



Robust Violence Detection in Heterogeneous Surveillance Videos using a Dual-Branch Spatiotemporal Network

Sai Thu Ya Aung¹, Worapan Kusakunniran², Yu Nandar Aung³ and Chattapatr Leeraha⁴

ABSTRACT

Automatic detection of violent events in surveillance video is a critical component of modern public safety systems. In contrast, existing methods often show limited generalization, performing well on individual datasets but degrading under domain shifts in real-world environments. To address this generalization challenge, this paper proposes an Asymmetric Dual-Branch Network trained on a large-scale Multi-Source Dataset integrating nine heterogeneous benchmark datasets. The proposed architecture combines a motion branch based on ResNet-18 trained from scratch to capture motion-specific representations with a spatial branch employing an X3D backbone selected through comparative evaluation of four state-of-the-art architectures, to learn complementary spatial semantic representations. This asymmetric design effectively balances detection accuracy and real-time performance while improving robustness across heterogeneous scenarios. Extensive experiments using 5-fold cross-validation demonstrate stable performance across multiple datasets. Notably, the proposed framework achieves state-of-the-art results on the Bus-violence dataset (89.07%) and RLVS dataset (96.65%), while maintaining a high processing rate of 76.61 FPS. These empirical results, including validated performance on the unseen Kranok-NV dataset, confirm that the asymmetric integration of decoupled features offers a practical and scalable solution for real-time urban surveillance in uncontrolled environments.

Article information:

Keywords: Violence Detection, Asymmetric Dual-branch Architecture, Multi-source Learning, Cross-Dataset Generalization

Article history:

Received: February 21, 2026

Revised: April 6, 2026

Accepted: July 1, 2026

Published: July 18, 2026

(Online)

DOI: 10.37936/ecti-cit.2026203.266320

1. INTRODUCTION

Video surveillance systems have become ubiquitous in modern urban infrastructure and generating massive volumes of data daily. Manually analyzing this footage is labor-intensive and error-prone, creating a demand for automated solutions capable of detecting violent events in real time. Violence detection, unlike general action recognition, presents unique challenges. Violent acts are often short-lived, involve rapid and subtle motions, and can occur in highly cluttered or occluded environments. While recent advancements in Deep Learning (DL) have yielded promising results, existing methods often suffer from limited generalization. Most state-of-the-art models are trained and evaluated on single, homogeneous datasets, such as Hockey Fight or Movie Fight,

where the background, lighting, and camera angles are relatively consistent. Consequently, these models frequently fail when deployed in real-world scenarios that differ from their training distribution, a problem known as a domain shift.

To address this limitation, this study proposes a robust violence detection framework designed for high generalization across diverse surveillance environments. A large-scale Multi-Source Dataset was prepared by integrating nine benchmark datasets, covering a wide spectrum of scenarios from sports and movies to real-world CCTV footage.

The primary methodological contribution is an asymmetric dual-branch architecture that explicitly decouples the kinetic motion from spatiotemporal feature extraction. Unlike symmetric two-branch networks that implement identical backbones or com-

^{1,2,3,4}The authors are with the Faculty of Information and Communication Technology, Mahidol University, Salaya, Phutthamonthon, Nakhon Pathom, Thailand 73170, Email: saithuyaaung1@gmail.com, worapan.kun@mahidol.ac.th, yunandar.yn@gmail.com and chattapatr.ler@student.mahidol.ac.th

²Corresponding author: worapan.kun@mahidol.ac.th

putationally expensive optical flows, the proposed approach integrates a lightweight backbone (trained from scratch) with a powerful spatiotemporal branch (implementing pre-trained Transformers or CNNs). By substituting dense optical flow with a Temporal Difference Volume, the framework effectively isolates violent dynamics from background noise while maintaining the efficiency required for real-time edge deployment. Extensive experiments using 5-fold cross-validation demonstrate that this approach not only achieves competitive accuracy on standard benchmarks but also outperforms single-source baselines in cross-domain generalization.

The main contributions of this study are summarized as follows:

Asymmetric Dual-Branch Integration: A specialized asymmetric framework is proposed that tackles motion and spatial feature learning. A lightweight ResNet branch is trained from scratch to capture pure motion dynamics from temporal difference maps, whereas the second branch with a pre-trained backbone extracts complex spatiotemporal features. This design provides a more efficient alternative to symmetric or optical-flow-based architectures.

Robust Multi-Source Generalization: A unified learning strategy across nine diverse benchmarks is implemented to reduce single-source bias. The framework’s robustness is validated through state-of-the-art performance in difficult environments such as Bus-violence (89.07%) and proven zero-shot generalization on the unseen Kranok-NV dataset, confirming its adaptability to unencountered real-world scenarios.

High-Efficiency Performance: Through an extensive comparative analysis of state-of-the-art backbones (including TimeSformer, Swin-T, and MobileNetV3), the proposed framework demonstrates that X3D-based integration and feature fusion via concatenation achieve a processing rate of 76.61 FPS. This provides a quantitative justification for real-time deployment on edge-surveillance hardware, significantly outperforming transformer-based architectures in both latency and memory stability.

2. RELATED WORK

2.1 Hybrid Modeling and Spatiotemporal Features Fusion

Deep learning has significantly advanced video analysis by enabling the extraction of high-level semantic representations beyond handcrafted features, including Optical Flow [1], Histogram of Oriented Gradients (HOG) [2], and Motion Interaction Fields (MIF) [3]. Recent developments emphasize the use of ensemble strategies to improve robustness by aggregating predictions from multiple models, as utilized by maximum occurrence and probability scores to stabilize classification performance [4]. In the specific domain of violence detection, Convolutional

Long Short-Term Memory (Conv-LSTM) networks have been employed to capture spatiotemporal dependencies by processing video segments and aggregating features for final decision-making [5]. Similarly, the deep ensemble classifier introduced in [6] employs heterogeneous models and majority voting for detection in surveillance scenarios. Unlike these traditional ensemble methods [4],[6] that stack heavy, independent models resulting in linear scaling of computational costs, our framework decouples spatial and temporal processing into a streamlined, highly integrated dual-branch design. This architecture effectively captures both static scene contexts and pure dynamic motion cues without the burden of redundant feature extraction.

