mathematics

MDPI

Article

Two-Stage Video Violence Detection Framework Using GMFlow and CBAM-Enhanced ResNet3D

Mohamed Mahmoud [1,2], Bilel Yagoub [1], Mostafa Farouk Senussi [1,2], Mahmoud Abdalla [1], Mahmoud Salaheldin Kasem [1,3] and Hyun-Soo Kang [1,*]

[1] Department of Information and Communication Engineering, School of Electrical and Computer Engineering, Chungbuk National University, Cheongju-si 28644, Republic of Korea
[2] Information Technology Department, Faculty of Computers and Information, Assiut University, Assiut 71526, Egypt
[3] Multimedia Department, Faculty of Computers and Information, Assiut University, Assiut 71526, Egypt
* Correspondence: hskang@cbnu.ac.kr

Abstract: Video violence detection has gained significant attention in recent years due to its applications in surveillance and security. This paper proposes a two-stage framework for detecting violent actions in video sequences. The first stage leverages GMFlow, a pre-trained optical flow network, to capture the temporal motion between consecutive frames, effectively encoding motion dynamics. In the second stage, we integrate these optical flow images with RGB frames and feed them into a CBAM-enhanced ResNet3D network to capture complementary spatiotemporal features. The attention mechanism provided by CBAM enables the network to focus on the most relevant regions in the frames, improving the detection of violent actions. We evaluate the proposed framework on three widely used datasets: Hockey Fight, Crowd Violence, and UBI-Fight. Our experimental results demonstrate superior performance compared to several state-of-the-art methods, achieving AUC scores of 0.963 on UBI-Fight and accuracies of 97.5% and 94.0% on Hockey Fight and Crowd Violence, respectively. The proposed approach effectively combines GMFlow-generated optical flow with deep 3D convolutional networks, providing robust and efficient detection of violence in videos.

Keywords: video violence detection; GMFlow; optical flow; CBAM (convolutional block attention module); ResNet3D; anomaly detection

MSC: 68T07

Academic Editors: Chengwei Pan, Hongyi Li, Di Zhao and Marjan Mernik

Received: 20 January 2025
Revised: 15 March 2025
Accepted: 7 April 2025
Published: 8 April 2025

Citation: Mahmoud, M.; Yagoub, B.; Senussi, M.F.; Abdalla, M.; Kasem, M.S.; Kang, H.-S. Two-Stage Video Violence Detection Framework Using GMFlow and CBAM-Enhanced ResNet3D. *Mathematics* **2025**, *13*, 1226. https://doi.org/10.3390/math13081226

1. Introduction

Violence detection is a critical application in video analysis, offering significant benefits in various domains. It is a powerful tool for filtering sensitive media content, protecting users from exposure to unwanted material, and supporting law enforcement in forensic investigations. Beyond their role in public safety, violence detection systems can prevent the dissemination of violent content on social networks, forums, and educational platforms. These systems are equally valuable for maintaining safe environments by restricting violent material in sensitive settings such as workplaces, schools, and public spaces. With the increasing prevalence of video surveillance systems, automated violence detection has become indispensable for crime prevention, crowd management, and enhancing the effectiveness of intelligent security systems. This study aims to develop a robust and scalable deep learning framework that effectively integrates motion and spatial features

for accurate and generalizable video violence detection. Beyond its strong classification performance, our framework has practical implications for real-time security and surveillance systems. The GMFlow-based optical flow extraction mechanism operates efficiently in sequential video frames, while CBAM-enhanced ResNet3D processes motion and spatial cues in parallel, making the model well-suited for real-time violence detection in public spaces, transportation hubs, and law enforcement applications. Using weak supervision, the approach reduces the need for detailed frame-level annotations, facilitating easier deployment in large-scale monitoring systems. Future optimizations, such as model pruning and quantization, could further enhance the inference speed, making the system even more viable for real-world applications.

Deep learning has revolutionized video analysis, particularly in violence detection, by enabling models to automatically learn hierarchical and discriminative representations from raw data, surpassing traditional handcrafted feature-based methods that often suffer from limited generalizability [1,2]. Recent advancements in convolutional and recurrent neural networks have significantly enhanced the ability to extract robust spatiotemporal features and model temporal sequences, while attention mechanisms further improve the focus on critical spatial and temporal regions [3–6]. These innovations have set new benchmarks in violence detection and classification, showcasing the effectiveness of deep learning in handling complex motion dynamics in diverse scenarios. Beyond violence detection, deep learning has achieved remarkable success in various domains, including image inpainting, especially object removal [7,8], object detection and recognition [9–11], and security applications [12], further demonstrating its versatility and transformative impact in computer vision and artificial intelligence research.

Research in video violence detection spans three primary paradigms: supervised, unsupervised, and weakly supervised learning. Supervised methods dominate the field, leveraging frame-level annotations to achieve high accuracy, but are often constrained by their reliance on exhaustive manual labeling and their limited generalizability to unseen scenarios. Unsupervised methods, which focus on anomaly detection by assuming that violent events deviate from normal patterns, eliminate the need for explicit labels but struggle to capture the nuanced characteristics of violence. Weakly supervised approaches provide a scalable alternative, utilizing video-level annotations to balance performance and resource efficiency. By reducing dependency on detailed labels while maintaining competitive accuracy, weakly supervised learning is particularly suited for large-scale applications, addressing the challenges posed by limited annotations and the growing volume of video data. Nevertheless, video violence detection remains a complex task, further complicated by the subjective nature of violence, variations in environmental conditions, and the ever-growing volume of video data being generated.

In this paper, we propose a two-stage network architecture for video violence detection that seamlessly integrates motion and appearance cues to achieve robust and balanced performance. The first stage utilizes the GMFlow pre-trained network [13] to generate optical flow images, effectively capturing the temporal dynamics of motion within video sequences. Optical flow serves as a detailed representation of movement patterns, providing crucial insights into temporal behavior for analyzing violent actions. Recognizing that motion alone is insufficient to detect violence accurately, the second stage incorporates RGB frames alongside optical flow images to leverage both spatial and temporal information. Specifically, the first frame of every two-frame sequence used for optical flow computation is selected, ensuring an efficient yet comprehensive spatial representation. These inputs are then processed by a ResNet3D-based architecture [14], enhanced with Convolutional Block Attention Modules (CBAM) [15] integrated into the residual blocks. CBAM improves the network's ability to focus on critical spatial regions and channel features, enabling

more effective extraction of spatiotemporal patterns essential for distinguishing violent actions from non-violent ones. The overall design of this two-stage framework is depicted in Figure 1, providing an overview of its components and workflow. This integration of motion and spatial cues addresses key challenges in violence detection, ensuring a scalable and effective solution for real-world applications.

