Complex Systems Informatics and Modeling Quarterly (CSIMQ)
eISSN: 2255-9922
Published online by RTU Press, https://csimq-journals.rtu.lv
Article 243, Issue 44, September/October 2025, Pages 31–51
https://doi.org/10.7250/csimq.2025-44.03

Hierarchical Fusion of 3D CNNs with Confidence Awareness for Violence Recognition in Videos

Nadjia Khatir[1,2⋆] and Hassina Meziane[2]

[1] Higher School of Electrical and Energetic Engineering (ESGEE), Oran, Algeria
[2] LITIO Laboratory, University Oran1 Ahmed Ben Bella, Oran 31000, Algeria

khatirnadjia@esgee-oran.dz, meziane.hassina@univ-oran1.dz

Abstract. The deployment of surveillance networks in smart cities plays a pivotal role in enhancing public safety through the monitoring of various environments such as roads, airports, residential areas, and establishments. Nevertheless, the vast volumes of video data generated daily by these networks present both opportunities and challenges in terms of information management and analytical processing. In this study, we propose a novel trust-aware fusion framework of video-based violence and threat modeling by combining two state-of-the-art models. I3D, which excels in overall spatio-temporal reasoning, and C3D, which learns short-term motion behaviors. In Stackelberg's game theory, the process of fusion outlines inference as a sequential decision-making process, wherein the leader is I3D, and C3D acts as a follower. A dynamic confidence threshold governs the prediction delegation power, enabling adaptive decision-making based on model confidence. Extensive experiments on a three-class dataset (*Normal*, *Violence*, *Weaponized*) prove that the introduced fusion strategy significantly outperforms single models. Setting the confidence threshold to 0.5 achieves 97.27% peak of overall accuracy. In addition, class-wise performance reveals considerable improvements, especially in the *Violence* class, where precision is 99% and the F1 score is 94%, versus 82% and 85% when using I3D individually. The experiments confirm the performance of the confidence-aware fusion for robust and context-adapted threat detection in smart-city surveillance.

Keywords: Violence Detection, Smart City Surveillance, Deep Learning, Inflated 3D ConvNet, 3D Convolutional Network, Confidence-Aware Fusion, Game Theory.

1 Introduction

Along with the expansion of smart city infrastructure, a large number of surveillance cameras have been deployed and generate volumes of video data on a daily basis. Conventionally, video data

⋆ Corresponding author

Reference: N. Khatir and H. Meziane, "Hierarchical Fusion of 3D CNNs with Confidence Awareness for Violence Recognition in Videos", Complex Systems Informatics and Modeling Quarterly, CSIMQ, no. 44, pp. 31–51, 2025. Available: https://doi.org/10.7250/csimq.2025-44.03

Additional information. Author ORCID iD: N. Khatir – https://orcid.org/0000-0003-1073-3438, H. Meziane – https://orcid.org/0000-0002-4376-0785. PII S225599222500243X. Received: 5 August 2025. Accepted: 21 October 2025. Available online: 31 October 2025.

monitoring is highly reliant upon human operators, which is time-consuming, error-prone, and non-scalable in large surveillance networks.

Intelligent Video Analytics (IVA) deployments employ machine learning to provide automated surveillance, determining suspicious behavior without human attention being devoted all the time, thereby better allocating resources [1]. In recent years, the integration of Artificial Intelligence (AI) techniques, particularly Deep Learning (DL), significantly improved the effectiveness and accuracy of anomaly detection for video surveillance scenarios [2].

Anomalies refer to rare or out-of-pattern activities such as traffic accidents, quarrels, theft, or unusual behaviors different from standard patterns. Due to their rarity and uncertainty, anomaly detection is a challenging problem, particularly in highly complex real-world scenarios where standard activities predominantly make up the data distribution [3]. Of all types of anomalies, violent actions are one of the most serious public safety issues. Rapid and real-time detection of violent actions from surveillance videos plays an important role in crime prevention and evasion, such as assault and robbery.

Automated Teller Machine (ATM) violence detection is one of the main application areas where human subjects are exposed to, and early intervention can prevent serious injury or financial loss [4].

Furthermore, violence detection now finds an appropriate role in highly sensitive environments, such as schools, where early detection of bullying or fights can save injured students and prevent further escalation [5].

Even hospitals, which are considered safe and quiet places, become hotspots of violence against doctors, medical personnel, and even patients, and automated systems to detect violence become a necessary part of the provision of safe working conditions. In addition to these environments, public environments, such as parks, public transport hubs, nightclubs, and bars, are generally areas where violent fights can occur, especially during weekends or festive periods. Human monitoring of such sites is usually inefficient due to the large number of video feeds and the limited human vigilance capacity. Automated video analytics technology based on sophisticated AI algorithms can help security staff by providing real-time alerts, which would facilitate faster responses and possibly save lives. Several AI-based models have been proposed to deal with this type of scenario. For instance, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and 3D Convolutional Networks (3D-CNNs) that can extract spatial and temporal features from video sequences [6].

Even weakly supervised and unsupervised learning methods have become popular due to their ability to be trained from unlabeled or partially labeled data, decreasing the reliance on expensive annotations [7].

In this study, the recognition of anomaly and aggressive behavior of security cameras is addressed through a deep learning framework. We seek to create an automated feature extraction framework that can efficiently recognize abnormal and aggressive behaviors from live video streams, aiding human operators during real-time surveillance and boosting security management overall. We design a hybrid model where deep spatial-temporal feature extraction is integrated with a hierarchical game-theoretic decision process.

The Stackelberg game theory (or the Stackelberg model) is a branch of non-cooperative game theory that analyzes situations in which one player (the leader) decides first, while the other players (the followers) react after observing this decision.

This framework is widely used in economics, industrial strategy, optimization, and even in the control of electrical networks or multi-agent systems. It describes hierarchical decision-making where anticipation and reaction are key elements of strategic interaction. However, this method has never been used in the field of vision or image processing.

In the Stackelberg game framework, the interaction between the leader and the follower can be expressed as a hierarchical optimization process. A Stackelberg equilibrium corresponds to a pair

of strategies (q_1^*, q_2^*) that satisfy the following conditions, where the symbol R denotes the reaction function, representing the optimal response of the follower to the leader's decision:

$$\begin{cases} q_2^* = R_2(q_1^*), & \text{the follower maximizes its payoff given } q_1^*, \\ q_1^* = \arg\max_{q_1} f_L(q_1, q_2^*), & \text{the leader maximizes its profit by anticipating this reaction.} \end{cases}$$

Inspired by this theory, we introduce an original fusion framework, leveraging two complementary deep 3D Convolutional Neural Networks, namely I3D (Inflated 3D ConvNet) and C3D (Convolutional 3D Network). Inspired by the Stackelberg game model, our method describes the inference pipeline as a sequential decision-making game in which the I3D model is described as a leader due to its strong ability to model deep spatio-temporal dynamics, and the C3D model is described as a follower, which intervenes judiciously when the leader exhibits uncertainty during prediction.

A pre-established confidence threshold adaptively governs this interaction: high-confidence leader outputs are accepted without modification, whereas low instances are forwarded to the follower, which provides a second level of analysis. This confidence-aware Stackelberg mechanism optimally balances its detection performance with its computational complexity, with this balance being sensitively tuned according to changes in scene complexity and video quality. Our proposed system aims to provide a scalable and robust solution for violence detection in real-world surveillance scenarios that can be practically deployed.

The paper is structured as follows. First, we provide a comprehensive review of related work to situate our contribution in the context of existing approaches. We then describe the proposed methodology in detail, including the theoretical background and implementation of the fusion strategy. This is followed by an experimental evaluation on a real-world dataset to assess the performance of our method relative to baseline models. Finally, we discuss the implications of our findings and suggest directions for future research.

2 Related Work

Smart city video surveillance networks increasingly utilize automated classification to identify violence in real time, with the potential to improve public safety and reduce human operator workload. Recent advances in AI and machine learning have led to faster and more accurate violence detection even in challenging public areas. This section presents a comprehensive review of deep learning and computer vision techniques deployed on video for violence detection, which seek to solve fundamental issues of accuracy, effectiveness, and real-time application in challenging public settings.

2.1 Feature Extraction Techniques and Machine Learning

Previous works employ attributes such as the Histogram of Oriented Gradients (HOG) and Optical Flow, and machine learning-based classifiers such as Support Vector Machines (SVM) are commonly used to identify violence from video frame features [8].

HOG features are also used to extract low-level features from video frames, which are further classified through several machine learning algorithms, such as SVM. This method has revealed notable increases in the accuracy to detect violence in surveillance applications, reaching 86% accuracy when Random Forest classifiers were used [9].