2.2 Multi-Branch Architectures

The dual-branch approach, introduced by the Two-Stream ConvNet [7], separates spatial and motion processing to capture different features. Building on this, the convolutional fusion architecture proposed in [8] demonstrated that feature-level fusion at intermediate layers significantly outperforms late-stage decision fusion, enabling the learning of richer spatiotemporal information. Architectures such as SlowFast [9] further extend this concept by processing video at multiple temporal resolutions through parallel pathways. Similarly, multi-branch designs have proven effective for identifying sudden movements. For instance, the FallNeXt model [10] aggregates features from multiple convolutional branches to capture diverse motion patterns in sensor-based fall detection. Despite their effectiveness, these models typically rely on large-scale pre-training on image datasets, which can introduce a bias toward appearance features rather than motion representations. Furthermore, widely used multi-branch frameworks such as SlowFast [9] require large-scale, 3D convolutional networks in both streams, leading to intense resource consumption. In contrast, the proposed approach minimizes this footprint by pairing an ultra-lightweight, expanded Conv1 ResNet-18 motion branch with an optimized spatiotemporal stream, achieving high-throughput validation without relying on heavy parallel architectures.

2.3 Asymmetric Designs for Violence Detection

Understanding violence requires not only motion cues but also the spatial relationships between interacting subjects. The interpretation of spatial relationships, such as proximity and relative motion between objects, is essential for scene understanding in complex video streams [11]. In the context of violence detection, prior work has incorporated spatial interaction cues through a dual-branch framework that fuses human pose estimation with motion dynamics

[12]. This method employs a weighted feature fusion strategy to prioritize relevant features, achieving strong performance by reducing background interference through skeletal isolation. While these methods can reduce background noise by tracking skeletons, they often struggle with low-quality video or the high computational cost and latency of running extra pose detection calculations.

In contrast to existing frameworks that rely on heavy multi-model ensembles [4],[6] or extra pre-processing steps like optical flow and skeletal tracking [12], the proposed dual-branch architecture introduces three key contributions. First, it uses an asymmetric design that completely separates tasks, using one backbone for spatial scene context and a different, specialized backbone for motion features. Second, to prevent the model from overfitting to background textures, the motion branch utilizes a modified ResNet-18 trained entirely from scratch, forcing it to focus purely on dynamic motion movements. Finally, by replacing complex motion-estimation pipelines with lightweight inter-frame difference maps, the framework minimizes processing overhead. This streamlined ensures the system achieves high-throughput, real-time performance while maintaining strong generalization capabilities across unseen surveillance environments.

3. METHODOLOGY

In this section, the proposed Asymmetric Dual-Branch Fusion Network is presented for robust violence detection. The framework is designed to explicitly decouple the learning of static appearance cues and dynamic motion patterns, fusing them only at the semantic level to ensure high generalization across diverse surveillance environments. Unlike symmetric two-branch architectures, this model implements asymmetric integration of two branches from different backbones to optimize the trade-off between representational capacity and inference latency. Figure 1 illustrates the overall architecture.

3.1 Problem Formulation

Let V be an input video clip consisting of T sequential frames, represented as a four-dimensional tensor $X \in \mathbb{R}^{T \times H \times W \times 3}$ where R denotes the set of real numbers. Here, T is the temporal dimension (number of frames), H and W represent the spatial height and width, respectively, and $C = 3$ corresponds to the RGB color channels. The objective is to learn a mapping function $F(X) \rightarrow \hat{y}$ where $\hat{y} \in \{0, 1\}$ is the predicted class label representing non-violent or violent events, respectively.

3.2 Asymmetric Dual-Branch Architecture

To capture both the semantic context and rapid temporal dynamics, the input X is processed through

two distinct backbones: a high-capacity Appearance Branch and a lightweight Motion Branch. This asymmetric design is a deliberate methodological choice to ensure real-time feasibility on edge hardware.

3.2.1 Appearance Branch (Spatial and Spatiotemporal Semantics)

The appearance branch extracts high-level spatial features from individual frames. A pre-trained backbone is employed as a feature extractor φ_{rgb} .

Given the input clip X , the encoder processes each frame x_t independently to generate a sequence of frame-level feature vectors as follows:

$$f_t = \varphi_{rgb}(x_t), \quad t \in \{1, \dots, T\} \quad (1)$$

where $f_t \in \mathbb{R}^{D_{rgb}}$ is the latent representation of the t -th frame. To obtain a global representation of the video content, temporal average pooling is employed across the sequence:

$$v_{rgb} = \frac{1}{T} \sum_{t=1}^T f_t \quad (2)$$

This aggregation ensures that the model retains semantic information about salient objects (e.g., weapons), regardless of their temporal position within the clip.

3.2.2 Motion Branch (Temporal Motion Dynamics)

Standard 3D convolutions are computationally expensive and often struggle to isolate pure motion from the background texture. To address this, the lightweight motion branch processes motion cues via inter-frame differencing. Let d_t represent the pixel-wise difference between the consecutive frames as follows:

$$d_t = |x_{t+1} - x_t|, \quad t \in \{1, \dots, T-1\} \quad (3)$$

The selection of frame differencing over dense optical flow or background subtraction is a strategic design choice driven by the requirement for real-time inference on surveillance edge devices. Although optical flow is robust to complex camera ego-motion, its high computational overhead introduces significant latency. Similarly, background subtraction methods (e.g., Gaussian Mixture Models) are effective at isolating foreground objects; however, they are highly sensitive to illumination changes and typically require a static background model, making them unsuitable for the heterogeneous datasets used in this study, which feature diverse lighting. In the context of fixed-mount surveillance devices, frame differencing acts as an efficient, high-pass temporal filter. It effectively suppresses static background information while highlighting the kinetic motion of the moving subjects.