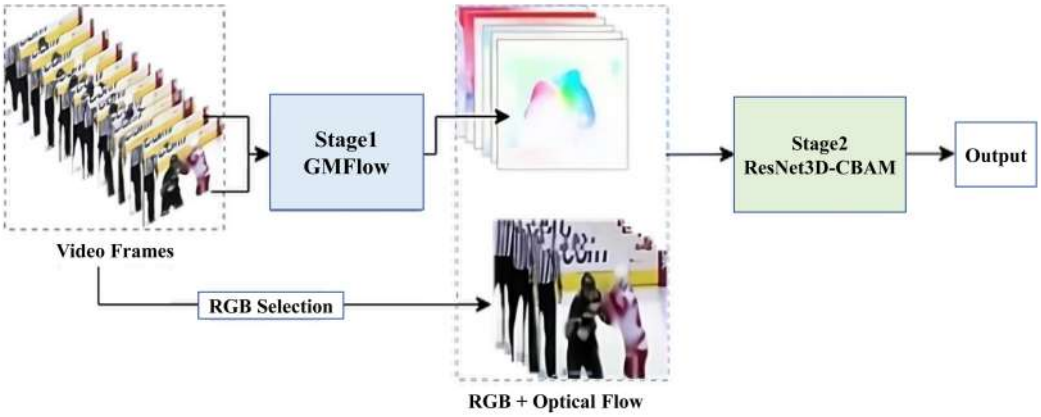

Figure 1. Illustration of our two-stage video violence detection framework. Stage 1 extracts motion cues using GMFlow optical flow, while Stage 2 integrates these cues with RGB frames in a CBAM-enhanced ResNet3D network for improved spatiotemporal feature extraction.

We evaluate the effectiveness of the proposed two-stage framework on three benchmark datasets: Hockey [16], Crowd [17], and UBI-Fights [18]. These datasets encompass diverse scenarios and challenges, providing a robust platform to assess the generalization and adaptability of our method. Experimental results reveal that the framework excels in the weakly supervised setting, showcasing its ability to capture complex violence patterns while minimizing reliance on detailed annotations. This highlights the potential of the proposed approach for scalable and practical applications in real-world violence detection tasks.

The key contributions of this paper are summarized as follows:

1. A two-stage deep learning framework for video violence detection, integrating GMFlow-based motion encoding with a CBAM-enhanced ResNet3D network to capture both temporal and spatial features effectively.
2. Robust motion representation using GMFlow, which generates dense optical flow features, improving the model's ability to understand movement patterns critically for detecting violent actions.
3. Enhanced spatiotemporal feature extraction with CBAM allows the network to selectively focus on the most relevant regions, leading to improved discrimination between violent and non-violent activities.
4. Extensive evaluation on three diverse benchmark datasets (Hockey Fight, Crowd Violence, and UBI-Fight), demonstrating the generalizability of our method across different violence scenarios and environmental conditions.
5. Competitive state-of-the-art performance, achieving high accuracy and AUC scores while reducing reliance on detailed frame-level annotations, making the framework suitable for scalable real-world applications.

The remainder of this paper is organized as follows: Section 2 reviews related work, including existing approaches in supervised, unsupervised, and weakly supervised video violence detection. Section 3 details the proposed two-stage framework, including the optical flow computation and attention-enhanced ResNet3D architecture. Section 4 describes

the experimental setup, datasets, evaluation metrics, and results and analysis. Finally, Section 5 concludes the paper and discusses future research directions.

2. Related Work

Video violence detection has gained important attention because of its critical applications in public safety and surveillance systems. The advancements in deep learning have helped the modeling of complex spatiotemporal patterns and motion dynamics in video data, significantly improving the field. Supervised methods, which are based on frame-level annotations, have shown remarkable accuracy by leveraging architectures such as 3D Convolutional Neural Networks (3D CNNs) [19] and hybrid models integrating Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks [20]. These methods are very effective at capturing complex motion patterns and temporal dependencies, which are essential for distinguishing violent from non-violent actions. However, their dependence on frame-level annotated datasets, which are often scarce, limits their scalability and applicability in real-world scenarios. To address these challenges, weakly supervised approaches, which utilize video-level annotations, have gained importance by reducing the annotation load while maintaining competitive performance. This section reviews state-of-the-art methods in video violence detection, with a particular focus on weakly supervised and unsupervised learning approaches.

2.1. Unsupervised Methods

Unsupervised methods are designed to detect anomalies without relying on labeled data, either through fully unsupervised learning or semi-supervised learning, also known as one-class classification. In fully unsupervised frameworks, the goal is to separate normal and abnormal video segments by employing techniques such as clustering, pseudo-labeling, and anomaly scoring, enabling their application in scenarios where annotations are unavailable.

For instance, DyAnNet [21] and C2FPL [22] generate pseudo-labels to distinguish between normal and abnormal segments in videos. DyAnNet employs isolation tree-based clustering to create pseudo anomaly and dynamicity scores from both RGB and optical flow streams, refining these scores with a cross-branch feed-forward network based on I3D. C2FPL [22], on the other hand, proposes a two-stage pseudo-label generation framework using hierarchical clustering and statistical testing to generate segment-level labels for training an anomaly detector. Hu et al. [23] introduce Temporal Masked Auto-Encoding (TMAE), a method that masks temporal patches in spatial-temporal cubes and uses a vision transformer to predict them. Anomalies are detected when these predictions are inaccurate, and TMAE is further enhanced by applying it to optical flow for better performance. Additionally, Tao et al. [24] present FRD-UVAD, a model for unsupervised video anomaly detection that employs cascade cross-attention transformers and a disruption process to refine features and improve pseudo-label generation.

Another common approach in unsupervised methods is reconstruction-based anomaly detection, which utilizes models like autoencoders to learn normal patterns and reconstruct video frames. Anomalies are detected when the reconstruction error overrides a predefined threshold. Luo et al. [25] propose a ConvLSTM-AE framework that combines Convolutional Neural Networks (CNNs) for appearance encoding with ConvLSTM for motion encoding, capturing both spatial and motion regularities of normal events. Gong et al. [26] introduce MemAE, a memory-augmented autoencoder that strengthens anomaly detection by using a memory module to retrieve relevant normal data for reconstruction, amplifying the reconstruction error for anomalies. Wang et al. [27] present DF-ConvLSTM-VAE, a model designed to address challenges related to unbalanced data and time-series issues, thus

enhancing video anomaly detection. However, these methods are prone to overfitting, particularly when identifying subtle violent behaviors.

Prediction-based methods, another class of unsupervised approaches, aim to predict future frames based on past video sequences, detecting anomalies when predicted frames deviate significantly from the ground truth. Liu et al. [28] propose a video prediction-based method that compares predicted future frames with actual frames, incorporating both spatial and temporal (optical flow) constraints to improve normal event prediction and anomaly detection. Li et al. [29] introduce a context-based anomaly detection method using a generative adversarial network (GAN), where a two-branch generator network predicts future frames by considering both preceding and succeeding video frames. The final anomaly score is derived from these predictions. While unsupervised methods eliminate the need for labeled data, they often struggle with distinguishing subtle violent actions due to their reliance on anomaly-based assumptions. Weakly supervised learning offers a middle ground by leveraging video-level labels, reducing annotation effort while retaining a degree of supervision to improve discrimination between violent and non-violent activities.

2.2. Weakly Supervised Methods

Unlike unsupervised methods, which attempt to detect anomalies without explicit labels, weakly supervised approaches utilize video-level annotations to provide indirect supervision. This allows the model to learn meaningful patterns without requiring frame-level labels, striking a balance between performance and annotation efficiency. By focusing on coarse labels rather than detailed per-frame supervision, weakly supervised learning enables scalable video analysis while still achieving competitive accuracy. This characteristic makes weak supervision particularly suitable for large-scale applications, where annotating every frame of a video is impractical. As a result, these methods are gaining popularity due to their ability to reduce annotation effort while maintaining strong performance in violence detection tasks.

A prominent strategy in weakly supervised learning is Multiple Instance Learning (MIL) [30], where a video is treated as a collection of instances and instance-level labels are inferred based on video-level annotations. MIL-based frameworks prioritize the top-K video segments most likely to contain violent content, optimizing the learning of discriminative features. Sultani et al. [31] proposed a weakly supervised deep multiple-instance ranking framework for anomaly detection, utilizing video-level labels in a MIL setup. Their method employs sparsity and temporal smoothness constraints for improved anomaly localization while introducing a large-scale dataset featuring 128 h of surveillance videos across 13 anomaly categories.