SVM, a popular supervised classification model, was successfully implemented in tandem with dynamic image methods and the Bag of Visual Words framework to identify violence and achieve high accuracy from varied data sets [10]. Optical flow features were utilized to obtain motion information, which is a pivotal element in identifying violent behavior. Visual features were further refined through techniques that utilize the Motion-Guided Attention Module (MGAM) based on optical flow, increasing the accuracy of the recognition of violent action of video sequences [11].

2.2 Deep Learning Approaches

Current developments have progressed to deep learning models that exhibit higher accuracy than conventional approaches. Deep learning technologies have enabled real-time video surveillance to develop intelligent video surveillance systems that independently classify human behavior [12]. Convolutional Neural Networks (CNNs) automatically extract spatial features from video frames and significantly improve detection performance. CNNs are also used to extract spatial features from individual frames, which are then sent through additional layers to be classified. Pre-trained networks like VGG-19, Inception-v3, ResNet-50, DenseNet121, and MobileNetV2 are popular feature extractors [13]. For instance, a pre-trained model of ResNet-50 extracts features, which can be sent through a ConvLSTM block [14].

3D CNNs and Long Short-Term Memory (LSTM) Models such as VioNet combine 3D CNNs with Bidirectional LSTM to capture spatio-temporal features, achieving high accuracy rates (up to 97.85%) on various datasets [6].

Aremu et al. [15] introduced SSIVD-Net, a novel framework for the detection of weaponized violence in surveillance footage. Their work proposed the Smart-City CCTV Violence Detection (SCVD) dataset, specifically designed to capture both weaponized and non-weaponized violent behaviors. To address the computational and spatial–temporal challenges of 3D CNNs, they developed the Salient Super Image (SSI) technique, which converts video sequences into salient 2D composite images, thus reducing data dimensionality and improving inference efficiency. Furthermore, they designed the *Salient-Classifier*, a kernelized residual architecture that combines polynomial and Gaussian kernels for enhanced discrimination between classes of close violence. Experimental results demonstrated that SSIVD-Net significantly outperformed existing 3D CNNs and ConvLSTM-based models, achieving state-of-the-art accuracy on the SCVD and benchmark datasets.

2.3 Hybrid Models

The issue of recognizing violence in video recordings through an aggregation of disparate kinds of features has remained paramount in recent research, particularly in application scenarios involving youth protection agencies and police departments [16]. The latest research focuses more on the integration of disparate types of data inputs, specifically visual information, audio signals [17], and behavioral movements, in an effort to achieve more robust and accurate detection systems. One common strategy is to use deep learning architectures to individually extract each modality and then fuse the said representations to produce a classification [18]. The integration of visual and auditory information is done with greater precision, considering the complementary nature of these cues. Recent and more sophisticated methods of neural network architecture with attention modules [17], bilinear pooling, and mutual learning to combine these features are more efficient compared to previous methods.

Multilevel Feature Fusion focuses on the extraction of subtle details and complex features of violent actions [19]. Systems that extract appearance and motion features using CNNs and LSTM networks on sequences of frames exhibit robust performance.

Recent research combines lightweight CNNs such as MobileNetV2, EfficientNet-B0, and ResNet50V2, with layers of LSTM (Long Short-Term Memory) or GRU (Gated Recurrent Unit) to achieve efficient real-time performance, even on embedded devices [20].

Graph neural networks and action knowledge graphs model relationships between features and temporal context, enabling real-time and robust violence detection [21].

3 Proposed Methodology

For video data classification, we use the two well-known 3D Convolutional Neural Networks (3D CNNs): I3D (Inflated 3D ConvNet) [22] and C3D [23]. Both models have spatial-temporal

processing capabilities, although they differ substantially in design and operational characteristics. To improve the system's robustness, we introduce a hierarchical fusion approach based on predictive confidence. With this model, fusion defines the I3D model as the primary predictor and C3D as the adaptive secondary model for uncertain cases, making the system resilient for the detection of a real-life video violence.

3.1 Overview of I3D and C3D Architectures

A key distinction between image-based and video-based data lies in the temporal dimension present in videos. To effectively capture this temporal information, deep learning architectures have evolved to incorporate 3D processing techniques. This study highlights the architectural shift from static image analysis to spatio-temporal video analysis, with a particular focus on the I3D and C3D models as representative examples.

I3D model: proposed by DeepMind and researchers at Oxford University in the article *"Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset"* [22], puts forward a different way of action recognition from videos. The authors discuss early methods and describe an architecture that expands ordinary 2D convolutional networks to the time domain. In particular, their strategy inflates all 2D convolutional and pooling kernels by adding a third dimension, time-changing square filters of size $N \times N$ to cubic filters of size $N \times N \times N$. In this way, the mentioned inflation creates spatio-temporal patterns fundamental to video comprehension that the model can capture. The Key Features of this model are: (1) based on an Inflated Inception-v1 backbone, (2) pre-trained on large-scale video data such as Kinetics-400[†], (3) handles short-term as well as long-term video sequence dependencies, and (4) high computational complexity yet better generalization.

C3D model: The C3D model was developed in video action recognition and was described in the article 'Learning Spatiotemporal Features with 3D Convolutional Networks' by Tran et al. [23]. Unlike 2D CNNs, which only analyze still images or short clips, C3D processes sections of video with 3D convolution. It captures visual information and motion dynamics over time in a set of adjacent frames since its architecture extends 2D convolutions to 3D. C3D obtains meaningful spatiotemporal features in fixed-length clips and learns from the raw RGB data, which means it does not rely on optical flow or hand-crafted features relying on prior work. The relative lightness of its architecture as compared to deeper models is offset by the fact that its span of captured context is limited to the short term.

3.2 Our Confidence-Guided Model Fusion Approach

Although both C3D and I3D showed excellent video comprehension, each model has its intrinsic shortcomings. I3D tends to perform poorly in noisy inputs, whereas C3D usually lags behind in modeling long-range temporal relationships. Since no model is optimal for all cases, we introduce a confidence-based fusion methodology that adaptively exploits both models according to their respective strengths. The system responds variably to different input circumstances and model performance: when I3D makes a high-confidence decision, its output dominates to take advantage of its powerful spatio-temporal reasoning. However, if its confidence is below a pre-specified value, the fusion process transitions towards C3D, which, although incapable of modeling long-term relationships, provides stable short-term motion comprehension. The adaptive system responds flexibly to different input circumstances and model performance. Our fusion mechanism is formally grounded in the Stackelberg game model, originally proposed by Heinrich von Stackelberg in 1934 [24]. As already was mentioned, this model describes situations with a leader and a follower

[†] Kinetics400 Dataset: The Kinetics Human Action Video Dataset
https://academictorrents.com/details/184d11318372f70018cf9a72ef867e2fb9ce1d26

acting in a hierarchy, in which the leader makes the first move and the follower optimally reacts after seeing the leader move.

We model the inference process as a sequential decision game (see the pipeline in Figure 1):

Collaboration is regulated by a confidence threshold θ: the I3D model's confidence score surpassing θ means its prediction is considered the most accurate. Otherwise, the C3D model is granted the opportunity to make a compensatory decision based on its distinct feature representation in relation to the leader's uncertainty.

Our inference procedure in this example is structured following a hierarchical method. This structuring minimizes dependence on any single model and smooths out issues that arise when a model performs sub-optimally. Additionally, this form of combining models is amenable to practical real-world operational limitations, allowing a balancing of computational cost, time to inference, and classification robustness.

In general, this proposed fusion method improves reliability and provides an extendable framework that is suitable for high-complexity, high-stakes applications, such as recognizing violent activity from surveillance videos.