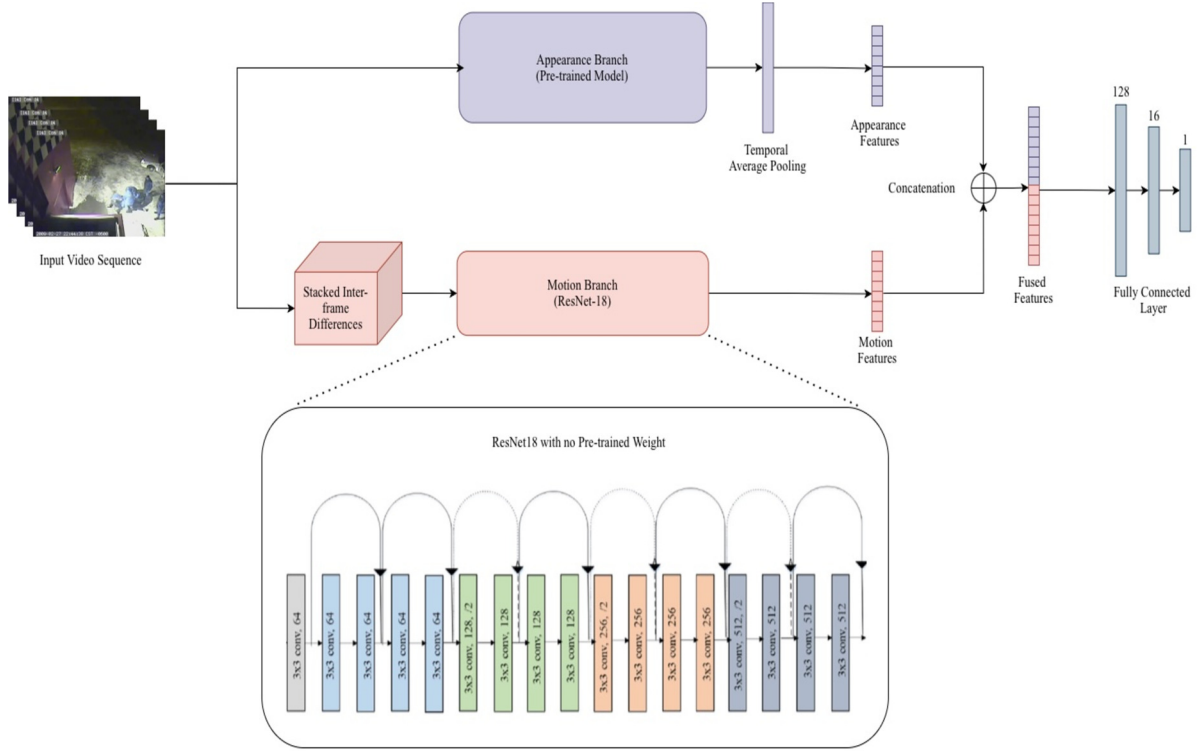


Fig.1: Architecture of the proposed Asymmetric Dual-Branch Network. The top branch (appearance branch) extracts spatial semantics using a pretrained backbone. The bottom branch (motion branch) extracts motion information through inter-frame differencing and processes the resulting motion volume using ResNet-18 [13]. Late fusion via concatenation combines the appearance and motion cues for final violence classification.

Temporal modeling within the motion branch is achieved by processing the stacked frame difference maps through a lightweight ResNet-18 backbone to form a single motion volume $M \in R^{H \times W \times 3(T-1)}$. The first convolutional layer was modified to accept $3(T-1)$ input channels, allowing the network to observe the entire temporal window simultaneously.

The entire motion volume is processed by the ResNet-18 backbone φ_{motion} rather than frame-by-frame and, accepts the motion feature vector, which is obtained as

$$v_{motion} = \varphi_{motion}(M) \quad (4)$$

where $v_{motion} \in R^{D_{motion}}$ encodes the dynamic patterns of the entire sequence, such as the velocity and trajectory of moving subjects.

3.2.3 Feature Fusion and Classification

The distinct representations from the two branches were fused to form a comprehensive spatiotemporal descriptor. A late fusion strategy was explored through three element-wise operations: concatenation, multiplication, and summation.

$$v_{fused} = v_{rgb} \oplus \varphi_{motion} \quad (5)$$

$$v_{fused} = v_{rgb} \odot \varphi_{motion} \quad (6)$$

$$v_{fused} = v_{rgb} + \varphi_{motion} \quad (7)$$

where $\oplus$ denotes concatenation, $\odot$ denotes multiplication, and $+$ denotes the summation operation, resulting in the combined feature v_{fused} .

Based on the empirical results presented in Table 3, concatenation was identified as the most stable and effective integration method among the evaluated operations. Furthermore, to prevent overfitting and enhance generalization, a dropout layer with probability $p = 0.5$ is applied to v_{fused} . Finally, the regularized vector is passed through a fully connected classification head (MLP) to produce the class logit:

$$\hat{y} = \text{softmax}(\text{MLP}(v_{fused})) \quad (8)$$

3.3 Architectural Selection for Appearance Branch

To evaluate different spatiotemporal extraction paradigms, four state-of-the-art backbones representing distinct computational and modeling trade-offs were selected for the appearance branch. Specifically, X3D [14] was chosen for its efficient progressive expansion across spatial, temporal, and channel axes, making it highly suited for high-throughput surveillance. TimeSformer [15] was integrated to assess the impact of pure attention via divided space-time op-

erators capable of capturing long-range historical dependencies. Swin Transformer [16] was selected for its hierarchical structure and shifted window mechanism to extract multi-scale contextual features. At the same time, MobileNetV3 [17] served as an edge-deployment baseline designed via platform-aware Architecture Search to establish a lower bound for computational costs.

Moreover, these models were pretrained on large-scale datasets. The **X3D** and **TimeSformer** models were trained using the **Kinetics-400** dataset [18], which contains over 400,000 video clips covering 400 human action classes. For the MobileNetV3, weights were pre-trained on **ImageNet-1K** [19], which contains over 1 million images of 1,000 object categories. The **Swin Transformer** was pretrained using the **Something-Something V2** dataset [20], which contains 220,847 video clips across 174 classes of human-object interactions. This pretraining phase is critical because it provides the models with a foundational understanding of complex human kinematics and interactions with objects.

4. MULTI-SOURCE DATASET CONSTRUCTION

To address the limited generalization capability of existing violence detection models, a large-scale Multi-Source Dataset was constructed by integrating eight distinct benchmarks. Most prior studies rely on single, homogeneous datasets (e.g., Hockey Fight or Movie Fight), which fail to capture the wide variance in lighting, camera angles, and background clutter found in real-world surveillance scenarios.

The combined dataset encompasses a diverse range of scenarios, ensuring high heterogeneity:

Surveillance Footage: Real-world CCTV clips from RWF-2000, RLVS, Crowd Violence, VioPeru, SCFD, and CamNuVem, capturing authentic violent events in unconstrained environments.

Sports & Movies: High-quality, action-rich sequences from Hockey Fight and Movie Fight, providing clear examples of aggressive motion dynamics.

Staged Scenarios: Controlled enactments from the Bus Violence dataset, offering consistent interactions in a specific public transport context.

This integration resulted in a total of 8,398 videos, balanced between the violent and non-violent classes. The heterogeneity of this collection ensures that the model learns robust features applicable to various domains, from crowded streets to confined indoor spaces to outdoor environments. Table 1 presents the composition of the integrated dataset.

4.1 Diversity of Violence Classes and Dataset Analysis

To enforce robust cross-domain generalization, the integrated Multi-Source Dataset is structured to re-

fect high heterogeneity across visual qualities and behavioral contexts, mitigating single-source bias. As classified in Table 1, the repository balances three distinct domain paradigms: (1) General Surveillance (RWF-2000, RLVS, VioPeru, SCFD) which captures real-world physical altercations in diverse geographic settings characterized by unscripted body language, variable resolutions, and frequent passenger/seat occlusions; (2) High-Dynamics Benchmarks (Hockey Fight, Movie Fight) featuring high-definition, fast-paced aggressive motions with distinct foreground-background separation; and (3) Specialized Scenarios (Crowd Violence, Bus Violence, CamNuVem) targeting mass crowd panic, confined public transport interactions, and instrumental violent actions like armed robberies. This comprehensive integration ensures the model learns invariant, actionable features across unconstrained settings rather than overfitting to fixed lighting, camera viewpoints, or isolated visual styles.