Generative models have also been explored in weakly supervised settings. Hasan et al. [32] proposed a model for detecting regular motion patterns in long video sequences, addressing challenges such as ambiguous and cluttered scenes. Their approach uses two autoencoder-based models: one leveraging handcrafted spatiotemporal features and the other employing a fully convolutional autoencoder for end-to-end learning. Wang et al. [33] introduced S^2-VAE, which combines a shallow Stacked Fully Connected Variational AutoEncoder (S F-VAE) for modeling data distributions with a deep Skip Convolutional VAE (S C-VAE), integrating CNNs, VAEs, and skip connections to enhance anomaly detection.

Recent advancements in attention mechanisms and transformer-based architectures have further enhanced weakly supervised methods. Lee et al. [6] proposed a three-stage architecture combining feature extraction via CNNs, temporal sequence modeling with LSTMs and attention mechanisms, and long-range temporal analysis through a transformer encoder. Deshpande et al. [34] utilized transformer-based Videoswin features along with an attention layer integrating dilated convolutions and self-attention to capture temporal de-

pendencies for generating frame-level anomaly scores from video-level labels. Jin et al. [35] extended transformer-based approaches to aerial video anomaly detection, proposing Anomaly Detection with Transformers (ANDTs) for UAV-captured aerial videos. Their method treats consecutive frames as tubelets and applies a transformer encoder for feature extraction and a decoder for frame prediction, detecting anomalies with unpredictable temporal dynamics.

To improve temporal feature modeling, Degardin et al. [36] introduced an iterative self-supervised learning framework for anomaly detection. This framework employs two experts to iteratively expand the training dataset by incorporating confidently classified instances. Their approach integrates a Bayesian framework for filtering instances, a novel loss function for optimizing score distribution, and a decision fusion scheme using decision trees. Hwang et al. [5] explored real-time violent crime detection by reconstructing frames into smaller image sizes and applying a Convolutional Block Attention Module (CBAM) [15] to a 3D convolutional residual neural network to focus on key spatiotemporal regions. Chen et al. [37] proposed the Glance and Focus Network (GFN), which integrates spatiotemporal information to enhance anomaly localization in surveillance videos.

Finally, methods focusing on robust temporal modeling have emerged to detect subtle anomalies. Tian et al. [38] introduced Robust Temporal Feature Magnitude (RTFM) learning, which employs dilated convolutions and self-attention to capture temporal dependencies. RTFM enhances weakly supervised learning by effectively identifying subtle anomalies within video sequences.

Table 1 provides a comparative analysis of existing supervised, unsupervised, and weakly supervised methods, outlining their strengths, limitations, and how our proposed framework addresses these challenges.

Table 1. Comparison of existing methods and our proposed approach.

Method Type	Strengths	Weaknesses	Our Approach (Advantages)
Supervised Learning	High accuracy; learns fine-grained features.	Requires frame-level annotations, and poor generalization to unseen data.	Reduces annotation effort by using weak supervision while maintaining accuracy.
Unsupervised Learning	No need for labeled data; can detect anomalies.	Struggles with distinguishing subtle violent actions; may misclassify normal motion.	Uses optical flow to enhance motion representation, improving detection of subtle violent cues.
Weakly Supervised Learning (Existing)	Balances annotation effort and accuracy; more scalable.	Still lacks fine-grained temporal localization; performance varies across datasets.	Enhances weak supervision by integrating both motion (GMFlow) and spatial attention (CBAM-enhanced ResNet3D) for better spatiotemporal feature extraction.

Weakly supervised methods and the integration of motion and appearance features offer complementary strategies for video violence detection. While weakly supervised approaches leverage video-level annotations to reduce labeling effort, they often lack accurate temporal and spatial localization. In contrast, integrating motion (e.g., optical flow) and appearance (e.g., RGB frames) enhances spatiotemporal modeling, enabling the detection of fine-grained violent actions.

Multiple-instance learning (MIL) has facilitated scalable video analysis, but its effectiveness can be further improved by incorporating detailed motion and spatial features. By bridging the gap between weak supervision and fine-grained feature extraction, combining these approaches enhances detection accuracy and enables more precise anomaly localization in large-scale datasets. This synergy demonstrates that weakly supervised learning and feature integration can be jointly leveraged to develop robust systems capable of handling large-scale datasets while achieving fine-grained anomaly detection.

2.3. Integration of Motion and Appearance Features

Recent works emphasize the importance of integrating motion and appearance features for robust violence detection. Optical flow captures temporal motion dynamics but often lacks spatial context, making its integration with RGB data crucial. Two-stream networks, which process RGB frames for spatial features and optical flow for temporal dynamics, have emerged as a popular solution. Carreira et al. [39] introduced the Two-Stream Inflated 3D ConvNet (I3D), which extends 2D ConvNet filters into 3D to learn spatiotemporal features from videos. Pre-trained on the large-scale Kinetics dataset, I3D effectively captures both spatial and temporal patterns. Zhang et al. [40] proposed a model integrating YOLO-v3 with FlowNet 2.0 for video object detection, improving detection accuracy through flow-guided partial warp and optical flow compression. Yi Zhu et al. [41] advanced this approach with hidden two-stream CNNs, an end-to-end architecture that extracts motion information directly from raw video inputs, simplifying preprocessing while maintaining competitive performance. Park et al. [4] demonstrated the importance of combining optical flow and RGB data for accurate violence detection, underscoring the critical role of integrating motion and appearance features in real-world scenarios.

3. Approach

This section introduces the proposed two-stage video violence detection framework, designed to effectively integrate motion and spatial features for robust recognition of violent actions. The framework leverages optical flow to capture temporal dynamics and RGB frames to extract spatial details, which are processed using a CBAM-enhanced ResNet3D network. By analyzing complementary spatiotemporal features, the method identifies key patterns indicative of violent actions, such as sudden aggressive movements and chaotic crowd behavior.

A key aspect of our approach is its weakly supervised nature, which enables learning from video-level labels without the need for explicit frame-wise annotations. This is particularly advantageous for large-scale video datasets where obtaining fine-grained annotations is impractical. Our method addresses the inherent challenges of weakly supervised learning, particularly the difficulty of associating specific frames with violent actions. GMFlow provides motion representations that capture dynamic changes, while CBAM-enhanced ResNet3D focuses on spatial features, enhancing the identification of violence-relevant patterns. The model is trained to associate violence with particular frame segments by combining motion cues and spatial-temporal features, effectively bridging the gap between coarse video-level labels and fine-grained frame-wise predictions. During training, the model optimizes its classification by learning which spatiotemporal features contribute most to distinguishing violent from non-violent actions. GMFlow generates motion representations emphasizing dynamic changes, guiding the network to focus on movement-heavy regions. The CBAM module enhances attention to spatially crucial areas, reinforcing the identification of violence-relevant patterns.

Figure 1 illustrates the workflow, showcasing the integration of the framework's components for effective violence detection. In the first stage, GMFlow [13], a state-of-the-art optical flow estimation network, generates dense optical flow images to represent temporal dynamics and movement patterns within video sequences. The framework processes the input video by extracting frames at intervals of n rather than using all frames. For each pair of consecutive frames I and $I + 1$, optical flow is estimated to detect motion dynamics, while the Ith RGB frame is retained for the next stage to serve as appearance cues. In the second stage, the optical flow images are combined with the retained RGB frames to leverage both spatial and temporal information simultaneously. The CBAM-enhanced ResNet3D network processes these inputs, focusing on critical spatiotemporal

regions to distinguish violent actions from non-violent ones. This integrated approach addresses challenges such as complex motion patterns, subtle aggression cues, and varying environmental conditions.