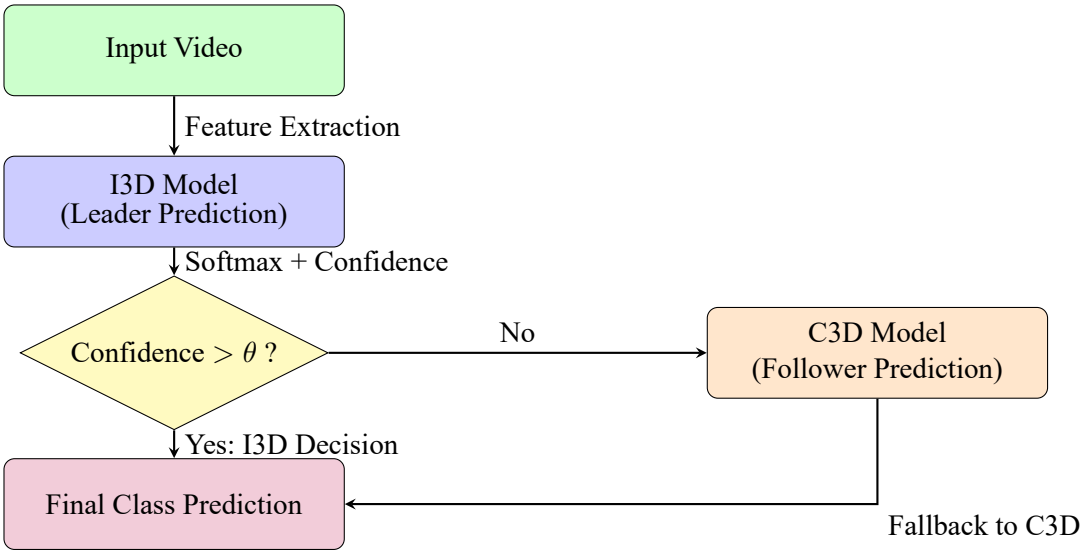


Figure 1. Pipeline of Fusion Strategy

3.3 Mathematical Formulation of the Theoretical Contribution

Let $\mathcal{X}$ denote the input video space and $\mathcal{Y} = \{1, \ldots, K\}$ the set of class labels. Given a video sample $x \in \mathcal{X}$, two deep models are used:

$$f_L(x) : \mathcal{X} \to \mathbb{R}^K \quad \text{(I3D: Leader model)}, \tag{1}$$

$$f_F(x) : \mathcal{X} \to \mathbb{R}^K \quad \text{(C3D: Follower model)}. \tag{2}$$

Each model produces a vector of logits that represents the unnormalized evidence for each class. The corresponding probability distributions are obtained by applying the softmax operator:

$$p_L(x) = \sigma(f_L(x)), \qquad p_F(x) = \sigma(f_F(x)), \tag{3}$$

where $\sigma(\cdot)$ denotes the softmax function.
The leader model's confidence is defined as the maximum softmax probability:

$$c_L(x) = \max_{y \in \mathcal{Y}} p_L^y(x), \tag{4}$$

and its predicted label is given by:

$$\hat{y}_L(x) = \arg\max_{y \in \mathcal{Y}} p_L^y(x). \tag{5}$$

Analogously, the follower prediction is the following:

$$\hat{y}_F(x) = \arg\max_{y \in \mathcal{Y}} p_F^y(x). \tag{6}$$

The proposed fusion framework models inference as a **sequential decision process**:

$$\hat{y}(x) = \begin{cases} \hat{y}_L(x), & \text{if } c_L(x) \geq \theta, \\ \hat{y}_F(x), & \text{otherwise,} \end{cases} \tag{7}$$

where $\theta \in [0, 1]$ denotes a pre-specified *confidence threshold*. This threshold regulates the hierarchical collaboration between both models.

From a game-theoretical point of view, the decision-making can be formulated as follow:

$$\mathcal{G} = \langle \{L, F\}, \mathcal{S}_L, \mathcal{S}_F, U_L, U_F \rangle,$$

where:

- L (the leader) corresponds to the I3D model, responsible for global spatio-temporal reasoning,
- F (the follower) corresponds to the C3D model, which reacts to the leader's uncertainty,
- $\mathcal{S}_L$ and $\mathcal{S}_F$ are their respective strategy spaces (predict or defer),
- U_L and U_F denote their utility functions, related to confidence-based accuracy.

The leader acts first, generating an action $a_L = (\hat{y}_L, c_L)$. Upon observing a_L, the follower reacts according to:

$$a_F^*(a_L) = \begin{cases} \text{No Action}, & \text{if } c_L \geq \theta, \\ \text{Predict } \hat{y}_F, & \text{if } c_L < \theta. \end{cases} \tag{8}$$

The final decision of the system follows the equilibrium of this hierarchical interaction, as formalized in Equation 7. Operationally, the fusion algorithm implemented in our system can be described as:

$$\forall x \in \mathcal{X} : \begin{cases} \text{Compute } p_L(x) = \sigma(f_L(x)) \\ \text{If } \max(p_L(x)) \geq \theta, \text{ return } \hat{y}_L(x) \\ \text{Else, compute } p_F(x) = \sigma(f_F(x)) \text{ and return } \hat{y}_F(x) \end{cases} \tag{9}$$

This adaptive hierarchical fusion effectively leverages the complementary capabilities of both networks: the I3D (leader) dominates when confident, exploiting long-term temporal dependencies, while the C3D (follower) compensates in uncertain cases by emphasizing short-term motion dynamics.

Mathematically, the inference objective can be expressed as minimizing the expected classification risk under a hierarchical policy:

$$\hat{y}(x) = \arg\min_{y \in \mathcal{Y}} \mathbb{E}_{M \in \{L,F\}} \big[\mathcal{L}(y, M(x)) \big], \tag{10}$$

where model selection $M(x)$ follows the Stackelberg rule in Equation 7.

Therefore, the proposed fusion provides a rational and interpretable framework for combining heterogeneous video classifiers, uniting robustness, and spatio-temporal reasoning in a principled manner.

3.4 Algorithm and Illustrative Example

The fusion procedure is formally described in Algorithm 1. It ensures that predictions are primarily driven by the stronger model (I3D) while adaptively leveraging the secondary model (C3D) in ambiguous cases, yielding a robust framework for real-world violence detection in videos.

To better illustrate the decision-making process, we present a concrete numerical example using softmax output probabilities from both I3D and C3D models.

Scenario Description: Consider a video clip $\mathcal{V}$ processed by both I3D and C3D models. The softmax output probabilities from the I3D model are as follows:

$$\mathbf{p}_{\text{I3D}} = [0.10, 0.85, 0.05], \tag{11}$$

where each element corresponds to the predicted probability for the classes *Normal*, *Violence*, and *Weaponized*, respectively.

Step 1: Compute I3D Confidence. The confidence score of I3D is the maximum softmax probability:

$$C_{\text{I3D}} = \max(\mathbf{p}_{\text{I3D}}) = 0.85, \tag{12}$$

which corresponds to the predicted class $\hat{y}_{\text{I3D}} = \textit{Violence}$.

Step 2: Apply Decision Rule Based on Threshold θ.

- **Case 1:** If the confidence threshold is set to $\theta = 0.7$, since $C_{\text{I3D}} = 0.85 \geq 0.7$, the system directly accepts I3D's decision:

$$\boxed{\text{Prediction} = \textit{Violence}\ (\text{from I3D})}.$$

- **Case 2:** If the confidence threshold is set to $\theta = 0.9$, we have $C_{\text{I3D}} = 0.85 < 0.9$, meaning I3D's confidence is insufficient. The system defers the decision to the C3D model.
  Suppose the C3D model produces the following logits:

$$\mathbf{l}_{\text{C3D}} = [1.2, 0.5, 2.3], \tag{13}$$

  which, after applying the softmax function, yields:

$$\mathbf{p}_{\text{C3D}} = [0.19, 0.11, 0.70], \tag{14}$$

  resulting in the final class prediction $\hat{y}_{\text{C3D}} = \textit{Weaponized}$.
  Therefore, the final prediction is:

$$\boxed{\text{Prediction} = \textit{Weaponized}\ (\text{from C3D})}.$$

Summary: This example highlights the adaptive behavior of the Stackelberg fusion strategy:

- When I3D is sufficiently confident ($C_{\text{I3D}} \geq \theta$), its prediction is accepted directly.
- When I3D is uncertain ($C_{\text{I3D}} < \theta$), the system relies on C3D, potentially correcting misclassifications in low-confidence situations.

Algorithm 1: Confidence-Aware Stackelberg Fusion Strategy

Input: Video $\mathcal{V}$, I3D Model M_{I3D}, C3D Model M_{C3D}, Confidence Threshold θ
Output: Predicted Class y

Extract frames $\mathbf{X}$ from $\mathcal{V}$;
Compute logits $\mathbf{l}_{\text{I3D}} = M_{\text{I3D}}(\mathbf{X})$;
Calculate softmax probabilities $\mathbf{p}_{\text{I3D}} = \text{Softmax}(\mathbf{l}_{\text{I3D}})$;
Identify predicted class $y_{\text{I3D}} = \arg\max(\mathbf{p}_{\text{I3D}})$ and confidence $C_{\text{I3D}} = \max(\mathbf{p}_{\text{I3D}})$;
if $C_{I3D} \geq \theta$ **then**
    Assign $y = y_{\text{I3D}}$;
else
    Compute logits $\mathbf{l}_{\text{C3D}} = M_{\text{C3D}}(\mathbf{X})$;
    Predict $y = \arg\max(\mathbf{l}_{\text{C3D}})$;
return y;

3.5 Computational Complexity Analysis

In order to compare the computational efficiency of the method, we analyze its algorithmic complexity in terms of time and space complexity.

Let M_{I3D} and M_{C3D} denote the two models involved in the hierarchical fusion process, where M_{I3D} acts as the leader and M_{C3D} as the follower. The proposed inference mechanism evaluates an input video $\mathcal{V}$ consisting of T frames, each of spatial dimension $H \times W$ and with C channels.

