaly Detector $S(\cdot)$. It is trained with feature vectors from normal videos and with anomalous vectors from the subdataset, labeled as $y = 0$ and $y = 1$, respectively. Training batches are randomly sampled from the combined set without preserving temporal order. The optimization is performed using the Binary Cross-Entropy (BCE) loss function:

$$\mathcal{L}_{BCE} = -\frac{1}{B}\sum_{i=1}^{B}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad (1)$$

where B is the batch size, y_i is the ground-truth label, and $\hat{y}_i$ is the predicted probability of the feature vector being anomalous.

Multi-class Classifier $\mathcal{C}(\cdot)$. It is trained exclusively with the anomalous features vectors from the subdataset, assigning each feature vector the label corresponding to its abnormal action class. Training batches are randomly sampled from this set without preserving temporal order. The optimization is performed using the Categorical Cross-Entropy (CE) loss function:

$$\mathcal{L}_{CE} = -\sum_{k=1}^{K} y_k \log(\hat{y}_k) \quad (2)$$

where K is the number of anomaly classes, y_k is the ground-truth label for class k, and $\hat{y}_k$ is the predicted probability for that class.

3) Integration Step: Once trained, the detector $S(\cdot)$ and classifier $\mathcal{C}(\cdot)$ are integrated into the feature extraction pipeline for the inference stage described in Sussection III-1.

IV. EXPERIMENTS

1) Dataset: In this study, we use the UCF-Crime dataset [3], which consists of 1,900 videos split into 1,610 for training (810 anomalous and 800 normal) and 290 for testing (140 anomalous and 150 normal). The dataset covers 13 anomaly categories, and all videos have a resolution of 240×320 pixels at 30 fps. The main focus of this work is on the detection and classification of human-related anomalous events, given their high incidence and impact on public safety, particularly in urban environments [1], [2]. We specifically selected nine categories: *abuse*, *assault*, *fighting*, *robbery*, *shooting*, *burglary*, *shoplifting*, *stealing*, and *vandalism*. These represent criminal or violent behaviors involving direct human interaction or intentional property damage, which typically require immediate response in monitoring systems. However, to ensure fair comparative evaluation with previous works, the complete set of 13 UCF-Crime categories was also used in the validation experiments, maintaining consistency with the standard evaluation protocol reported in the SOTA [3]–[11].

2) Evaluation Metric: The evaluation of the proposed framework follows the metrics commonly adopted in the current state of the art. For detection, we report the frame-level Area Under the ROC Curve (AUC), widely used in anomaly detection tasks to measure the ability to discriminate between normal and anomalous frames [3]–[9]. For classification, we use video-level accuracy as the main metric, following the current SOTA [3], [6], [11]. Finally, efficiency is evaluated through frames processed per second (FPS), number of model parameters, and floating-point operations (GFLOPs), standard indicators for assessing suitability in online inference environments and edge devices [9].

3) Implementation Details: X3D-S was used as feature extractor with $T' = 13$ and $G = 6$. The input frames was center-cropped to 182×182 ($h \times w$). Length fix segment $\hat{T} = 32$, using $K = 9$ classes for our method and $K = 13$ for state-of-the-art comparison, both from UCF-Crime. Feature dimensionality was D $= 192$ and $\lambda = 0.2$. The anomaly detector used Adam optimizer (lr $= 0.001$, batch size = 64, 100 epochs, dropout = 60%), while the classifier used AdamW (lr $= 0.0001$, batch size = 64, 2000 epochs). The threshold thr was set to the 90th percentile based on ROC-derived AUC scores. Training was performed on an Intel Core i7-13700F with 32 GB RAM and RTX 4060 Ti (16 GB). Inference and efficiency tests were run on a Jetson Orin NX (8 GB, 20 W).

4) Comprehensive Performance Analysis of the Proposed Framework: To evaluate the performance of the proposed architecture, its detection and classification capabilities were analyzed in two scenarios: one focused on *Abnormal Human Actions* and another using the complete 13-class configuration of UCF-Crime. In addition, the associated computational cost was considered to determine its initial feasibility for integration into edge devices and, therefore, its potential use in real environments. Furthermore, this section examines how the proper selection of the feature extractor influences the balance between accuracy and efficiency of the framework. For this purpose, our proposed configuration using X3D-S was compared against state-of-the-art reference architectures (UniFormer-S [18] and I3D [17]) and a higher-capacity variant from the same family (X3D-M [12]). These feature extractors were chosen because they have been previously employed in solutions designed for online environments [5], [6], [9], making them potentially competitive alternatives to our proposal. For each feature extractor, the anomaly detection module $S(\cdot)$ and the multi-category classification module $C(\cdot)$ were trained under the same conditions defined by the proposed method (Section III), ensuring a fair and consistent comparison across all evaluated configurations.