3.1. Stage 1: Motion Feature Extraction with GMFlow

In this subsection, we present a detailed overview of GMFlow [13], the first stage of our framework, which serves as a pre-trained model for dense optical flow estimation. Violent actions in videos are often characterized by abrupt and chaotic motion patterns, which optical flow effectively captures. By estimating the magnitude and direction of movement between consecutive frames, GMFlow identifies significant motion variations that may indicate violent behavior, offering a robust representation of temporal dynamics.

3.1.1. GMFlow Framework Overview

GMFlow [13] redefines optical flow estimation as a global matching task, making it particularly effective for identifying large displacements and abrupt movements, which are characteristic of violent actions. Its architecture integrates a transformer-based module for feature enhancement, a global correlation layer for feature matching, and a self-attention mechanism for efficient flow propagation. These components work together to accurately capture motion dynamics in complex scenarios. The detailed structure of GMFlow is illustrated in Figure 2.

A. Feature Extraction and Enhancement: GMFlow extracts dense features from two consecutive video frames using a shared convolutional backbone. To enhance these features, a transformer-based module applies both self- and cross-attention mechanisms enriched with fixed 2D positional encodings. This module models spatial and temporal dependencies between frames effectively. A shifted local window attention strategy, inspired by the Swin Transformer [42], balances computational efficiency and accuracy by operating within adaptive local regions.
B. Global Matching and Flow Estimation: Enhanced features from both frames are compared using a global correlation matrix, measuring the similarity of each pixel in one frame with all pixels in the other. This matrix is normalized with a softmax operation, enabling differentiable training and sub-pixel accuracy. The resulting matching distribution determines pixel correspondences, which compute the optical flow as the displacement between matched pixel coordinates.
C. Flow Propagation and Refinement: To handle occluded and out-of-boundary regions, GMFlow employs a self-attention-based flow propagation mechanism. This ensures consistent flow estimation across the frame. Refinement is achieved by upscaling the initial flow predictions to higher resolution and applying residual learning, improving accuracy without significant computational overhead.

One of the key advantages of GMFlow is its robustness in handling challenging scenarios such as occlusion and low-frame-rate videos. The global matching mechanism enables the model to infer motion in occluded regions by leveraging contextual information from non-occluded areas, reducing the impact of missing pixel correspondences. Additionally, GMFlow's ability to process long-range dependencies allows it to maintain motion continuity even in low-frame-rate videos, where conventional flow estimation methods struggle with large temporal gaps. This ensures reliable motion encoding, enhancing the detection of violent actions in complex environments.

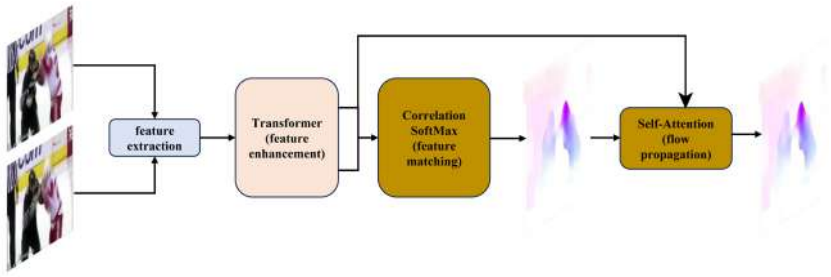

Figure 2. Overview of GMFlow: a global matching-based framework for optical flow estimation. The architecture integrates transformer-based feature enhancement, global feature matching, and flow refinement.

3.1.2. Motion Representation Output

In this subsection, we present the outputs generated by the GMFlow module, which are subsequently used in conjunction with RGB frames in the next stage to differentiate between violent and normal behaviors. The GMFlow module produces dense optical flow images for each consecutive frame pair, capturing the magnitude and direction of motion with high accuracy. These optical flow images serve as a detailed representation of temporal dynamics, enabling the framework to effectively analyze complex motion patterns often associated with violent actions.

By leveraging GMFlow's precise motion estimation, the framework minimizes the need for explicit temporal annotations, ensuring a robust and efficient approach to violence detection. Figure 3 illustrates examples from multiple datasets, showcasing the generated optical flow outputs. In each example, the leftmost images represent consecutive frames from a video sequence, while the image on the right depicts the corresponding optical flow, providing a visual representation of the captured temporal motion dynamics.

Figure 3. Illustrative examples of input video frames and their corresponding optical flow outputs. The top row displays samples from the Hockey dataset, while the bottom rows present samples from the UBI-Fight dataset (**left**) and the Crowd dataset (**right**). Each example consists of two consecutive input frames (**left**) and the computed optical flow representation (**right**), highlighting motion dynamics.

3.2. Stage 2: Spatial-Temporal Feature Integration with CBAM-Enhanced ResNet3D

In this subsection, we provide a detailed explanation of the second stage of our network, which is built upon the ResNet architecture with enhanced residual blocks integrated with the CBAM. While optical flow effectively captures temporal dynamics, it inherently lacks the spatial context required to interpret complex scene details and subtle actions. To overcome this limitation, the second stage of the framework combines RGB frames with the optical flow images generated by GMFlow in the first stage. Specifically, the first frame from each pair used in optical flow computation is selected to provide complementary spatial information.

This dual-input approach allows the network to leverage both spatial and temporal features, enabling a more thorough analysis of video sequences to detect violent actions. First, we provide an overview of the entire second stage of the CBAM-enhanced ResNet3D architecture. Subsequently, in a dedicated subsection, we delve into the CBAM mechanism, detailing its two components: channel attention and spatial attention.

3.2.1. CBAM-Enhanced ResNet3D Architecture

The CBAM-enhanced ResNet3D network effectively processes the combined RGB and optical flow inputs to model spatial-temporal features. As depicted in Figure 4, the architecture leverages 3D convolutions to capture temporal dependencies while incorporating the Convolutional Block Attention Module (CBAM) [15] to enhance feature representation through attention mechanisms. The second-stage network comprises three primary components:

1. Input Block: The input data, consisting of RGB frames and optical flow images, are initially processed through a 3D convolutional layer. This layer has a kernel size of $7 \times 7 \times 7$, with a temporal stride of 1 and spatial stride of 2. The layer is designed to extract low-level spatial-temporal features from consecutive frames. After convolution, the output is normalized using batch normalization to stabilize the learning process and improve convergence. A ReLU activation function is then applied element-wise to introduce non-linearity. This block is followed by a max-pooling layer with a kernel size of 3, a stride of 2, and padding of 1, which further reduces spatial dimensions and enhances the representation of the input data.
2. Residual Blocks with CBAM: The core of the architecture is composed of a series of residual blocks, each augmented with the Convolutional Block Attention Module (CBAM) [15]. These blocks are responsible for capturing hierarchical spatial-temporal features by utilizing 3D convolutions. The network employs a ResNet34-3D backbone, which is structured into four stages of residual blocks with configurations of $[3, 4, 6, 3]$, corresponding to the respective layers in the architecture. As shown in Figure 5, each residual block consists of two 3D convolutional layers with kernel sizes of $3 \times 3 \times 3$, with batch normalization applied between the convolutional layers, followed by a ReLU activation. The CBAM block is integrated between the second batch normalization and the ReLU activation, enhancing feature representation through a two-fold attention mechanism: channel attention and spatial attention. Additionally, each residual block incorporates a skip connection that connects the output of the CBAM block to the input after applying a $1 \times 1 \times 1$ 3D convolution, followed by batch normalization, facilitating the learning of residuals.
3. Prediction Block: Following the feature extraction process, the network passes the output through an adaptive average pooling layer to reduce dimensionality. The pooled features are then flattened and fed into a fully connected (FC) layer for the classification task, where the network categorizes the input sequence as either violent or non-violent.