4.2 Data Partitioning and Training Protocol

To evaluate cross-domain generalization capabilities, the integrated dataset (Table 1) was partitioned into an 80% training set (6,718 clips; $n = 3,359$ per class) and an independent 20% hold-out testing set (1,680 clips; $n = 840$ per class). This split was executed independently for each underlying source using strict class stratification to ensure proportional representation. Hyperparameters were optimized using a stratified 5-fold cross-validation protocol on the training partition. In each iteration, the framework was trained on 80% of the training data, while the remaining 20% served as an independent fold for model selection and checkpoint monitoring. This process was repeated five times so that every training clip served as validation data exactly once, with the optimal checkpoint from each fold retained for final evaluation on the unseen hold-out testing set.

The complete framework was implemented using the PyTorch library and executed on an HPC cluster node equipped with a single NVIDIA A100 GPU. The temporal volume was constructed using a fixed window size of $T = 8$ consecutive frames to ensure robust kinetic context. All architectural variants were trained for 20 epochs using the standard cross-entropy loss function and the AdamW optimizer with a constant learning rate of 1×10^{-4} , a batch size of 16 clips, and a weight decay of 1×10^{-2} .

5. EXPERIMENTS AND RESULTS

5.1 Evaluation on Standalone Appearance and Motion Branches

To validate the effectiveness of the dual-branch design, an ablation study on the performance of each branch was first conducted. The evaluation compared appearance-only branches (using both pre-trained and scratch setups) against the motion-

Table 1: Summary of the Multi-Source Violence Detection Dataset.

Dataset	Type	Non-Violent	Violent	Total	Train Set	Test Set
1. Movie Fight [3]	Movie	100	100	200	160	40
2. RLVS [21]	Surveillance	1,000	1,000	2,000	1,600	400
3. Hockey Fight [3]	Sport	500	500	1,000	800	200
4. Crowd Violence [22]	Surveillance	123	123	246	196	50
5. RWF-2000 [23]	Surveillance	1,000	1,000	2,000	1,600	400
6. SCFD [24]	Surveillance	150	150	300	240	60
7. VioPeru [25]	Surveillance	140	140	280	225	55
8. Bus Violence [26]	Staged	700	700	1,400	1,120	280
8. Bus Violence [27]	Surveillance	486	486	972	777	195
Total	Mixed	4,199	4,199	8,398	6,718	1,680

Table 2: Performance Comparison of Standalone Branches on the Hold-Out Test Set.

Configuration	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Accuracy
Appearance-Only (RGB, Pretrained)						
X3D (Pretrained)	0.8942	0.9083	0.9039	0.8867	0.9031	(0.8993±77)
MobileNetV3 (Pretrained)	0.8869	0.8804	0.8881	0.8929	0.8869	(0.8870±40)
Swin-T (Pretrained)	0.8869	0.8863	0.9023	0.8917	0.8929	(0.8920±58)
TimeSformer (Pretrained)	0.6321	0.6196	0.6607	0.6899	0.6952	(0.6595±301)
Appearance-Only (RGB, Training from Scratch)						
X3D (Scratch)	0.7863	0.7958	0.7792	0.7946	0.7958	(0.7904±66)
MobileNetV3 (Scratch)	0.7738	0.7679	0.7643	0.7589	0.7744	(0.7679±58)
Swin-T (Scratch)	0.5788	0.6036	0.5786	0.5702	0.5768	(0.5815±114)
TimeSformer (Scratch)	0.6405	0.6685	0.6440	0.6440	0.6542	(0.6502±10)
Motion-Only (Diff, Pretrained)						
ResNet-18 (Pretrained)	0.7845	0.7869	0.7785	0.7827	0.7851	(0.7835±28)
Motion-Only (Diff, Training from Scratch)						
ResNet-18 (Scratch)	0.8036	0.7726	0.8018	0.7952	0.8018	(0.7950±115)

Table 3: Accuracy of Feature Fusion Strategies in the X3D-Based Dual-Branch Framework on the Hold-Out Test Set.

Fusion Method	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Accuracy
Summation	0.9143	0.8952	0.8958	0.9071	0.9107	(0.9046±86)
Multiplication	0.9077	0.9042	0.9065	0.9060	0.9077	(0.9064±15)
Concatenation	0.9071	0.9089	0.9107	0.9024	0.9125	(0.9083±38)

only branch (ResNet-18 using pretrained with ImageNet1K and scratch setups) on the independent hold-out test set across all five folds.

As summarized in Table 2, the appearance-only branches exhibit severe performance degradation when trained from scratch. For instance, the top-performing X3D architecture drops from an average accuracy of 89.93% (pretrained) to 79.04% (scratch). This drop underscores the dependence of spatial backbones on large-scale pretraining to interpret static scene environments effectively.

Conversely, the motion-only branch (ResNet-18) demonstrates the exact opposite behavior; training from scratch yields an average accuracy of 79.50%, outperforming its pretrained counterpart (78.35%). This key finding quantitatively validates our asymmetric initialization strategy. It demonstrates that resetting the weights of the motion branch eliminates static appearance biases inherited from pretrained in-

formation, forcing the network to converge on highly specialized, context-independent spatiotemporal patterns essential for cross-domain violence detection.

5.2 Evaluation of Fusion Strategies

After identifying X3D as the optimal appearance backbone, the impact of different fusion mechanisms was investigated to integrate the appearance and motion branches. Three fundamental operations were evaluated on the dual-branch architecture: element-wise summation, element-wise multiplication, and feature concatenation. Each approach was benchmarked using the independent hold-out test set across all five folds.

As summarized in Table 3, feature concatenation emerged as the most effective strategy, yielding the highest mean accuracy of 90.83% and outperforming both standalone branches and alternative fusion

methods. To evaluate the mathematical stability of these gains, a paired-samples effect size analysis was conducted across the cross-validation folds between the top-performing Concatenation strategy and the Multiplication baseline. Specifically, Cohen's d was computed by dividing the mean of the fold-wise accuracy differences by their standard deviation. This statistical analysis yielded an effect size of $d \approx 0.50$, which represents a robust, medium practical effect. This finding demonstrates that the absolute accuracy improvements brought by concatenation are consistent and meaningful across the evaluation folds. This margin confirms that concatenation successfully preserves a distinct, non-redundant feature space from both branches, allowing the final classifier to optimize between spatial scene context and motion dynamics.

5.3 Evaluation of Proposed Architecture with Different Appearance Backbones

Following the architectural validation in Section 5.2, which shows that the dual-branch framework with concatenation fusion provides better performance than single-branch architectures, this section evaluates the effectiveness of various appearance backbones in the dual-branch framework. Table 4 presents the performance assessment of four dual-branch architectures with different appearance branches on nine heterogeneous test sets and one unseen dataset. The results indicate that the choice of appearance backbone significantly influences the model's ability to generalize across diverse surveillance environments.