1) Time Complexity. The overall time complexity of the fusion can be expressed as:

$$\mathcal{O}(f_{\text{I3D}}(T, H, W)) + \mathbb{I}_{(C_{\text{I3D}} < \theta)} \cdot \mathcal{O}(f_{\text{C3D}}(T, H, W)),$$

where f_{I3D} and f_{C3D} denote the computational costs of the I3D and C3D forward passes, respectively, and $\mathbb{I}_{(C_{\text{I3D}} < \theta)}$ is an indicator function that activates the follower model only when the leader's confidence C_{I3D} falls below the confidence threshold θ.

Since the follower is invoked only under uncertainty, the expected complexity is significantly reduced compared to a naïve ensemble strategy where both models are executed for every input. In practice, this leads to an average inference cost close to that of a single model, i.e.,

$$\mathbb{E}[\text{Time}] \approx \mathcal{O}(f_{\text{I3D}}) + p_u \cdot \mathcal{O}(f_{\text{C3D}}),$$

where p_u is the empirical probability that the leader's confidence is below θ.

2) Space Complexity. Both models are loaded in memory simultaneously during inference, resulting in a space complexity of

$$\mathcal{O}(\text{Params}_{\text{I3D}} + \text{Params}_{\text{C3D}}) + \mathcal{O}(T \times H \times W \times C),$$

which accounts for model parameters and input tensors. However, only one model at a time performs backpropagation during training, maintaining feasible GPU memory usage.

3) Discussion. The hierarchical structure of the fusion minimizes redundant computations while maintaining robustness. Compared to parallel ensemble methods, it achieves a favorable trade-off between accuracy and inference time, as only one model is executed in a majority of cases. Thus, the proposed method exhibits sub-linear computational growth with respect to the number of models and maintains practical deployment feasibility on standard hardware.

4 Experiments

4.1 Dataset Description

In this study, we used the Smart City CCTV (Closed Circuit Television) Violence Detection (SCVD) dataset [15][‡], a recent benchmark focused solely on CCTV-based surveillance. Unlike prior datasets that involve handheld or phone footage, SCVD aligns with real-world urban security conditions, reducing distributional bias. A key feature of SCVD is the inclusion of a weaponized activity class, which offers annotated video sequences of harmful handheld objects beyond typical firearms or knives. To our knowledge, SCVD is the first public video dataset that combines violence and weapon detection in CCTV contexts, making it a relevant benchmark for smart city threat analysis.

4.2 Development Environment

All experiments were conducted on a workstation with Ubuntu 22.04.5 LTS (64-bit) and Linux kernel 6.8.0-85-generic. The hardware configuration included an Intel Core i7-7700K CPU (8 cores, 4.5 GHz), 32 GB RAM, and a dual-GPU setup comprising an integrated Intel HD Graphics 630 and a discrete ASUS DUAL GeForce RTX 3060 12G GDDR6.

For comparison purposes, some initial experiments were also performed on Google Colab with a Tesla T4 GPU on the free-tier configuration. Yet this arrangement was considerably slower in latency and more time-consuming in training, mainly due to resource sharing and session time limitations of the free version. Additionally, from both a practical and ethical point of view, it is more appropriate to process video streams directly on a secured local server for surveillance, where data sensitivity and privacy are of paramount importance.

The proposed deep learning framework was implemented in Python 3.10 with the support of core libraries such as PyTorch, TorchVision, and PyTorchVideo for model construction and training, NumPy and Pandas for data manipulation, and Matplotlib and Seaborn for statistical visualization. The OpenCV library was used for frame extraction, temporal sampling, and image preprocessing from raw video feeds. This computing environment ensured both efficiency and reproducibility for all experiments.

4.3 Training and Fine-Tuning Models

Although both I3D and C3D are pre-trained on large-scale generic video datasets, domain shift is inevitable when applying these models to specialized tasks such as violence detection. Pre-trained features often capture general motion patterns and visual semantics that may not directly translate to the unique characteristics of violent or weaponized scenes.

Fine-tuning is therefore essential to bridge this domain gap. By updating all model parameters during training on the target dataset, the models can adapt their internal representations to capture task-specific discriminative cues such as aggressive movements, human interactions, and contextual objects indicative of violence. This process allows the models to retain valuable low-level spatio-temporal priors from pretraining while refining their high-level feature extraction towards the target classes.

Figure 2 illustrates the progression of training loss for the I3D and C3D models over 10 epochs. In general, both models exhibit a sharp decrease in loss within the first few epochs, indicating rapid learning during the initial training phase. Specifically, the I3D model demonstrates a steeper decline in loss, dropping from 0.5379 in epoch 1 to 0.0437 by epoch 5, followed by smaller fluctuations in later epochs. In particular, a minor increase is observed at epoch 6 (0.0590), which might be attributed to overfitting or local minima, before decreasing again.

[‡] https://github.com/tolusophy/Violence_Detection

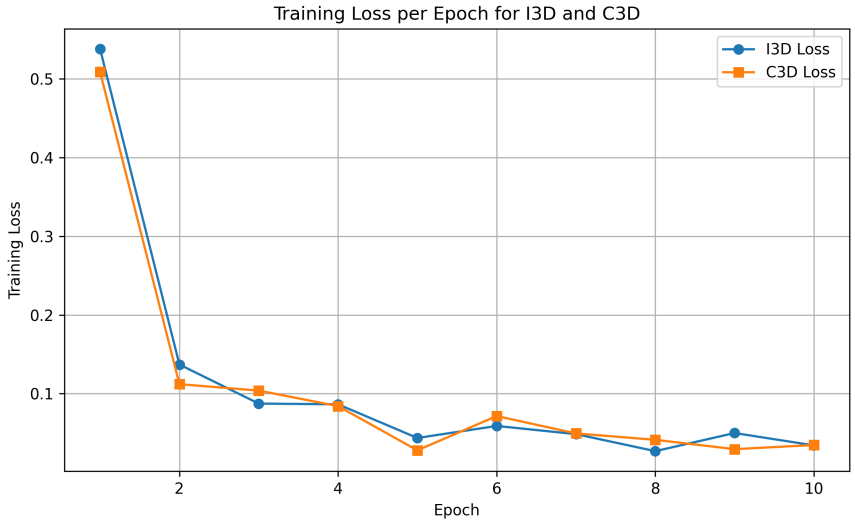


Figure 2. Training loss per epoch for I3D and C3D models over 10 epochs

Similarly, the C3D model also shows a significant drop in loss from 0.5087 to 0.0280 within the first five epochs. However, its loss fluctuates slightly in later epochs, with minor increases at epochs 6 (0.0716) and 7 (0.0495), before stabilizing.

The two models have similar levels of loss convergence; however after a certain number of epochs we observed that the I3D model had a generally smoother and more stable decline. This suggests that the potentially deeper spatiotemporal feature extraction ability of the I3D model will allow it to converge more stably on this dataset. The moderate amounts of loss oscillation in both models after epoch 5, suggests that these are indications of an eventual learning saturation point in the dataset, whereby either more regularization or adjusting the learning rate would have been required to make any further learning progress.

The fine-tuning process follows these key principles:

- The **final classification layer** of each model is modified by replacing the last fully connected layer with a new `linear` layer matching the number of target classes (three in our case).
- We apply **full fine-tuning**, meaning that all layers of the network are updated during training to allow optimal adaptation to the new domain.
- An **Adam optimizer** with a low learning rate (1×10^{-4}) is used to ensure smooth and stable convergence.
- We employ a **Cross-Entropy Loss** function, and training is performed for a fixed number of **epochs** (5 epochs), with temporal subsampling of the videos (extracting 8 frames per clip).

This fine-tuning strategy enables each model to capture task-specific discriminative features while retaining general spatiotemporal knowledge from pretraining, resulting in enhanced classification performance on the violence detection task.

4.4 Evaluation Metrics

The performance of each model and fusion configuration was assessed using the following metrics [25]:

- **Accuracy**: Overall proportion of correctly classified images.

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{y_i = \hat{y}_i\}$$

where N is the total number of images, y_i is the true label of image i, $\hat{y}_i$ is the predicted label, and $\mathbf{1}\{\cdot\}$ is the indicator function (equal to 1 if the condition is true and 0 otherwise).

- **Precision per class**:

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k}$$

where TP_k is the number of true positives for class k, and FP_k is the number of false positives for class k.

- **Recall per class**:

$$\text{Recall}_k = \frac{TP_k}{TP_k + FN_k}$$

- **F1-score per class**:

$$F1_k = 2 \times \frac{\text{Precision}_k \times \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}$$

- **Macro-averaged metrics**:

$$\text{Precision}_{macro} = \frac{1}{K} \sum_{k=1}^{K} \text{Precision}_k$$

$$\text{Recall}_{macro} = \frac{1}{K} \sum_{k=1}^{K} \text{Recall}_k$$

$$F1_{macro} = \frac{1}{K} \sum_{k=1}^{K} F1_k$$

where $K = 4$ is the total number of classes.