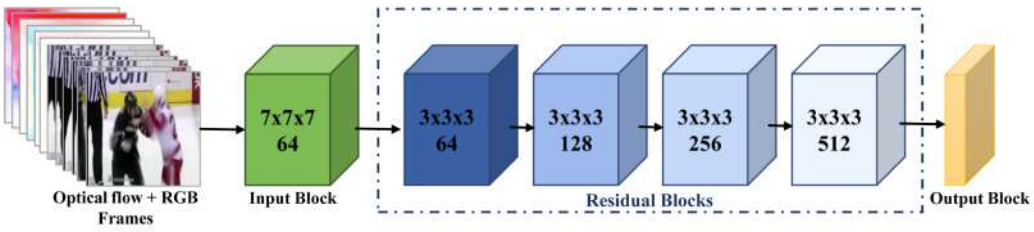

Figure 4. Overview of the CBAM-enhanced ResNet3D architecture for violent action detection. The model integrates RGB frames and optical flow inputs to extract comprehensive spatial-temporal features.

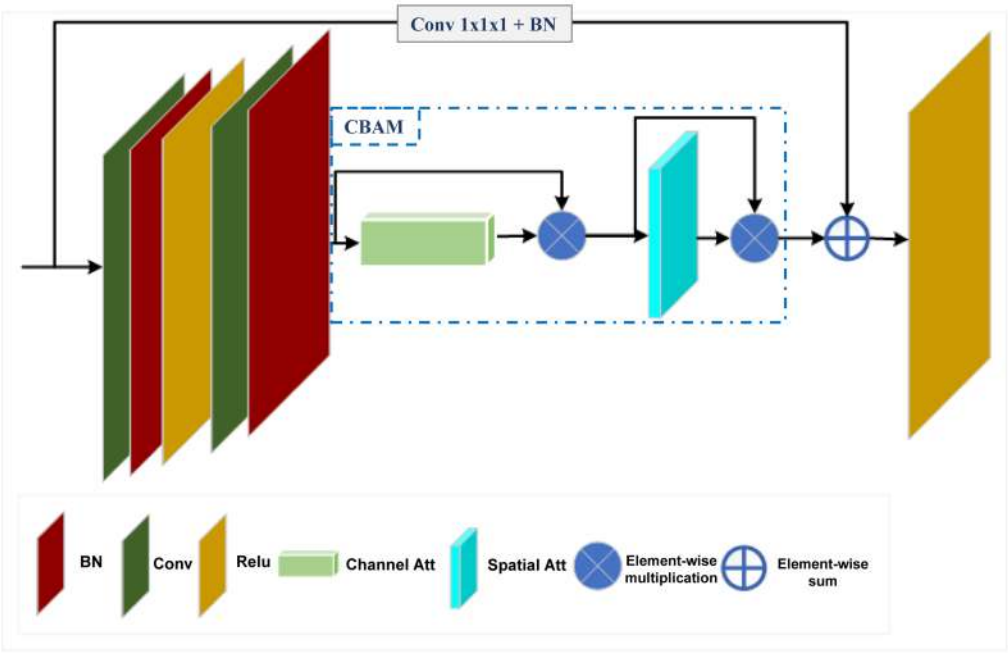

Figure 5. Structure of a residual block integrated with CBAM. The block combines residual learning with channel and spatial attention mechanisms to enhance feature selection.

3.2.2. CBAM Mechanisms for Attention-Enhanced Feature Extraction

CBAM refines feature representations in each residual block through two complementary attention mechanisms:

- Channel Attention: This mechanism prioritizes important feature channels by performing global pooling along spatial dimensions. It learns attention weights that highlight channels most relevant to violent action detection, as shown in Figure 6 (top).
- Spatial Attention: Complementing channel attention, spatial attention highlights key regions within each frame by generating an attention map based on inter-channel relationships. This mechanism enables the network to focus on interactions or movements indicative of violence, as depicted in Figure 6 (bottom).

CBAM enhances feature representation by applying two sequential attention modules: channel attention and spatial attention. The channel attention module computes global feature importance using both average pooling and max pooling operations, followed by a shared multi-layer perceptron (MLP), allowing the network to amplify the most informative feature channels. The spatial attention module then refines feature localization by applying a spatial weight map to highlight action-relevant regions, focusing on where violent movements occur.

The ResNet3D architecture effectively combines spatial and temporal features by integrating CBAM into the residual blocks. Leveraging both RGB and optical flow inputs,

the network achieves a holistic representation of video data. This dual-modality approach ensures robust detection of violent actions, even in challenging scenarios with subtle motion patterns, complex backgrounds, or occlusions. The enhanced spatial-temporal feature representation enables the framework to generalize effectively across diverse datasets and complex video sequences.

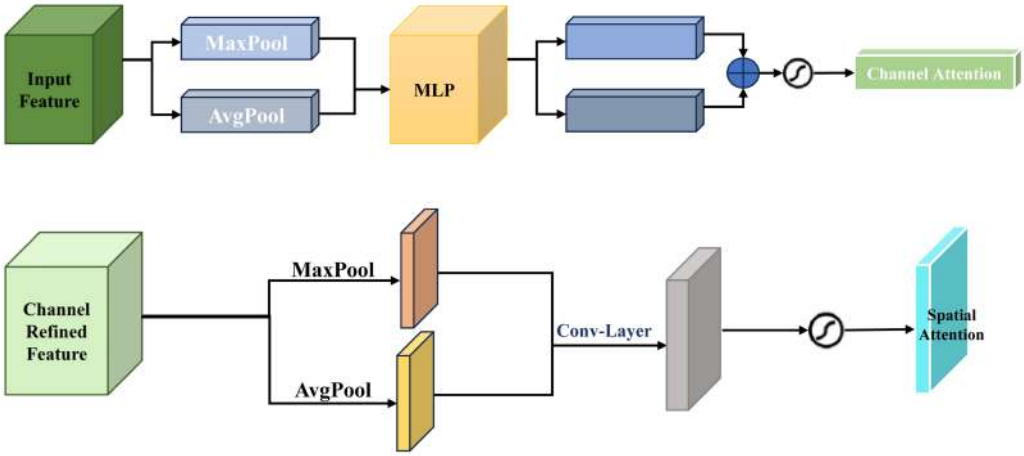

Figure 6. CBAM mechanisms: **Top**: Channel attention emphasizes relevant feature channels. **Bottom**: Spatial attention highlights key regions in the input.

At the end of the proposed approach, the entire two-stage framework can be summarized as follows: The input video is first processed by the GMFlow network in the first stage to generate optical flow images for each sequential frame pair. Specifically, the optical flow is computed for frame pairs corresponding to indices $N \times I$ and $N \times I + 1$, where I is an integer sequence $(0, 1, 2, \ldots)$. In the second stage , these generated optical flow frames are concatenated with the corresponding $N \times I$ RGB frames from the video. This combined input is then fed into the CBAM-enhanced ResNet3D network, which effectively learns complementary spatial and temporal features. Integrating these two stages, the framework achieves robust and efficient video violence detection.