The dual-branch architecture with the X3D backbone demonstrated the most consistent performance across the test sets. It achieves the highest accuracy on the Bus-violence (89.07%), Crowd Violence (92.00%), RWF-2000 (88.05%), and Kranok-NV (86.00%) datasets among dual-branch architectures. This performance suggests that the 3D-convolutional approach of X3D is more effective at capturing essential spatiotemporal features from limited and heterogeneous data than pure attention-based (TimeSformer and Swin-T) and CNN-based (MobileNetV3) models.

In contrast, while the dual-branch architecture with the MobileNetV3 backbone shows strong results on high-resolution datasets such as Hockey (96.12%) and Movie Fight (93.85%), its performance is less stable than the X3D-based architecture when applied to surveillance datasets such as SCFD (72.00%). This limitation may be attributed to pre-training on ImageNet-1K and its relatively lightweight architectural design. Although the TimeSformer architecture was pretrained with the Kinetics-400 dataset, which is the same as X3D, it consistently shows the lowest accuracy across all benchmarks, indicating that pure attention mechanisms may require more extensive training sets to learn robust features. In con-

trast, the transformer-based Swin-T architecture pretrained with Something-Something V2 showed the highest performance on the Movie (93.85%) and RLVS (96.95%) datasets. This can be attributed to its hierarchical design and shifted window mechanism, which enable the efficient modeling of both local and global contextual information.

5.4 Dataset-wise Analysis of Generalization Capability

The generalization performance of the proposed asymmetric dual-branch framework was evaluated across multiple heterogeneous test sets, with detailed results presented in Table 5. Following the validation of various appearance backbones, the X3D architecture was selected for this comprehensive analysis because it demonstrated superior performance in the initial validations. The X3D-based asymmetric dual-branch model was evaluated on each test set and on the entirely unseen Kranok-NV dataset. This external benchmark consists of 3,683 video clips, with 3,000 non-violent and 683 violent samples.

The model achieved exceptional performance on Hockey (96.02% accuracy) and RLVS (96.65% accuracy) datasets. In these scenarios, the F1-scores for both classes remained highly balanced, confirming that the model can extract discriminative motion dynamics and semantic representations from both low-quality surveillance and high-resolution sports footage. Consistent stability was also observed across other complex environments, including the Crowd (92.00%) and Movie (89.74%) datasets. The high precision in the Movie dataset (0.9567 for the violent class) indicates that the framework is robust against dramatic camera shifts and varied editing styles of cinematic content. Furthermore, the results for the Bus-violence dataset (89.07%) demonstrate the efficacy of the dual-branch integration in identifying aggressive interactions within the confined spatial constraints of public transportation.

A significant indicator of the framework's robustness is its performance on the unseen Kranok-NV dataset, where it achieved an average accuracy of 86.00%, although the model had no prior exposure to this environment's specific information. While the high recall for the non-violent class (0.9417) suggests a superior performance, the precision for the violent class (0.6763) reflects the inherent difficulty of zero-shot deployment. The qualitative examination of this performance gap reveals that it is heavily driven by a distinct shift in action dynamics. While our training datasets (such as Hockey Fight and RWF-2000) define violence through multi-person physical contact, occlusions, and chaotic interpersonal interactions, the violent class in Kranok-NV predominantly features individuals engaged in solo boxing practice or isolated martial arts drills. It occasionally misclassifies these clean, rapid, single-subject athletic movements

Table 4: Inference Accuracy Comparison of Dual-branch Architectures with Different Appearance Backbones.

Dataset	TimeSformer	Swin-T	MobileNetV3	X3D
Bus-violence	0.6536	0.8300	0.8564	0.8907
CamNuVem	0.7190	0.8718	0.8841	0.8426
Crowd Violence	0.6560	0.8760	0.9160	0.9200
Hockey Fight	0.8259	0.9711	0.9612	0.9602
Movie Fight	0.7487	0.9385	0.9385	0.8974
RLVS	0.7390	0.9695	0.9570	0.9665
RWF-2000	0.6930	0.8400	0.8595	0.8805
SCFD	0.7100	0.6933	0.7200	0.8667
VioPeru	0.6286	0.8571	0.8482	0.8357
Kranok-NV (Unseen)	0.7891	0.7530	0.7778	0.8600

Table 5: Dataset-wise inference performance of the X3D dual-branch model using concatenation fusion.

Dataset	Accuracy	Class	Precision	Recall	F1-Score
Bus-violence	(0.8907±82)	Non-Violent	(0.8988±201)	(0.8814±270)	(0.8896±93)
		Violent	(0.8843±209)	(0.9000±237)	(0.8917±80)
CamNuVem	(0.8426±205)	Non-Violent	(0.8672±193)	(0.8824±119)	(0.8439±119)
		Violent	(0.8298±172)	(0.8722±237)	(0.8502±112)
Crowd Violence	(0.9200±200)	Non-Violent	(0.8965±153)	(0.9680±335)	(0.9306±181)
		Violent	(0.9663±350)	(0.8880±179)	(0.9251±176)
Hockey Fight	(0.9602±70)	Non-Violent	(0.9407±98)	(0.9820±110)	(0.9609±69)
		Violent	(0.9815±110)	(0.9386±108)	(0.9595±72)
Movie Fight	(0.8974±363)	Non-Violent	(0.8529±497)	(0.9579±440)	(0.9013±338)
		Violent	(0.9567±445)	(0.8400±652)	(0.8928±400)
RLVS	(0.9665±78)	Non-Violent	(0.9682±126)	(0.9650±197)	(0.9664±82)
		Violent	(0.9655±185)	(0.9680±135)	(0.9666±75)
RWF-2000	(0.8805±118)	Non-Violent	(0.9178±143)	(0.8360±182)	(0.8749±127)
		Violent	(0.8496±145)	(0.9250±141)	(0.8856±111)
SCFD	(0.8667±118)	Non-Violent	(0.8452±301)	(0.9000±300)	(0.8710±110)
		Violent	(0.8944±285)	(0.8333±408)	(0.8618±141)
VioPeru	(0.8357±343)	Non-Violent	(0.7741±338)	(0.9500±407)	(0.8526±301)
		Violent	(0.9368±473)	(0.7214±530)	(0.8140±415)
Kranok-NV (Unseen)	(0.8600±615)	Non-Violent	(0.8924±350)	(0.9417±461)	(0.9162±378)
		Violent	(0.6763±2006)	(0.5010±1598)	(0.5715±1693)

Table 6: Comparison of inference efficiency across dual-branch models with concatenation fusion.