- **Confusion Matrix**: Provides a visual overview of classification performance across all classes, highlighting potential misclassifications.

5 Results and Discussion

Before using the confidence aware fusion method proposed above, we first evaluated the performance of the individual C3D and I3D pretrained networks on the new dataset. The main quantitative results are listed in Table 1.

Table 1. Performance metrics (%) of individual models on the test set before fusion

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
I3D	91.61	91.33	91.00	91.00
C3D	85.53	85.00	86.00	85.00

Table 2. Per class Precision, Recall, and F1-score (%) of I3D and C3D

Class	Support	I3D			C3D		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Normal	169	99.0	85.0	91.0	84.0	80.0	82.0
Violence	118	82.0	89.0	85.0	76.0	89.0	82.0
Weaponized	190	93.0	99.0	96.0	95.0	88.0	91.0

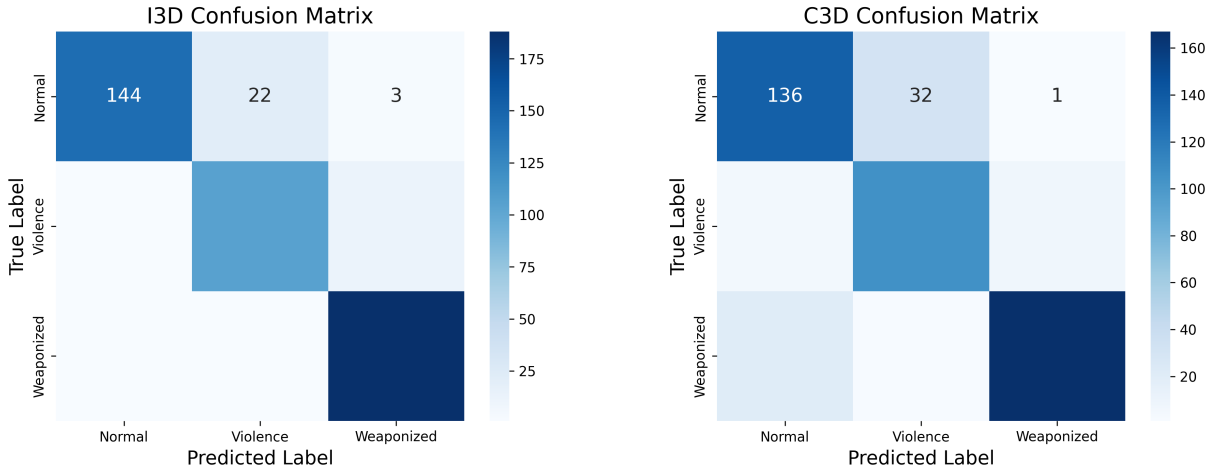
The I3D model achieved a global accuracy of 91.61%, demonstrating strong performance, particularly in recognizing the *Normal* and *Weaponized* classes with high precision (99% and 93%, respectively) and high recall (85% and 99%, respectively). However, for the *Violence* class, while the recall remained high, 89%, the precision was comparatively lower, 82%, suggesting a tendency

to produce more false positives for this category. The corresponding F1 scores indicate balanced performance in general, with a macro-average F1 score of 91%.

In contrast, the C3D model exhibited a lower overall accuracy of 85.53%. It maintained reasonable performance on the *Weaponized* class with a precision of 95% but showed a noticeable drop in precision for the *Normal* 84% and *Violence* 76% classes. Interestingly, the C3D model demonstrated a similar recall of 89% to I3D for the *Violence* class, but at the cost of precision, leading to a less balanced F1 score profile. The macro average F1 score reached 85%, confirming that while C3D remains effective in certain contexts, it is overall less robust than I3D in capturing the full spectrum of spatio-temporal features present in the dataset.

Figures 3a and 3b present the confusion matrices of the I3D and C3D models, respectively, illustrating their prediction behavior per class on the SCVD dataset. Both architectures show solid overall performance but differ in their handling of temporal information and inter-class ambiguity.

The I3D model, shown in Figure 3a, correctly classifies 144 Normal, 22 Violence, and 3 Weaponized samples. It demonstrates a strong ability to recognize weaponized activities, with very few misclassifications between the Violence and Weaponized classes. However, a small number of violent clips are incorrectly categorized as normal, indicating a limitation in capturing fine short-term motion details.



(a) Confusion matrix of the I3D model, showing per-class prediction accuracy across *Normal*, *Violence*, and *Weaponized* classes.

(b) Confusion matrix of the C3D model, emphasizing class-level misclassifications and temporal bias.

Figure 3. Side-by-side comparison of I3D and C3D confusion matrices illustrating complementary classification patterns.

The C3D model, illustrated in Figure 3b, records 136 correctly predicted Normal samples, 32 correctly identified Violence clips, and a single misclassified Weaponized instance. Compared to I3D, C3D is more sensitive to local motion variations, which results in better recognition of violent actions. However, it sometimes confuses violent and nonviolent sequences in situations involving occlusion or subtle motion, leading to slightly higher error rates in the normal category.

In general, the confusion matrices highlight the complementary nature of the two models. I3D performs better at capturing long-term temporal dependencies, while C3D focuses more effectively on short-term motion cues. These observations support the need for a hierarchical confidence-based fusion approach that unifies both perspectives to achieve balanced and robust performance across all classes.

5.1 Qualitative Evaluation of I3D and C3D Models Prior to Fusion

To complement quantitative performance analysis, we conducted a qualitative evaluation of both the I3D and C3D models before applying the fusion strategy. The objective of this qualitative

assessment is to visually examine the decision-making behavior of each model when faced with correctly and incorrectly classified video samples from the test set.

For this purpose, we randomly selected two test videos that were correctly classified and two test videos that were incorrectly classified by each model (I3D and C3D), without any retraining. The video samples were then visualized through representative keyframes to better understand the strengths and limitations of each model.

Keyframes were extracted from each video sequence by uniformly sampling a fixed number of frames across the temporal axis. Specifically, for each video clip consisting of T frames, we selected $k = 4$ *keyframes* at regular intervals to provide a compact yet informative summary of the temporal evolution of the video content. This approach allows for an efficient qualitative inspection of both spatial and temporal patterns without requiring full video playback.

The qualitative results are presented in Figures 4 and 5, where each figure displays the selected keyframes for correctly and incorrectly classified examples. The title of the image indicates the predicted and true class labels, facilitating an intuitive understanding of model behavior.