4. Experiments and Results

In this section, we provide a comprehensive evaluation of our proposed two-stage video violence detection framework. The discussion begins with an overview of the datasets utilized for training and evaluation, emphasizing their diversity and significance in the context of violence detection. We then detail the experimental setup, including network configurations, training strategies, and implementation specifics. Standard evaluation metrics are defined and employed to quantitatively measure the framework's performance. The results of our approach are analyzed and benchmarked against state-of-the-art methods, demonstrating the superiority of integrating GMFlow-generated optical flow with the CBAM-enhanced ResNet3D architecture for capturing spatial-temporal dynamics. Lastly, we offer an in-depth discussion of the framework's strengths, limitations, potential for practical applications, and opportunities for future enhancement.

4.1. Datasets

We evaluated the proposed framework on three widely recognized datasets tailored for violence detection: Hockey Fight [16], Crowd Violence [17], and UBI-Fight [18]. These datasets present diverse challenges, ensuring a comprehensive evaluation of our model's performance. Details of each dataset are outlined below:

- Hockey Fight Dataset: This dataset consists of 1000 video sequences evenly divided into two categories: fights and non-fights. The videos are captured during hockey matches, making the dataset ideal for detecting aggressive behaviors in confined and fast-paced environments.
- Crowd Violence Dataset: Specifically curated to address the detection and classification of violent crowd behavior, this dataset comprises 246 videos, evenly split into 123 violent and 123 non-violent scenarios. The videos, collected from YouTube, feature a diverse range of real-world challenges, including varying scene types, video qualities, and surveillance conditions. To enhance usability, all videos were de-interlaced, stored as AVI files, and resized to 320×240 pixels, making this dataset a robust benchmark for uncontrolled, in-the-wild conditions.
- UBI-Fight Dataset: The UBI-Fight dataset serves as a comprehensive benchmark for detecting fighting events in surveillance footage. It contains 1000 videos, including 216 sequences of real-life fighting scenarios and 784 showcasing everyday activities. To ensure relevance, extraneous content such as video introductions and news segments was meticulously removed during curation. The dataset spans diverse environments and scenarios, capturing footage under varying conditions such as different times of day, uncontrolled poses, inconsistent lighting, and varying spatial scales. Frequent occlusions and complex interactions further challenge the detection process, making this dataset a robust testbed for evaluating video-based anomaly detection models.

A statistical summary of the datasets is presented in Table 2, offering a clear comparison of the number of videos and their violent and non-violent categorizations.

Table 2. Statistical overview of the datasets used for violence detection.

Dataset	Total Videos	Violent	Non-Violent	FPS
Hockey	1000	500	500	25
Crowd	246	123	123	25
UBI-Fight	1000	216	784	30

4.2. Experimental Setup

Each video dataset was resized to an image resolution of 256×256 to maintain consistency across inputs. The datasets were split into 80% for training and 20% for testing to ensure a fair evaluation. The experimental workflow was divided into two stages: In the first stage, optical flow images were generated using the pre-trained GMFlow network, which required no additional training. Instead, its pre-trained weights were loaded to process the video datasets and extract optical flow images, representing the temporal motion of video sequences. In the second stage, the CBAM-enhanced ResNet3D network was trained using both RGB frames and the optical flow images generated in Stage 1. This integration allowed the model to effectively capture complementary spatial and temporal features crucial for detecting violent actions.

The training spanned 200 epochs using the AdamW optimizer with an initial learning rate of 4×10^{-4} , which was progressively reduced via a cosine annealing scheduler to stabilize convergence. The batch size was set to 4, balancing GPU memory constraints with effective model training. The loss function employed was cross-entropy loss, suitable for the binary classification task (violent vs. non-violent).

The training process was conducted on a high-performance NVIDIA GeForce RTX 4090 GPU, leveraging its computational power to expedite processing. The system ran on Windows 11, ensuring compatibility with the latest software libraries. The model was implemented in PyTorch [43], utilizing its flexibility and efficiency for seamless implementation and training. This structured setup, combined with advanced hardware and software,

ensured effective model training while optimizing spatial-temporal feature integration for robust violence detection.

Integrating GMFlow-generated optical flow with RGB frames posed several challenges. First, the temporal misalignment issue—optical flow represents movement across two consecutive frames, while RGB images capture static spatial features. To mitigate this, we selected the first frame of each pair used for optical flow computation, ensuring a consistent spatiotemporal representation. Second, differences in feature scales between RGB and optical flow maps required normalization. We standardized optical flow values and aligned their intensity range with RGB inputs to facilitate effective feature fusion. Finally, model complexity was a consideration, as incorporating two different data streams increased the network's computational cost. To optimize efficiency, we employed CBAM to selectively focus on essential spatial-temporal features, reducing unnecessary computations and improving detection performance.

4.3. Evaluation Metrics

To evaluate the performance of our proposed framework, we adopt two widely recognized metrics: Accuracy (ACC) and Area Under the Receiver Operating Characteristic Curve (AUC). These metrics are selected to provide a well-rounded assessment of the model's ability to classify violent and non-violent actions.

(a) Accuracy (ACC): Accuracy is a fundamental metric that measures the proportion of correctly classified instances out of the total instances. It is formally defined as:

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

Accuracy provides an overall measure of the classification performance and is useful for evaluating the proportion of correct predictions. However, it can be sensitive to class imbalance, as it treats both classes equally without considering their proportions in the dataset. In cases where the dataset is imbalanced, accuracy might be misleading, especially when the majority class dominates the classification. Therefore, accuracy alone is not sufficient, but it serves as a basic indicator of overall performance.

(b) Area Under the Curve (AUC): AUC is a more robust metric that assesses the ranking ability of a binary classifier across all possible decision thresholds. It is derived from the Receiver Operating Characteristic (ROC) curve, which plots the False Positive Rate (FPR) on the x-axis against the True Positive Rate (TPR) on the y-axis.

The TPR and FPR are defined as follows:

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN}. \tag{2}$$

AUC provides a scalar value ranging from 0 to 1, where a value closer to 1 indicates excellent classification performance. A high AUC signifies that the model effectively distinguishes between positive and negative classes, even in the presence of class imbalance. In contrast to ACC, AUC evaluates how well the model ranks the instances, regardless of the decision threshold, making it less sensitive to class distribution.

Using both ACC and AUC provides complementary insights into the model's performance. While ACC evaluates the overall correctness of predictions, AUC measures the model's ability to distinguish between violent and non-violent actions, particularly useful in the context of class imbalance. Together, these metrics offer a comprehensive view of the model's effectiveness in classifying actions, balancing both overall performance and discriminatory power.

4.4. Results

The performance of the proposed framework was evaluated on three benchmark datasets: Hockey Fight, Crowd Violence, and UBI-Fight. The evaluation metrics were selected based on the characteristics of each dataset. Accuracy (ACC) was used for the balanced Hockey Fight and Crowd Violence datasets, while the area under the receiver operating characteristic curve (AUC) was employed for the unbalanced UBI-Fight dataset.

4.4.1. Performance on UBI-Fight Dataset

To address the imbalance in the UBI-Fight dataset, the AUC metric was used for evaluation. As shown in Table 3, our method achieved an AUC of 0.963, outperforming the state-of-the-art (SOTA) methods: Hasan et al. [32], Ravanbakhsh et al. [44], YS Chong et al. [45], Wang et al. [33], Bruno et al. [18], Sultani et al. [31], Degardin et al. [36], and Park et al. [4]. This demonstrates the robustness of the proposed framework in handling unbalanced datasets and its ability to capture complementary spatiotemporal representations.