Model	Latency	FPS	Throughput	GFLOPs	Params (M)	Memory (MB)
X3D	(13.1324±1.1759)	(76.6147±6.5253)	6.3360	7.2507	16.5719	153.7818
MobileNetV3	(19.9877±5367)	(50.0308±1.1923)	5.3493	4.3156	14.9601	109.0718
Swin-T	(22.0054±1.3137)	(45.4435±2.9917)	8.1736	126.2611	97.9793	624.2310
TimeSformer	(15.9853±0.0327)	(62.5575±1278)	8.5594	198.5798	132.4943	626.6909

as ambient non-violent motion, leading to a higher rate of false negatives. Nevertheless, the results validate that the model successfully learned generalized motion and spatial patterns rather than overfitting to specific training backgrounds.

The analysis of the remaining performance variations highlights specific environmental boundaries. Within the VioPeru dataset, a notable difference between the recall of non-violent (0.9500) and violent (0.7214) events was due to severe street-level occlusions and background clutter. Similarly, the performance on the SCFD dataset (86.67%) remained sta-

ble despite the challenges of street scenes under low-light conditions. Collectively, these dataset-wise metrics confirm that the asymmetric dual-branch architecture, using the X3D backbone and concatenation fusion, provides a generalized representation of violence that remains effective across both seen and unseen domains.

Table 7: Detailed Comparison of Accuracy between the Proposed Dual-Branch X3D and Multiple SOTA Benchmarks.

Dataset	Ours (Avg)	SOTA Acc.	Method / Model	Reference
1. Bus-violence	0.8907	0.8000	ResNet-3D	[28]
		0.7597	I3D-RGB	[26]
		0.6300	ResNet50 + MIL	[29]
2. RLVS	0.9665	0.9610	CNN	[30]
		0.9522	DenseNet-121 + LSTM	[31]
		0.9505	CUE-Net (Pose + Motion)	[32]
3. RWF-2000	0.8805	0.9525	InternVL3 + labeling (Short-window sliding)	[33]
		0.9230	3D CNN	[34]
		0.8940	IDG-ViolenceNet (Graph + 3D-CNN)	[35]
4. Hockey Fight	0.9602	0.9966	BGRU-MTRA (Spatiotemporal)	[36]
		0.9750	IDG-ViolenceNet (3D-CNN)	[35]
		0.9650	HSTNet(Hybrid SNN-Transformer)	[37]
5. Movie Fight	0.8974	1.0000	CNN	[30]
		0.9950	IDG-ViolenceNet (3D-CNN)	[35]
		0.9800	HSTNet(Hybrid SNN-Transformer)	[37]
6. Crowd Violence	0.9200	0.9864	CNN-BiLSTM	[22]
		0.9400	GMFlow + CBAM (Two-Stage)	[38]
7. VioPeru	0.8357	0.9000	Pretrained CNN	[25]
		0.8923	Dual Branch VideoMamba	[39]
8. SCFD	0.8667	0.9610	CNN	[30]
		0.9130	Subgroup Analysis (Localization)	[40]
9. CamNuVem	0.8426	0.9487	3D-CNN	[41]
10. Kranok-NV	0.8600	0.9810	Pose Estimation + CNN	[42]

5.5 Evaluation of Inference Efficiency and Real-Time Performance

The computational efficiency of the dual-branch framework was evaluated to verify its suitability for real-time surveillance applications. Metrics, including inference latency, frames per second (FPS), throughput, and computational complexity (GFLOPs), were recorded across four distinct backbone architectures. All benchmarks were performed using the concatenation fusion strategy to ensure a consistent comparison of the integrated systems. To provide a standardized baseline for hardware performance, all inference metrics were measured on a single NVIDIA A100 GPU with 40GB of VRAM. The results of this efficiency analysis are summarized in Table 6.

The X3D-based dual-branch model demonstrated the highest operational efficiency among the evaluated architectures. A mean latency of 13.13 ms and a processing rate of 76.61 FPS were achieved, significantly exceeding the standard real-time threshold of 30 FPS. While the MobileNetV3 backbone yielded lower memory consumption (109.07 MB) and fewer GFLOPs (4.31), its inference speed was observed to be slower than that of the X3D model. This performance discrepancy suggests that the X3D architecture is better optimized for the violence detection task in terms of accuracy and efficiency.

Moreover, the Transformer-based architectures, Swin-T and TimeSformer, exhibited substantial computational overhead with over 100 GFLOPs and

memory consumption of over 600 MB. These high requirements are attributed to self-attention mechanisms, which heavily demand hardware resources. The superior throughput and low latency, combined with a low parameter count (16.57 M), confirm that the proposed framework is capable of processing videos with minimal delay.

5.6 Performance Comparison with State-of-the-Art

As shown in Table 7, the proposed dual-branch X3D framework demonstrates competitive results across ten distinct datasets. Notably, the model outperforms the existing benchmark for the Bus-violence dataset, achieving an accuracy of 0.8907 compared to the 0.7597 reported by the I3D-RGB baseline [26]. Additionally, on the *RLVS* dataset, it surpassed the current state-of-the-art accuracy of 0.9610, reaching 0.9665.

While some datasets like *Hockey* and *Movie Fight* exhibit SOTA accuracies near 100%, these often rely on staged or high-resolution footage. The model maintains a more balanced and robust generalization profile across both staged and real-world surveillance environments. Furthermore, while the SOTA for *Kranok-NV* (98.1%) [42] deploys computationally intensive human pose estimation, the proposed model achieves a respectable 86.00% (with a peak of 91.61%) using an end-to-end video classification approach. This trade-off emphasizes the model's scalability and suitability for real-time monitoring in di-

verse conditions where individual skeletal tracking is often unreliable.

To ensure a rigorous contextualization of these comparisons, it is important to highlight the variations in evaluation protocols and auxiliary supervision across prior works. Many top-performing models utilize distinct evaluation boundaries, such as localized, non-cross-validated random splits, or incorporate heavy auxiliary modules including pre-computed dense optical flow maps, pose estimations, and bounding-box annotations. In contrast, the proposed framework is designed strictly around a lightweight end-to-end pipeline using raw pixel difference maps to satisfy real-time application constraints. Accounting for these structural and protocol differences highlights the efficiency and background-robust generalization achieved by our architecture without relying on heavy external supervision layers.

6. DISCUSSION

The experimental results validate that the asymmetric integration of X3D and ResNet-18 provides a robust solution for violence detection in varying environments. The significant performance gains observed in the Bus-violence dataset are particularly noteworthy, as this environment contains complex variations, such as varying lighting conditions and frequent occlusions from bus seats and passengers.