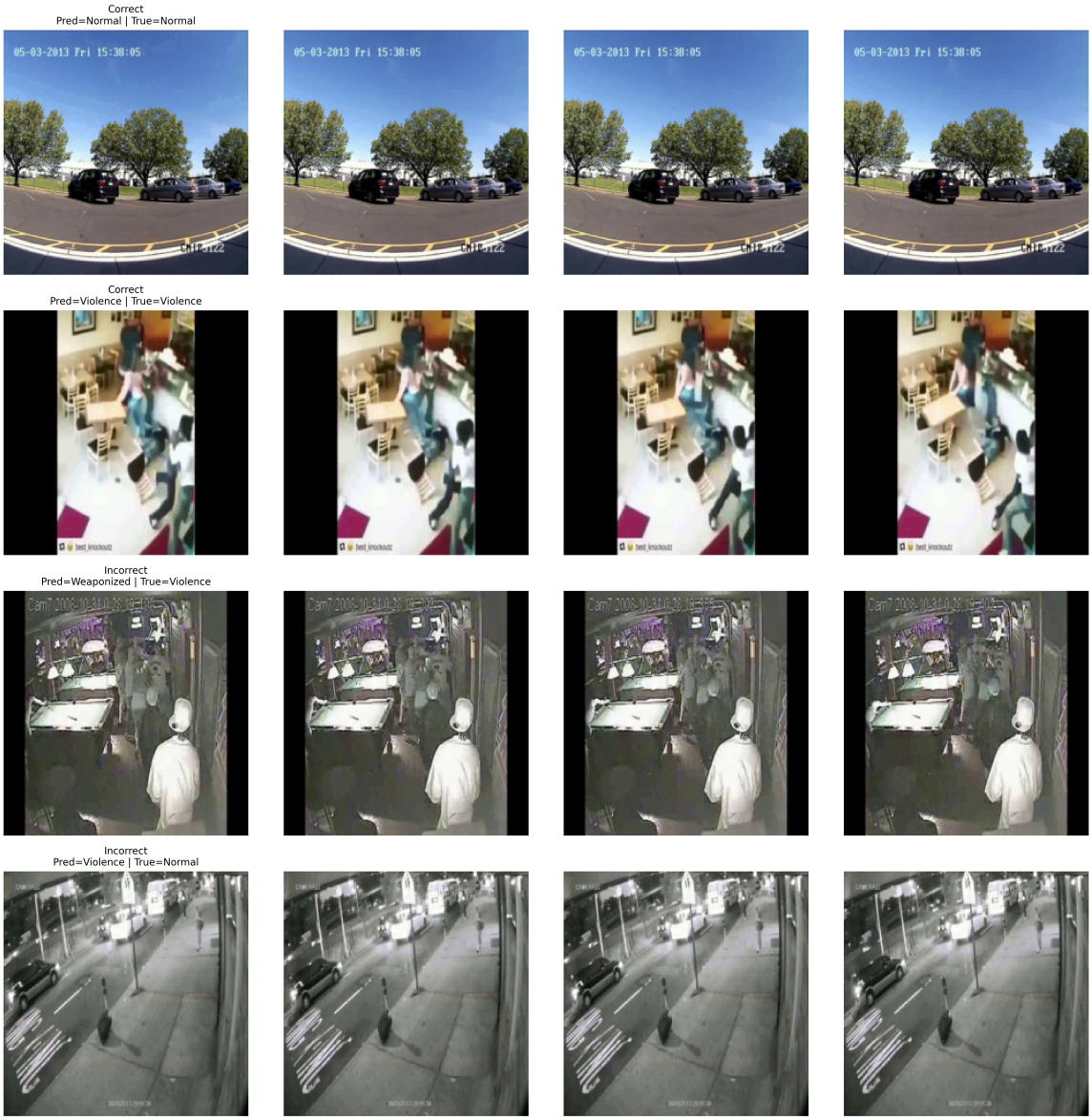


Figure 4. Representative keyframes of correctly and incorrectly classified examples by the I3D model. Each row corresponds to a video sequence, with frames sampled uniformly from the clip. The predicted and true class labels are indicated above each set of frames.

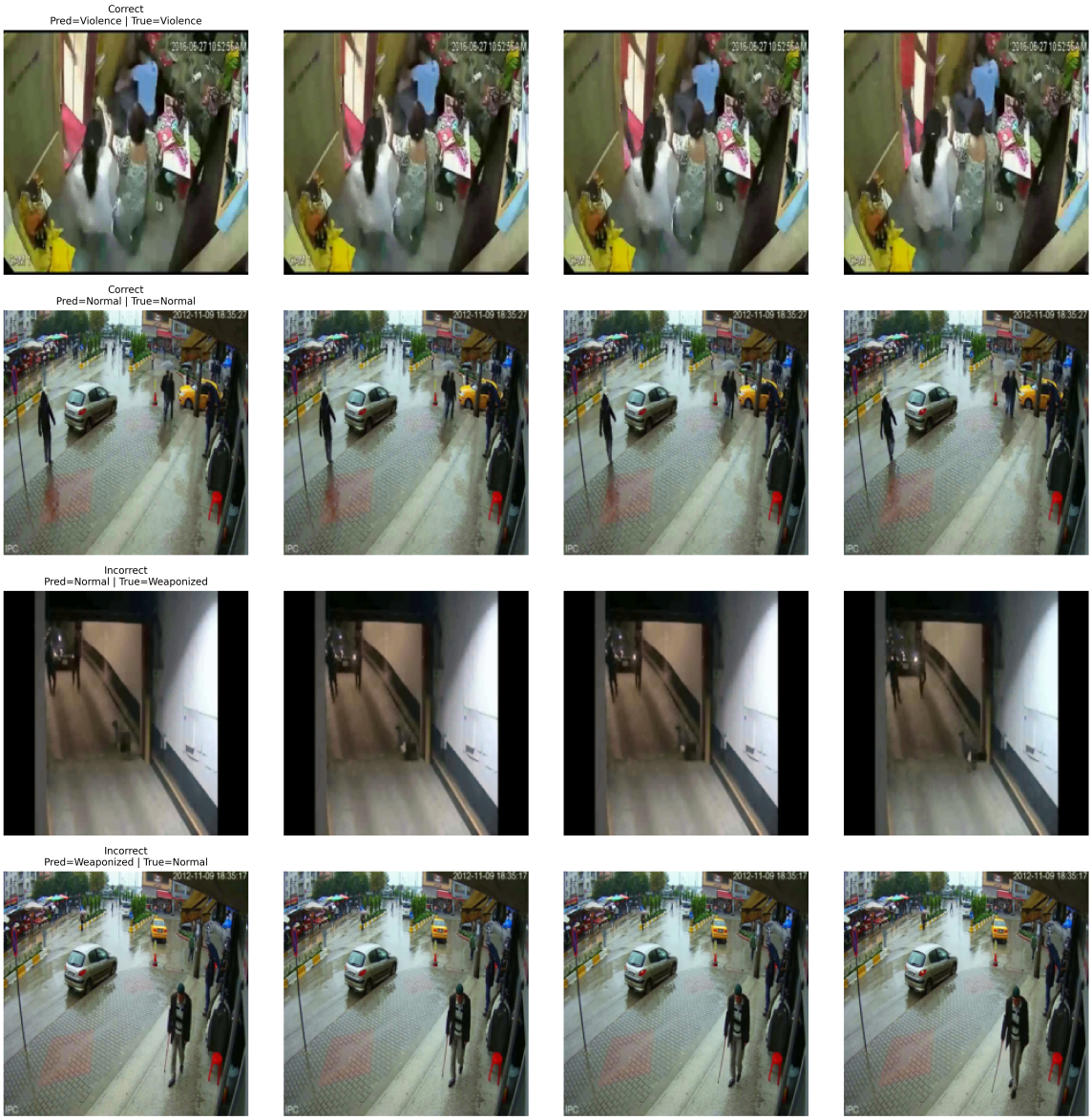


Figure 5. Representative keyframes of correctly and incorrectly classified examples by the C3D model. Each row corresponds to a video sequence, with frames sampled uniformly from the clip. The predicted and true class labels are indicated above each set of frames.

From these visualizations, we observed that the I3D model tends to capture more discriminative spatio-temporal patterns, particularly for complex actions, whereas the C3D model occasionally misinterprets scenes with ambiguous or subtle motion cues. This qualitative inspection corroborates the quantitative findings and further motivates the integration of a fusion strategy to leverage the complementary strengths of both models.

The images above illustrate a misclassification made by the models. For instance, in the case of the C3D (Convolutional 3D) model, the video sequence shows a normal scene featuring a senior individual walking on a rainy street using a cane for support. Despite the absence of aggressive or suspicious behavior, the model incorrectly predicted the scene as *Weaponized*. This false positive indicates that the model might have misinterpreted the walking cane as a potential weapon, probably due to limitations in its training data or the sensitivity to feature extraction. Such errors emphasize the necessity of improving context awareness in action recognition models, especially in safety-critical applications.

5.2 Quantitative Fusion Performance Results

Following the individual evaluation of the I3D and C3D models, we implemented a trust-based fusion strategy to combine their predictive outputs. The primary goal of this fusion approach is to capitalize on the complementary strengths of both models and mitigate their individual weaknesses. Specifically, we employed a confidence threshold mechanism, where the decision of the I3D model is prioritized when its confidence exceeds a predefined threshold, otherwise the C3D model's prediction is utilized.

The fusion method was evaluated with various confidence threshold values ranging from 0.4 to 0.8 to assess its impact on classification performance. The evaluation metrics included global accuracy, class-wise precision, recall, and F1 score. The results demonstrate a notable improvement in overall performance when combining the two models, with optimal results obtained at a threshold of 0.5.

In the following subsection, we provide a detailed quantitative comparison of global accuracy between individual models and the fusion approach, highlighting the performance gains achieved through the proposed strategy.

5.2.1 Global Accuracy Comparison

Table 3 summarizes the overall classification accuracy achieved by the individual models (I3D, C3D) and the proposed Stackelberg fusion strategy across different confidence thresholds.

Table 3. Global Accuracy (%) of Individual Models and Fusion Strategy

Model	**Accuracy (%)**
I3D	91.61
C3D	85.53
Fusion @ 0.4	92.24
Fusion @ 0.5	**97.27**
Fusion @ 0.6	97.06
Fusion @ 0.7	96.65
Fusion @ 0.8	96.65

As shown in Table 3, the fusion method outperforms both individual models, with the best accuracy of 97.27% obtained at a confidence threshold $\theta = 0.5$. This result demonstrates the effectiveness of Confidence-aware fusion in improving classification performance by leveraging the complementary strengths of I3D and C3D.

5.2.2 Accuracy Evolution with Confidence Threshold.

Figure 6 visualizes the evolution of global accuracy with a varying threshold θ.

The best performance is observed at $\theta = 0.5$. A threshold too low (0.4) or too high (0.7, 0.8) slightly degrades the accuracy, highlighting the importance of tuning θ for optimal performance.