Table 3. Performance comparison of the proposed framework with SOTA methods on the UBI-Fight dataset (evaluated using AUC).

Method	AUC
Hasan et al. [32]	0.528
Ravanbakhsh et al. [44]	0.533
YS Chong et al. [45]	0.541
Wang et al. [33]	0.610
Bruno et al. [18]	0.819
Sultani et al. [31]	0.892
Degardin et al. [36]	0.906
Park et al. [4]	0.952
Ours	**0.963**

4.4.2. Performance on Hockey Fight and Crowd Violence Datasets

For the balanced Hockey Fight and Crowd Violence datasets, ACC was used as the evaluation metric. Table 4 presents the results, where our framework achieved an accuracy of 97.5% on the Hockey Fight dataset and 94.0% on the Crowd Violence dataset. Compared with four SOTA methods—Hasan et al. [32], Sultani et al. [31], C3D [19], and Park et al. [4]—our approach demonstrated consistent improvement, showcasing its ability to generalize effectively to different violence detection scenarios.

Table 4. Performance comparison of the proposed framework with SOTA methods on Hockey Fight and Crowd Violence datasets (evaluated using ACC).

Method	Hockey Fight	Crowd Violence
Hasan et al. [32]	93.4	83.4
Sultani et al. [31]	96.8	**94.5**
Du Tran et al. [19]	96.5	84.44
Park et al. [4]	94.5	92.0
Ours	**97.5**	94.0

The results validate the proposed framework's ability to achieve consistently high performance across datasets with diverse characteristics. By integrating GMFlow-derived optical flow and spatial features from RGB frames, the framework effectively captures complementary spatiotemporal representations, achieving robust and versatile performance for video violence detection tasks.

To assess the generalization capability of our framework, we evaluated it on three benchmark datasets, each representing distinct challenges in violence detection. The Hockey Fight dataset captures structured, sports-based violent interactions, while the Crowd Violence dataset presents complex real-world group interactions in uncontrolled environments. UBI-Fight introduces additional diversity by including both staged and real-world violent incidents, incorporating varying lighting, background complexity, and camera perspectives. The strong performance of our model across these datasets (97.5% on Hockey, 94.0% on Crowd, and 0.963 AUC on UBI-Fight) highlights its robustness in handling different types of violent behaviors, demonstrating broad applicability beyond a single dataset.

4.5. Limitations

While the proposed framework demonstrates high accuracy across benchmark datasets, certain limitations must be acknowledged. The model's performance can be affected by environmental factors such as variations in lighting conditions, camera perspectives, and occlusions, which may impact the effectiveness of optical flow estimation. Additionally, our approach primarily focuses on detecting violent actions with clear motion cues, making it less effective in identifying subtle or non-physical aggression, such as intimidation or threats. Another limitation is the generalizability of our model to unseen real-world datasets, as the training datasets may not fully capture the diversity of violent scenarios encountered in real-world surveillance applications.

Despite its benefits, weakly supervised learning has limitations. One major challenge is the inability to localize violent segments within a video precisely, as labels are assigned at the video level. This can lead to misclassifications, particularly in scenes with mixed violent and non-violent activities. Additionally, weakly supervised models are more susceptible to biases in training data, making them less reliable when applied to unseen scenarios.

These challenges point to important areas for future research, which are explored in the following section.

4.6. Ablation Study: Evaluating the Impact of Optical Flow and CBAM

This subsection presents an ablation study to analyze the contributions of optical flow generation and the CBAM-enhanced ResNet3D network to the overall performance of the proposed framework. The objective is to examine how these components, individually and collectively, influence the effectiveness of video violence detection. We trained and tested four variations of the model to analyze their contributions:

1. Baseline: ResNet3D without CBAM and without the first stage (using RGB frames only).
2. ResNet3D + CBAM: ResNet3D with CBAM but without the first stage (using RGB frames only).
3. ResNet3D + Optical Flow: ResNet3D without CBAM but incorporating the first stage (optical flow combined with RGB frames).
4. Full Framework: ResNet3D with CBAM and the first stage (optical flow combined with RGB frames).

The evaluation metrics remain consistent across all configurations to ensure a fair comparison. Specifically, we report the model's accuracy (ACC) and improvement percentage (%) relative to the baseline. As shown in Table 5, incorporating optical flow and CBAM independently leads to performance gains, while the combination of both achieves the highest accuracy. This demonstrates the complementary strengths of spatial attention mechanisms (CBAM) and motion features (optical flow) in improving video violence detection.

Table 5. Ablation study evaluating the impact of optical flow and CBAM on video violence detection. Each configuration is assessed based on accuracy (ACC) and relative improvement over the baseline. ✓ indicates a component is used, while ✗ indicates it is not.

Config. No.	CBAM	Optical Flow	ACC (%)	Improvement (%)
1 (Baseline)	✗	✗	94.0	-
2 (ResNet3D + CBAM)	✓	✗	95.0	+1.0
3 (ResNet3D + Optical Flow)	✗	✓	95.5	+1.5
4 (Full Framework)	✓	✓	97.5	+3.5

The findings from Table 5 confirm the significant contributions of both optical flow and CBAM to the proposed framework's performance. The model achieves enhanced spatial-temporal feature representation by integrating these components, resulting in superior accuracy.

The ablation study highlights the importance of combining motion and spatial attention mechanisms for robust violence detection. While ResNet3D alone provides a strong baseline, integrating CBAM improves its ability to selectively focus on informative regions, and the addition of optical flow enhances temporal feature representation. This synergy underscores the necessity of both components for achieving state-of-the-art results on the Hockey dataset.

5. Conclusions and Future Work

In this paper, we propose a two-stage framework for video violence detection that integrates GMFlow-generated optical flow with a CBAM-enhanced ResNet3D architecture to capture both spatial and temporal features. Our approach demonstrated superior performance across multiple benchmark datasets, achieving high accuracy and AUC scores when compared to state-of-the-art methods. The results prove the robustness and versatility of our model in handling diverse video scenarios with varying complexities.

However, several options for future work could further enhance the effectiveness of the proposed framework. First, we plan to use additional video features like audio and motion-based cues to improve the model's contextual understanding. Additionally, we aim to explore applying semi-supervised or unsupervised learning techniques to strengthen large amounts of unlabeled data, which could significantly improve the model's generalization and scalability. Finally, we intend to expand our framework's applicability to a wider range of real-world scenarios, such as detecting violence in highly dynamic or occluded environments. By enhancing the robustness of the model to challenging situations, we believe the framework can be adapted to broader video surveillance applications, such as monitoring public spaces, events, and crowd behavior.

Furthermore, while our study focuses on model accuracy and robustness, a comprehensive computational cost analysis, including inference time and memory consumption, remains an essential area for future exploration. Unlike some SOTA methods that primarily report classification performance, future research will aim to provide a detailed evaluation of computational efficiency, facilitating real-time deployment and resource optimization.

Author Contributions: Conceptualization, M.M., M.S.K. and B.Y.; Methodology, M.M., M.S.K. and M.A.; Software, M.M., B.Y., M.F.S. and M.A.; Validation, M.M. and H.-S.K.; Formal analysis, M.M. and M.F.S.; Investigation, H.-S.K.; Resources, H.-S.K.; Data curation, M.M.; Writing—original draft preparation, M.M.; Writing—review and editing, M.M. and H.-S.K.; Visualization, H.-S.K.; Supervision, H.-S.K.; Project administration, H.-S.K.; Funding acquisition, H.-S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology

Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (IITP-2025-RS-2020-II201462).