The technical efficiency of the framework is largely driven by the architectural expansion of the X3D backbone. Unlike the fixed kernels used in conventional models, the X3D progressively scales across temporal and spatial axes, allowing it to capture motion and spatial cues with a lower parameter count. Furthermore, the use of Kinetics-400 pre-trained weights provided a foundational understanding of human actions, which, when paired with the motion branch trained from scratch, allowed the model to converge on generalized violence representations rather than overfitting to a specific environment. Overall, the ability of the proposed architecture to remain highly competitive across ten distinct datasets, ranging from high-resolution movies to low-quality surveillance videos, demonstrates its robust generalization capability. The proposed asymmetric dual-branch with the pre-trained X3D backbone and lightweight ResNet-18 offers a balance between accuracy and deployment feasibility, marking a step toward real-time violence detection in uncontrolled environments.

7. LIMITATIONS OF THE PROPOSED FRAMEWORK

While the proposed framework demonstrates robust generalization, several limitations remain. First, relying on lightweight frame differencing rather than dense optical flow makes the motion branch suscep-

tible to noise from extreme camera jitter or sudden illumination flashes. Second, zero-shot evaluation (e.g., on *Kranok-NV*) reveals a vulnerability to domain shifts in action dynamics; the model struggles to detect isolated, single-subject violence that lacks the multi-person interactive cues prevalent in the training data. Third, as a video-level classifier, the pipeline cannot output bounding-box coordinates to localize altercations in crowded scenes. Finally, synchronized dual-stream processing may exceed the memory bandwidth of ultra-low-power IoT edge devices. Future work will explore unsupervised domain adaptation, lightweight spatial attention for localization, and network quantization for extreme edge deployment.

8. CONCLUSION

This study presents an asymmetric dual-branch framework designed for robust violence detection across diverse surveillance contexts. By using a multi-source dataset comprising 8,398 video clips and validating the model against ten distinct benchmarks, the study demonstrated that the integration of appearance and motion information enhances the model’s ability to generalize to real-world scenarios. The findings indicate that the proposed architecture not only maintains high generalizability but also achieves accuracy in specialized environments, such as public transport.

The primary contribution of this work is the establishment of a new state-of-the-art benchmark for the Bus-violence dataset, achieving an accuracy of 0.8907, and the *RLVS* dataset, reaching 0.9665. Furthermore, the model’s performance on the imbalanced *Kranok-NV* dataset, specifically its high recall for non-violent actions (0.96), underscores its practical utility in reducing false alarms in automated security systems. By providing a “pose-free” alternative to traditional skeleton-based methods, this framework ensures reliable performance in crowded or low-resolution conditions where auxiliary data extraction is often infeasible.

Future work will focus on improving the real-time deployment capability of the proposed framework through model compression and computational optimization techniques. Specifically, lightweight inference strategies such as pruning, quantization, and efficient backbone adaptation will be investigated to reduce computational overhead while maintaining recognition performance. In addition, future studies will evaluate the trade-off between detection accuracy and processing efficiency under practical surveillance constraints. The optimized framework will be deployed and tested on edge-computing platforms to assess real-time inference capability. Furthermore, temporal attention mechanisms will be explored to enhance the motion branch sensitivity for distinguishing violent activities from visually similar high-intensity

but non-violent movements. Ultimately, this work establishes a scalable foundation for efficient and autonomous violence detection in uncontrolled real-world environments.

ACKNOWLEDGEMENT

This research project was partially supported by Scholarships for Ph.D. Student from Mahidol University and was partially supported by Faculty of Information and Communication Technology, Mahidol University.

AUTHOR CONTRIBUTIONS

Conceptualization, S.A., W.K., Y.A. and C.L.; methodology, S.A. W.K. and C.L.; software, S.A.; validation, S.A., and Y.A.; formal analysis, S.A.; investigation, S.A. and W.K.; data curation, S.A. and C.L.; writing—original draft preparation, S.A.; writing—review and editing, S.A., W.K. and Y.A.; visualization, Y.A.; supervision, W.K.; funding acquisition, W.K. All authors have read and agreed to the published version of the manuscript.

References

- [1] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1–3, pp. 185–203, 1981.
- [2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, vol. 1, pp. 886–893, 2005.
- [3] K. Yun, H. Jeong, K. M. Yi, S. W. Kim and J. Y. Choi, "Motion Interaction Field for Accident Detection in Traffic Surveillance Video," *2014 22nd International Conference on Pattern Recognition*, Stockholm, Sweden, pp. 3062–3067, 2014.
- [4] P. Khanarsa and S. Kitsiranuwat, "Deep learning-based ensemble approach for conventional pap smear image classification," *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 18, no. 1, pp. 101–111, 2024.
- [5] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Lecce, Italy, pp. 1–6, 2017.
- [6] D. Sriveni and R. Loganathan, "Multi-constraints active learning assisted deep-ensemble spatio-textural feature learning model for violence detection in surveillance dataset," *Connection Science*, vol. 37, no. 1, p. 2544539, 2025.
- [7] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the 28th International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 1, pp. 568–576, 2014.
- [8] C. Feichtenhofer, A. Pinz and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 1933–1941, 2016.
- [9] C. Feichtenhofer, H. Fan, J. Malik and K. He, "SlowFast Networks for Video Recognition," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 6201–6210, 2019.
- [10] S. Mekruksavanich and A. Jitpattanakul, "Fall-NeXt: A deep residual model based on multi-branch aggregation for sensor-based fall detection," *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 16, no. 4, pp. 352–364, 2022.
- [11] N. Alabid, "Interpretation of spatial relationships by objects tracking in a complex streaming video," *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 15, no. 2, pp. 245–257, 2021.
- [12] S. Thu Ya Aung, W. Kusakunniran and Y. Kwang Hooi, "Leveraging Fusion Methods of Human Pose and Motion Dynamics for Accurate Violence Detection in Video Surveillance," in *IEEE Access*, vol. 13, pp. 168116–168125, 2025.
- [13] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, 2016.
- [14] C. Feichtenhofer, "X3D: Expanding Architectures for Efficient Video Recognition," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 200–210, 2020.
- [15] G. Bertasius, H. Wang and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 813–824, 2021.
- [16] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 9992–10002, 2021.
- [17] A. Howard *et al.*, "Searching for MobileNetV3," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 1314–1324, 2019.
- [18] W. Kay *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.
- [19] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li