Table I presents the impact of feature extractor selection on detection performance (AUC), classification accuracy (ACC), and computational cost (GFLOPs) in both scenarios. In *Abnormal Human Actions*, X3D-S achieved the highest ACC (59.02%), outperforming all other configurations, with an AUC of 60.72% slightly below UniFormer-S (61.97%) and I3D (62.22%). Although the AUC is marginally lower, the computational cost of X3D-S (2.844 GFLOPs) is approximately 90% lower than that of UniFormer-S (28.871) and I3D (27.858), which could facilitate its use in resource-constrained environments, allowing more sequences to be processed and alerts to be issued more consistently for human-related events, even if detection occurs with a slight delay compared to other models.

TABLE I
PERFORMANCE COMPARISON OF THE MODULAR FRAMEWORK USING DIFFERENT BACKBONE NETWORKS ON THE UCF-CRIME DATASET [3].

Backbone	**GFLOPs** ↓	**9 Classes***		**13 Classes***	
		AUC (%) ↑	**ACC** (%) ↑	**AUC** (%) ↑	**ACC** (%) ↑
UniFormer-S	28.871	61.97	53.02	80.55	53.79
I3D	27.858	62.22	51.29	76.04	37.97
X3D-M	**6.473**	**60.47**	**54.31**	**83.52**	**54.14**
X3D-S	**2.844**	**60.72**	**59.02**	**80.00**	**54.48**

* In both settings, the evaluation includes complete test videos containing normal and abnormal instances. The 9-class setup considers a subset of Abnormal Human Actions, while the 13-class setting covers all categories in UCF-Crime.

In the complete 13-class UCF-Crime configuration, X3D-M achieved the highest AUC (83.52%), followed by UniFormer-S (80.55%) and X3D-S (80.00%). However, the ACC of X3D-S (54.48%) remained practically identical to that of X3D-M (54.14%), with a 56% lower computational cost (2.844 vs. 6.473 GFLOPs). The inclusion of non-human classes introduces greater visual and motion variability, leading to an overall reduction in accuracy across all configurations.

Overall, these results show that X3D-S combines higher classification accuracy in *Abnormal Human Actions* with a very low computational cost, which may be relevant for systems that prioritize classification reliability and alert triggering in critical human-related events, while maintaining an acceptable balance in detection performance.

5) Performance in Abnormal Human Action Identification Applications: While initial estimates of computational cost and classification accuracy are informative, they do not alone confirm suitability for deployment. This section evaluates the feasibility of integrating the proposed framework into edge devices, considering accuracy, processing time, and computational demand. Tests were conducted on an NVIDIA Jetson Orin NX (8 GB, 20 W), a platform relevant for public transportation, retail, and other resource-constrained environments. I/O measurements were also included using a generic camera to simulate realistic acquisition conditions. Table II reports results for the *Abnormal Human Actions* scenario. With X3D-S, the framework achieved the highest accuracy (59.02%) and the fastest inference speeds, both with I/O (27.42 fps) and without I/O (283.49 fps), outperforming UniFormer-S and I3D by 5.23 percentage points in accuracy and by approximately 52% in I/O speed. The X3D-M variant offered slightly lower accuracy (54.31%) but remained competitive in processing (26.94 fps with I/O).

TABLE II
PERFORMANCE COMPARISON OF THE PROPOSED FRAMEWORK FOR THE CLASSIFICATION OF ABNORMAL HUMAN ACTIONS ON JETSON ORIN NX (8 GB) [13].

Backbone	**ACC** (%) ↑	**With I/O Time** (fps) ↑	**Without I/O Time** (fps) ↑
UniFormer-S	53.79	18.06	67.56
I3D	37.97	23.54	235.35
X3D-M	**54.31**	**26.94**	**238.62**
X3D-S	**59.02**	**27.42**	**283.49**

For X3D-S, the I/O-inclusive speed approaches the capture rate used in testing ($\approx 29.37\ fps$), suggesting minimal end-to-end latency. The observed gap between with- and without-I/O results highlights the impact of data transfer and preprocessing. Combining these outcomes with detection performance and computational cost from Table I, X3D-S emerges as the most consistent configuration: 59.02% ACC, 60.72% AUC, 2.844 GFLOPs, and top inference speed (27.42 fps with I/O, 283.49 fps without I/O). This balance of accuracy, detection capability, efficiency, and processing speed reflects the integrated design of the proposed framework, making it suitable for deployment in resource-limited environments. Trained on real and weakly controlled data such as UCF-Crime [3], the system also shows inherent robustness to complex conditions, tolerating variations in illumination, crowd density, and occlusions with acceptable performance in real-world surveillance.

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE FULL UCF-CRIME DATASET [3].

Method	Approach	GFLOPs ↓	AUC (%)↑	ACC (%) ↑
Sultani et al. [3]	C3D + MIL + NN classifier*	∼61.638	75.41	23.00
Sultani et al. [3]	TCN + NN classifier*	-	-	28.40
Tan et al. [4]	I3D + MIL-BERT	∼38.95	82.69	-
Wu et al. [11]‡	CLIP + Top-K† + Soft attention‡‡	-	86.40	41.43
Gao et al. [7]	X3D-L + FC	∼26.503	91.34	-
Flores et al. [6]	UniFormer-S + C2FPL + FC	28.873	82.27	58.96
Al et al. [5]	I3D + C2FPL	∼17.824	83.40	-
Ours	**X3D-M + C2FPL + FC**	**6.473**	**83.52**	**54.14**
Ours	**X3D-S + C2FPL + FC**	**2.844**	**80.00**	**54.48**

* Nearest Neighbor classifier. † Top-K classifier. ‡ Soft-attention classifier. ‡‡ Evaluated only on anomalous events ("Normal" class excluded).