5.2.3 Class-Wise Precision, Recall, and F1-Score.

Tables 4, 5, and 6 report precision, recall, and F1 score per-class, respectively, for the three classes: *Normal*, *Violence*, and *Weaponized*.

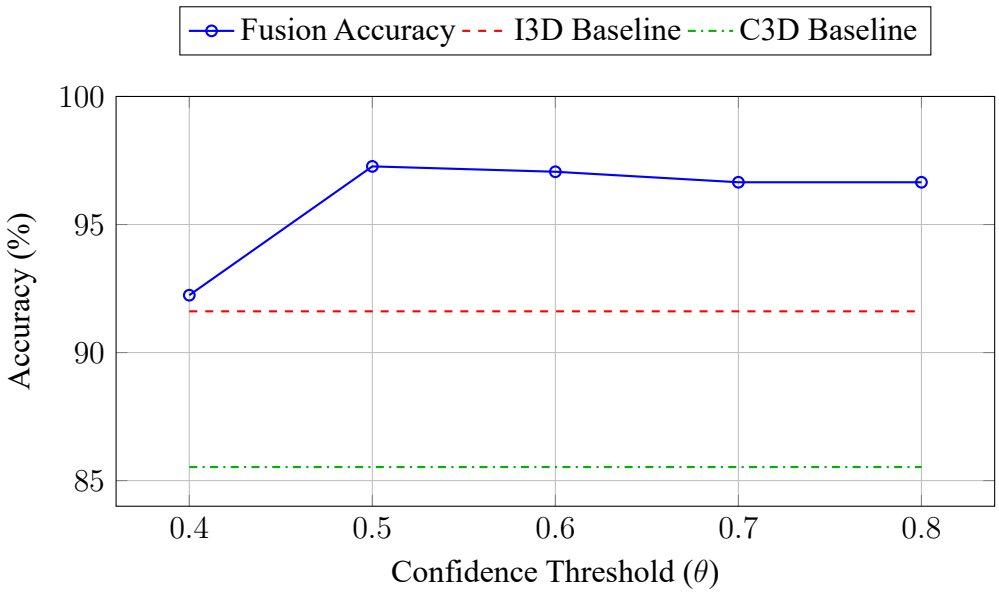


Figure 6. Global Accuracy Evolution across Confidence Thresholds

Table 4. Per class precision (%)

Model	Normal	Violence	Weaponized
I3D	99.0	82.0	93.0
C3D	84.0	76.0	95.0
Fusion@0.4	99.0	83.0	93.0
Fusion@0.5	**100.0**	**99.0**	94.0
Fusion@0.6	100.0	99.0	94.0
Fusion@0.7	99.0	99.0	94.0
Fusion@0.8	98.0	99.0	94.0

Table 5. Per class recall (%)

Model	Normal	Violence	Weaponized
I3D	85.0	89.0	99.0
C3D	80.0	89.0	88.0
Fusion@0.4	87.0	89.0	99.0
Fusion@0.5	**100.0**	90.0	99.0
Fusion@0.6	100.0	89.0	99.0
Fusion@0.7	100.0	88.0	99.0
Fusion@0.8	100.0	88.0	99.0

Table 6. Per class F1-score (%)

Model	Normal	Violence	Weaponized
I3D	91.0	85.0	96.0
C3D	82.0	82.0	91.0
Fusion@0.4	92.0	86.0	96.0
Fusion@0.5	**100.0**	**94.0**	**97.0**
Fusion@0.6	100.0	94.0	96.0
Fusion@0.7	99.0	93.0	96.0
Fusion@0.8	99.0	93.0	97.0

5.2.4 Discussion

The results indicate that the Stackelberg fusion significantly enhances both overall and class-wise performance metrics. The improvement is especially pronounced in the *Violence* class, where precision increases from 82% (I3D) and 76% (C3D) to 99% at $\theta = 0.5$, and the F1 score improves from 85% (I3D) to 94%.

Interestingly, for the *Normal* class, both precision and recall reach 100% at $\theta = 0.5$, showing a perfect classification for typical scenes. The *Weaponized* class also maintains a high recall 99% across all fusion thresholds while improving the F1 score compared to C3D alone.

5.3 Comparison with State-of-the-Art Approaches on the SCVD Dataset

To assess the effectiveness of the proposed hierarchical confidence-aware fusion framework, we conducted a comparative evaluation against several state-of-the-art approaches using the Smart City CCTV Violence Detection (SCVD) dataset. The selected baselines, also examined in [15], include the Flow-Gated Network (FGN) [26], ConvLSTM [27], and separable ConvLSTM [28], as well as the more recent SaliNet and SSIVD-Net models [15]. The results are reflected in Table 7.

Among earlier approaches, the Flow-Gated Network (FGN) [26], which integrates 3D convolutional operations to analyze motion through optical flow representations, achieved a peak accuracy of 74.4%. LSTM-based architectures, ConvLSTM [27] and SepConvLSTM [28], reported accuracies of 71.6% and 78.4%, respectively. These recurrent frameworks employ dynamic 2D convolutional filters derived from pre-trained backbones to capture spatial cues within individual frames, which are then temporally aggregated to model motion dynamics. Despite their computational efficiency, these methods exhibit limited capacity to capture the fine-grained temporal transitions that differentiate weaponized from non weaponized violent activities.

In contrast, recent SaliNet and SSIVD-Net architectures [15] demonstrated superior performance by adopting a novel Salient classifier built on a ResNet-inspired residual framework and Kervolution-based layers (KConv2D). Specifically, SaliNet introduces a minimal block structure designed to enhance the extraction of salient motion characteristics while maintaining energy efficiency. Among these variants, SaliNet-2m and SaliNet-4m achieved an accuracy of 86.6% and 83.1%, respectively, outperforming previous FGN and SepConvLSTM methods.

Our proposed method, the *Hierarchical Fusion of 3D CNNs with Confidence Awareness*, significantly extends these previous approaches by integrating two complementary 3D CNNs I3D and C3D through a Stackelberg game-theoretic fusion strategy. This confidence-aware hierarchical mechanism dynamically delegates the prediction task based on model confidence, enabling adaptive reasoning over complex spatio-temporal patterns. The approach achieved a maximum accuracy of 97.27% in the SCVD dataset, surpassing all baseline methods by a considerable margin.

Table 7. Comparison of state-of-the-art methods on the SCVD dataset in terms of classification accuracy (%)

Method	**Year**	**Accuracy (%)**
Flow-Gated Network (FGN) [26]	2020	74.4
ConvLSTM [27]	2017	71.6
Sep-ConvLSTM [28]	2021	78.4
SaliNet-4m [15]	2024	83.1
SaliNet-2m [15]	2024	86.6
Proposed Hierarchical Fusion (I3D + C3D)	2025	**97.27**

The comparison summarized in Table 7 highlights a consistent progression in accuracy across successive architectures, culminating in the proposed confidence-aware hierarchical fusion model. This substantial improvement demonstrates the benefit of incorporating multilevel temporal reasoning and adaptive confidence-based inference for real-world smart city surveillance applications.

6 Conclusion and Future Perspectives

Distinguishing aggressive and threatening behaviors within high-temporal-resolution urban surveillance networks remains a challenging task. This study proposed an innovative decision-level fusion framework that integrates confidence evaluation and hierarchical model synchronization. The approach is inspired by game theory and exploits the complementary strengths of two state-of-the-art deep learning architectures: I3D, which excels in deep spatial-temporal reasoning, and C3D, which specializes in capturing short-term motion dynamics.

In the proposed framework, the I3D model acts as the primary decision-maker, while the C3D model is invoked whenever the confidence score of the leader model falls below a predefined threshold. This confidence-driven delegation process enhances the system's robustness in uncertain environments, enabling accurate decision-making within near real-time constraints.

The system was evaluated on a three-class surveillance dataset comprising normal, violence, and weaponized categories. Experimental results demonstrated that the hierarchical fusion strategy performs consistently in individual models. The proposed method achieved an overall accuracy of 97.27% at a confidence threshold of 0.5, with significant improvements in the F1 scores for the Violence and Weaponized classes. Thus, it largely surpasses the best-performing modern baselines. Specifically, our approach improves on SaliNet-2m (86.6%) [15] by approximately +10.7%, and on SepConvLSTM (78.4%) [28] by about +18.9%, representing a consistent and significant performance gain over recent approaches. These findings confirm the effectiveness of confidence-based fusion in achieving stable and context-sensitive video understanding.

Beyond the current results, several directions can be explored in future work. First, the framework can be tested on additional public video surveillance datasets, such as UCF-Crime-DVS dataset [29] , to assess its generalization capabilities across different recording conditions and visual domains. Second, integrating distributed streaming platforms such as Apache Kafka would allow continuous data ingestion and adaptive model coordination, facilitating large-scale deployment in smart city infrastructures. Furthermore, utilizing multi-threading and asynchronous data pipelines could enhance the system's responsiveness, making it suitable for real-time monitoring scenarios with multiple simultaneous video streams.