- and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, pp. 248-255, 2009.
- [20] R. Goyal *et al.*, "The "Something Something" Video Database for Learning and Evaluating Visual Common Sense," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp. 5843-5851, 2017.
- [21] M. M. Soliman, M. H. Kamal, M. A. El-Massih Nashed, Y. M. Mostafa, B. S. Chawky and D. Khattab, "Violence Recognition from Videos using Deep Learning Techniques," *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, Cairo, Egypt, pp. 80-85, 2019.
- [22] T. Hassner, Y. Itcher and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Providence, RI, USA, pp. 1-6, 2012.
- [23] M. Cheng, K. Cai and M. Li, "RWF-2000: An Open Large Scale Video Database for Violence Detection," *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, pp. 4183-4190, 2021.
- [24] Ş. Aktı, G. A. Tataroğlu and H. K. Ekenel, "Vision-based Fight Detection from Surveillance Cameras," *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Istanbul, Turkey, pp. 1-6, 2019.
- [25] H. A. Huillcen Baca, F. L. Palomino Valdivia and J. C. Gutierrez Caceres, "Efficient human violence recognition for surveillance in real time," *Sensors*, vol. 24, no. 2, p. 668, 2024.
- [26] L. Ciampi *et al.*, "Bus violence: An open benchmark for video violence detection on public transport," *Sensors*, vol. 22, no. 21, p. 8345, 2022.
- [27] M. B. de Paula, D. H. Salvadeo and D. M. de Araujo, "CamNuVem: A robbery dataset for video anomaly detection," *Sensors*, vol. 22, no. 24, p. 10016, 2022.
- [28] M. Kulkarni and R. Chakraborty, "Violent Activity Detection on Public Transportation using Surveillance Footage," *2023 16th International Conference on Developments in eSystems Engineering (DeSE)*, Istanbul, Turkiye, pp. 737-742, 2023.
- [29] L. Ciampi, C. Santiago, F. Falchi, C. Gennaro, and G. Amato, "In the wild video violence detection: An unsupervised domain adaptation approach," *SN Computer Science*, vol. 5, no. 7, p. 834, 2024.
- [30] A. Kulkarni, S. Modi, O. Vyawahare, A. Dubbe-war and N. L. Pariyal, "Real-time quarrel detection in surveillance videos," *International Journal of Novel Research and Development (IJNRD)*, vol. 10, no. 12, Dec. 2025.
- [31] A. Akbaş, H. Üçgün and A. A. Ali, "A Hybrid Deep Learning Approach for Violence Detection in Videos," *2025 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Bursa, Turkiye, pp. 1-6, 2025.
- [32] D. C. Senadeera, X. Yang, D. Kollias and G. Slabaugh, "CUE-Net: Violence Detection Video Analytics with Spatial Cropping, Enhanced Uni-formerV2 and Modified Efficient Additive Attention," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, pp. 4888-4897, 2024.
- [33] S. Jung, T. Song, Y. Lee and S. Lee, "Short-window sliding learning for real-time violence detection via LLM-based auto-labeling," *arXiv preprint arXiv:2511.10866*, 2025.
- [34] J. Chen, "Method and analysis of violent behavior recognition based on multimodal information fusion," *Highlights in Science, Engineering and Technology*, vol. 161, pp. 34-42, 2026.
- [35] H. Huang and Q. Jiang, "IDG-ViolenceNet: A video violence detection model integrating identity-aware graphs and 3D-CNN," *Sensors*, vol. 25, no. 20, p. 6272, 2025.
- [36] A. Pandey and P. Kumar, "BGRU-MTRA: Bi-linear GRU networks with multi-path temporal residual attention for suspicious activity recognition," *Neural Computing and Applications*, vol. 37, no. 1, pp. 185-212, 2025.
- [37] F. Meng, L. Zou, J. Lin and Z. Liu, "HSTNet: Violent action detection," *Applied Sciences*, vol. 16, no. 4, p. 1825, 2026.
- [38] M. Mahmoud, B. Yagoub, M. F. Senussi, M. Abdalla, M. S. Kasem and H.-S. Kang, "Two-stage video violence detection framework using GM-Flow and CBAM-enhanced ResNet3D," *Mathematics*, vol. 13, no. 8, p. 1226, 2025.
- [39] D. C. Senadeera, X. Yang, S. Li, M. Awais, D. Kollias, and G. Slabaugh, "Dual branch Video-Mamba with gated class token fusion for violence detection," *arXiv preprint arXiv:2506.03162*, 2025.
- [40] E. Veltmeijer, M. Franken and C. Gerritsen, "Real-time violence detection and localization through subgroup analysis," *Multimedia Tools Appl.*, vol. 84, no. 7, pp. 3793-3807, 2025.
- [41] A. Kavathia and S. Sayer, "Optimizing violence detection in video classification accuracy through 3D convolutional neural networks," *arXiv preprint arXiv:2411.01348*, 2024.
- [42] K. B. Kwan-Loo, J. C. Ortíz-Bayliss, S. E. Conant-Pablos, H. Terashima-Marín and P. Rad, "Detection of Violent Behavior Using Neural Networks and Pose Estimation," in *IEEE Access*, vol. 10, pp. 86339-86352, 2022.



Sai Thu Ya Aung received his B.C.Tech (Hons) degree in Computer Technology from the University of Computer Studies, Yangon, Myanmar, and an M.Sc. degree in Computer Science from the Faculty of ICT, Mahidol University, Thailand. He is currently pursuing his Ph.D. degree in Computer Science at the Faculty of Information and Communication Technology, Mahidol University, Thailand. His research

interests focus on developing and applying machine learning, deep learning, and image and video analysis to solve problems in a variety of domains, including smart surveillance, smart agriculture, and bioinformatics.



Yu Nandar Aung received the B.E. degree in Information Technology from Technological University (Hmawbi), Yangon, Myanmar, the M.E. degree in Information Technology from Yangon Technological University, and the M.Sc. degree in Cyber Security and Information Assurance from the Faculty of Information and Communication Technology, Mahidol University, Thailand, where she is currently pursuing the Ph.D. degree

in computer science. Her research interests include embedded systems and IoT smart home security, blockchain technology, and efficient, lightweight machine learning models for computer vision.



Worapan Kusakunniran received the B.Eng. degree in computer engineering from the University of New South Wales (UNSW), Sydney, NSW, Australia, in 2008, and the Ph.D. degree in computer science and engineering from UNSW, in cooperation with Neville Roach Laboratory, National ICT Australia, Sydney, NSW, Australia, in 2013. He is currently an Associate Professor at the Faculty of Information and Communication Tech-

nology, Mahidol University, Nakhon Pathom, Thailand. He is the author of several papers in top international conferences and journals. He was a program committee member of many international conferences and workshops. He was also a reviewer of several international conferences and journals, such as ICPR, ICIP, PR, TIP, and TIFS. His research interests include biometrics, pattern recognition, medical image processing, computer vision, multimedia, and machine learning. He was the recipient of the ICPR Best Biometric Student Paper Award in 2010 and has won several national and international innovation contests.



Chattapatr Leeraha received the Bachelor of Science degree in Physics from Mahidol University International College, and M.Sc. degree in computer science from the Faculty of Information and Communication Technology, Mahidol University, Thailand, where he is currently pursuing the Ph.D. degree in computer science. His research interests include developing and applying machine learning to solve problems in a

variety of domains.