6) Comparison with State-of-the-Arts: Table III presents the comparison of our best configurations, X3D-M (6.47 GFLOPs, AUC = 83.52%, ACC = 54.14%) and X3D-S (2.84 GFLOPs, AUC = 80.00%, ACC = 54.48%), against state-of-the-art methods on the UCF-Crime dataset. These results show that our framework achieves a balanced trade-off between classification accuracy, detection performance, and computational efficiency, making it suitable for edge deployment. Compared to Gao et al. [7] (26.50 GFLOPs, AUC = 91.34%), which also uses an X3D variant, our method is more efficient, includes classification, and operates clip-by-clip without requiring full-video processing. Wu et al. [11] (AUC = 86.40%, ACC = 41.43%) achieves higher detection scores but depends on complete video access for both detection and classification, which limits integration into streaming-based pipelines, and does not report GFLOPs. In contrast, our method supports online processing with substantially lower computational requirements. Relative to Flores et al. [6] (28.87 GFLOPs, AUC = 82.27%, ACC = 58.96%), our framework reduces computational cost while achieving superior detection performance among online methods. Finally, compared to Al et al. [5] (17.82 GFLOPs, AUC = 83.40%), which is limited to detection only, our approach also performs classification, increasing its applicability in scenarios where the anomaly type must be identified.

V. CONCLUSION

This work presented a modular framework for the detection and classification of *Abnormal Human Actions*, evaluated in both reduced and complete scenarios using the UCF-Crime dataset. The configuration with the X3D-S backbone achieved the best balance between accuracy, computational cost, and inference speed on edge devices, positioning it as a viable option for resource-constrained environments. However, the detection and classification of Abnormal Human Actions remains a complex challenge that requires further research to improve robustness against the variability present in real-world scenarios. As future work, we plan to optimize the selection of training data to better represent both normal and abnormal actions, refine the detection and classification modules or explore improved feature selection strategies to increase accuracy without significantly increasing computational cost, and include an error analysis to identify and understand common misclassification patterns, maintaining their applicability in edge devices.

REFERENCES

[1] INEGI, "Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública (ENVIPE) 2024." [Online]. Available: https://www.inegi.org.mx/programas/envipe/2024/

[2] GALLUP, "2018 GlobalLaw and Order", 2018. [Online]. Available: https://www.saferspaces.org.za/uploads/files/Gallup_Global_Law_And_Order_Report_2018.pdf#page=11

[3] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 6479–6488.

[4] W. Tan, Q. Yao, and J. Liu, "Overlooked video classification in weakly supervised video anomaly detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2024, pp. 202–210.

[5] A. Al-Lahham, N. Tastan, M. Z. Zaheer, and K. Nandakumar, "A coarse-to-fine pseudo-labeling (C2FPL) framework for unsupervised video anomaly detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 6793–6802.

[6] J. Flores-Monroy, G. Benitez-Garcia, M. Nakano-Miyatake, and H. Takahashi, "An online modular framework for anomaly detection and multiclass classification in video surveillance," *Appl. Sci.*, vol. 15, no. 17, p. 9249, 2025.

[7] A. Gao, "STEAD: Spatio-temporal efficient anomaly detection for time and compute sensitive applications," M.S. thesis, San Jose State Univ., 2025.

[8] H. Karim, K. Doshi, and Y. Yilmaz, "Real-time weakly supervised video anomaly detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 6848–6856.

[9] J. Flores-Monroy, G. Benitez-Garcia, M. Nakano, and H. Takahashi, "Optimal Feature Extractor for Video Anomaly Detection in Public Transportation Applications," in *New Trends in Intelligent Software Methodologies, Tools and Techniques*, IOS Press, 2024, pp. 249–262.

[10] S. Majhi, S. Das, F. Brémond, R. Dash, and P. K. Sa, "Weakly-supervised joint anomaly detection and classification," in *Proc. 16th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, 2021, pp. 1–7.

[11] P. Wu, X. Zhou, G. Pang, Y. Sun, J. Liu, P. Wang, and Y. Zhang, "Open-vocabulary video anomaly detection," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 18297–18307.

[12] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 203–213.

[13] NVIDIA. Jetson Orin Nano Series.[Online]. Available: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, vol. 1, 2019, pp. 4171–4186.

[15] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, *et al.*, "Rethinking attention with performers," *arXiv preprint arXiv:2009.14794*, 2020.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[17] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6299–6308.

[18] K. Li *et al.*, "UniFormer: Unified transformer for efficient spatiotemporal representation learning," *arXiv preprint arXiv:2201.04676*, 2022.