References

[1] K. Shankar, V. Iyer, K. Iyer, and A. Pandhare, "Intelligent video analytics (IVA) and surveillance system using machine learning and neural networks," in *2020 International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 2020, pp. 623–627. Available: https://doi.org/10.1109/ICICT48043.2020.9112527

[2] B. Ardabili, A. Pazho, G. Noghre, C. Neff, S. Bhaskararayuni, A. Ravindran, and H. Tabkhi, "Understanding policy and technical aspects of AI-enabled smart video surveillance to address public safety," *Computational Urban Science*, vol. 3, no. 21, 2023. Available: https://doi.org/10.1007/s43762-023-00097-8

[3] T. Manesh, N. Nataraj, A. Jayaraj, A. Joby, P. Ananthakrishnan, and B. Thankachan, "A survey on video anomaly detection in surveillance system," in *2024 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. IEEE, 2024, pp. 1–5. Available: https://doi.org/10.1109/RAICS61201.2024.10690095

[4] N. Abirami, G. Radhika, and N. Radhika, "Automated teller machine security and robbery prevention based on human behaviour analysis," in *2023 Innovations in Power and Advanced Computing Technologies (i-PACT)*. IEEE, 2023, pp. 1–6. Available: https://doi.org/10.1109/i-PACT58649.2023.10434470

[5] T. Pham, H. Vu, T. Nguyen, S. Phan, and V. Pham, "Utilizing Deep Learning Models to Develop a Human Behavior Recognition System for Vision-Based School Violence Detection," in *2024 7th*

International Conference on Green Technology and Sustainable Development (GTSD). IEEE, 2024, pp. 189–193. Available: https://doi.org/10.1109/GTSD62346.2024.10674972

[6] M. Iftee, M. Rahman, and S. Das, "VioNet: An enhanced violence detection approach for videos using a fusion model of Vision Transformer with Bi-LSTM and 3D convolutional neural networks," in *Proceedings of the 2nd International Conference on Big Data, IoT and Machine Learning. BIM 2023. Lecture Notes in Networks and Systems*. Springer, 2023, vol. 86, pp. 139–151. Available: https://doi.org/10.1007/978-981-99-8937-9_10

[7] W. Jin, L. Zhu, and J. Sun, "Aligning First, Then Fusing: A novel weakly supervised multimodal violence detection method," *Knowledge-Based Systems*, vol. 322, article 113709, 2025. Available: https://doi.org/10.1016/j.knosys.2025.113709

[8] M. Ramzan, A. Abid, H. Khan, S. Awan, A. Ismail, M. Ahmed, and A. Mahmood, "A review on state-of-the-art violence detection techniques," *IEEE Access*, vol. 7, pp. 107 560–107 575, 2019. Available: https://doi.org/10.1109/ACCESS.2019.2932114

[9] S. Das, A. Sarker, and T. Mahmud, "Violence detection from videos using HOG features," in *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*. IEEE, 2019, pp. 1–5. Available: https://doi.org/10.1109/EICT48899.2019.9068754

[10] A. Guedes and G. Chávez, "Real-time violence detection in videos using dynamic images," in *2020 XLVI Latin American Computing Conference (CLEI)*. IEEE, 2020, pp. 503–511. Available: https://doi.org/10.1109/CLEI52000.2020.00065

[11] N. Su, L. Sun, Y. Gao, J. Wu, and X. Wu, "Violence detection in videos via motion-guided global and local views," in *2023 8th International Conference on Data Science in Cyberspace (DSC)*. IEEE, 2023, pp. 437–442. Available: https://doi.org/10.1109/DSC59305.2023.00069

[12] E. AlQaralleh, F. Aldhaban, H. Nasseif, M. Alksasbeh, and B. Alqaralleh, "Smart deep learning-based human behaviour classification for video surveillance," *Computers, Materials & Continua*, vol. 72, no. 3, pp. 5593–5605, 2022. Available: https://doi.org/10.32604/cmc.2022.026666

[13] S. Putri, A. Rifai, and I. Nawawi, "Physical violence detection system to prevent student mental health disorders based on deep learning," *Jurnal Pilar Nusa Mandiri*, vol. 19, no. 2, pp. 103–108, 2023. Available: https://doi.org/10.33480/pilar.v19i2.4600

[14] K. Sahay, B. Balachander, B. Jagadeesh, G. Kumar, R. Kumar, and L. Parvathy, "A real-time crime scene intelligent video surveillance system in violence detection framework using deep learning techniques," *Computers and Electrical Engineering*, vol. 103, article 108319, 2022. Available: https://doi.org/10.1016/j.compeleceng.2022.108319

[15] T. Aremu, Z. Li, R. Alameeri, M. Khan, and A. El Saddik, "SSIVD-Net: A Novel Salient Super Image Classification and Detection Technique for Weaponized Violence," in *Intelligent Computing. SAI 2024. Lecture Notes in Networks and Systems, vol. 1018*, K. Arai, Ed. Springer, 2024, pp. 16–35. Available: https://doi.org/10.1007/978-3-031-62269-4_2

[16] H. Jahlan and L. Elrefaei, "Mobile neural architecture search network and convolutional long short-term memory-based deep features toward detecting violence from video," *Arabian Journal for Science and Engineering*, vol. 46, no. 9, pp. 8549–8563, 2021. Available: https://doi.org/10.1007/s13369-021-05589-5

[17] W. Pang, W. Xie, Q. He, Y. Li, and J. Yang, "Audiovisual dependency attention for violence detection in videos," *IEEE Transactions on Multimedia*, vol. 25, pp. 4922–4932, 2022. Available: https://doi.org/10.1109/TMM.2022.3184533

[18] H. Mohammed and L. Elrefaei, "Detecting violence in video based on deep features fusion technique," *arXiv preprint arXiv:2204.07443*, pp. 1–4, 2022.

[19] M. Asad, J. Yang, J. He, P. Shamsolmoali, and X. He, “Multi-frame feature-fusion-based model for violence detection,” *The Visual Computer*, vol. 37, no. 6, pp. 1415–1431, 2021. Available: https://doi.org/10.1007/s00371-020-01878-6

[20] M. Abdullah, H. Karim, and N. AlDahoul, “A combination of light pre-trained convolutional neural networks and long short-term memory for real-time violence detection in videos,” *International Journal of Technology*, vol. 14, no. 6, pp. 1228–1236, 2023. Available: https://doi.org/10.14716/ijtech.v14i6.6655

[21] M. Khan, W. Gueaieb, A. Elsaddik, G. De Masi, and F. Karray, “Graph-based knowledge driven approach for violence detection,” *IEEE Consumer Electronics Magazine*, vol. 14, no. 1, pp. 77–85, 2024. Available: https://doi.org/10.1109/MCE.2024.3446192

[22] J. Carreira and A. Zisserman, “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4724–4733. Available: https://doi.org/10.1109/CVPR.2017.502

[23] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning Spatiotemporal Features with 3D Convolutional Networks,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 4489–4497. Available: https://doi.org/10.1109/ICCV.2015.510

[24] X. S. Gao, S. Liu, and L. Yu, “Achieve Optimal Adversarial Accuracy for Adversarial Deep Learning using Stackelberg Game,” *arXiv preprint arXiv:2207.08137*, pp. 1–12, 2022. Available: https://doi.org/10.48550/arXiv.2207.08137

[25] D. M. W. Powers, “Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation,” *International Journal of Machine Learning Technology*, vol. 2, no. 1, pp. 37–63, 2011. Available: https://doi.org/10.48550/arXiv.2010.16061

[26] M. Cheng, K. Cai, and M. Li, “RWF-2000: An open large scale video database for violence detection,” in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 4183–4190. Available: https://doi.org/10.1109/ICPR48806.2021.9412502

[27] S. Sudhakaran and O. Lanz, “Learning to detect violent videos using convolutional long short-term memory,” in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2017, pp. 1–6. Available: https://doi.org/10.1109/AVSS.2017.8078468

[28] Z. Islam, M. Rukonuzzaman, R. Ahmed, M. H. Kabir, and M. Farazi, “Efficient two-stream network for violence detection using separable convolutional LSTM,” in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8. Available: https://doi.org/10.1109/IJCNN52387.2021.9534280

[29] Y. Qian, S. Ye, C. Wang, X. Cai, J. Qian, and J. Wu, “UCF-Crime-DVS: A Novel Event-Based Dataset for Video Anomaly Detection with Spiking Neural Networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 6, 2025, pp. 6577–6585. Available: https://doi.org/10.1609/aaai.v39i6.32705