Data Availability Statement: The datasets used in this paper are public datasets.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CBAM	Convolutional Block Attention Module
AUC	Area Under the Curve
ACC	Accuracy
CNNs	Convolutional Neural Networks
LSTM	Long Short-Term Memory
GAN	Generative Adversarial Network
MIL	Multiple Instance Learning
RTFM	Robust Temporal Feature Magnitude
I3D	Inflated 3D ConvNet
SOTA	State-Of-The-Art
RGB	Red-Green-Blue

References

1. Stauffer, C.; Grimson, W.E.L. Adaptive background mixture models for real-time tracking. In Proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), Fort Collins, CO, USA, 23–25 June 1999; Volume 2, pp. 246–252.
2. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
3. Gao, J.; Shi, J.; Balla, P.; Sheshgiri, A.; Zhang, B.; Yu, H.; Yang, Y. Camera-Based Crime Behavior Detection and Classification. *Smart Cities* **2024**, *7*, 1169–1198. [CrossRef]
4. Park, J.H.; Mahmoud, M.; Kang, H.S. Conv3D-based video violence detection network using optical flow and RGB data. *Sensors* **2024**, *24*, 317. [CrossRef] [PubMed]
5. Hwang, I.C.; Kang, H.S. Anomaly detection based on a 3d convolutional neural network combining convolutional block attention module using merged frames. *Sensors* **2023**, *23*, 9616. [CrossRef]
6. Lee, J.W.; Kang, H.S. Three-stage deep learning framework for video surveillance. *Appl. Sci.* **2024**, *14*, 408. [CrossRef]
7. Senussi, M.F.; Kang, H.S. Occlusion Removal in Light-Field Images Using CSPDarknet53 and Bidirectional Feature Pyramid Network: A Multi-Scale Fusion-Based Approach. *Appl. Sci.* **2024**, *14*, 9332. [CrossRef]
8. Mahmoud, M.; Kang, H.S. Ganmasker: A two-stage generative adversarial network for high-quality face mask removal. *Sensors* **2023**, *23*, 7094. [CrossRef]
9. Bauer, A.; Nakajima, S.; Müller, K.R. Self-Supervised Autoencoders for Visual Anomaly Detection. *Mathematics* **2024**, *12*, 3988. [CrossRef]
10. Mahmoud, M.; Kasem, M.S.; Kang, H.S. A Comprehensive Survey of Masked Faces: Recognition, Detection, and Unmasking. *arXiv* **2024**, arXiv:2405.05900. [CrossRef]
11. Song, W.; Liao, B.; Ning, K.; Yan, X. Improved Real-Time Detection Transformer-Based Rail Fastener Defect Detection Algorithm. *Mathematics* **2024**, *12*, 3349. [CrossRef]
12. Althiyabi, T.; Ahmad, I.; Alassafi, M.O. Enhancing IoT Security: A Few-Shot Learning Approach for Intrusion Detection. *Mathematics* **2024**, *12*, 1055. [CrossRef]
13. Xu, H.; Zhang, J.; Cai, J.; Rezatofighi, H.; Tao, D. Gmflow: Learning optical flow via global matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8121–8130.
14. Li, H.; Li, X.; Su, L.; Jin, D.; Huang, J.; Huang, D. Deep spatio-temporal adaptive 3d convolutional neural networks for traffic flow prediction. *ACM Trans. Intell. Syst. Technol. (TIST)* **2022**, *13*, 1–21. [CrossRef]
15. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
16. Mukherjee, S.; Saini, R.; Kumar, P.; Roy, P.P.; Dogra, D.P.; Kim, B.G. Fight detection in hockey videos using deep network. *J. Multimed. Inf. Syst.* **2017**, *4*, 225–232.

17. Hassner, T.; Itcher, Y.; Kliper-Gross, O. Violent flows: Real-time detection of violent crowd behavior. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 1–6.
18. Degardin, B.; Proença, H. Iterative weak/self-supervised classification framework for abnormal events detection. *Pattern Recognit. Lett.* **2021**, *145*, 50–57. [CrossRef]
19. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
20. Hochreiter, S. Long Short-term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
21. Thakare, K.V.; Raghuwanshi, Y.; Dogra, D.P.; Choi, H.; Kim, I.J. Dyannet: A scene dynamicity guided self-trained video anomaly detection network. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 5541–5550.
22. Al-Lahham, A.; Tastan, N.; Zaheer, M.Z.; Nandakumar, K. A coarse-to-fine pseudo-labeling (c2fpl) framework for unsupervised video anomaly detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 6793–6802.
23. Hu, J.; Yu, G.; Wang, S.; Zhu, E.; Cai, Z.; Zhu, X. Detecting anomalous events from unlabeled videos via temporal masked auto-encoding. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022; pp. 1–6.
24. Tao, C.; Wang, C.; Lin, S.; Cai, S.; Li, D.; Qian, J. Feature Reconstruction with Disruption for Unsupervised Video Anomaly Detection. *IEEE Trans. Multimed.* **2024**, *26*, 10160–10173. [CrossRef]
25. Luo, W.; Liu, W.; Gao, S. Remembering history with convolutional lstm for anomaly detection. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 439–444.
26. Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M.R.; Venkatesh, S.; Hengel, A.v.d. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 1705–1714.
27. Wang, L.; Tan, H.; Zhou, F.; Zuo, W.; Sun, P. Unsupervised anomaly video detection via a double-flow convlstm variational autoencoder. *IEEE Access* **2022**, *10*, 44278–44289. [CrossRef]
28. Liu, W.; Luo, W.; Lian, D.; Gao, S. Future frame prediction for anomaly detection—A new baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6536–6545.
29. Li, D.; Nie, X.; Li, X.; Zhang, Y.; Yin, Y. Context-related video anomaly detection via generative adversarial network. *Pattern Recognit. Lett.* **2022**, *156*, 183–189. [CrossRef]
30. Wang, X.; Yan, Y.; Tang, P.; Bai, X.; Liu, W. Revisiting multiple instance neural networks. *Pattern Recognit.* **2018**, *74*, 15–24. [CrossRef]
31. Sultani, W.; Chen, C.; Shah, M. Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6479–6488.
32. Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A.K.; Davis, L.S. Learning temporal regularity in video sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2016; pp. 733–742.
33. Wang, T.; Qiao, M.; Lin, Z.; Li, C.; Snoussi, H.; Liu, Z.; Choi, C. Generative neural networks for anomaly detection in crowded scenes. *IEEE Trans. Inf. Forensics Secur.* **2018**, *14*, 1390–1399. [CrossRef]
34. Deshpande, K.; Punn, N.S.; Sonbhadra, S.K.; Agarwal, S. Anomaly detection in surveillance videos using transformer based attention model. In *International Conference on Neural Information Processing*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 199–211.
35. Jin, P.; Mou, L.; Xia, G.S.; Zhu, X.X. Anomaly detection in aerial videos with transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
36. Degardin, B.M. Weakly and Partially Supervised Learning Frameworks for Anomaly Detection. Master's Thesis, Universidade da Beira Interior, Covilhã, Portugal, 2020.
37. Chen, Y.; Liu, Z.; Zhang, B.; Fok, W.; Qi, X.; Wu, Y.C. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 387–395.
38. Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J.W.; Carneiro, G. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 4975–4986.
39. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.

40. Zhang, S.; Wang, T.; Wang, C.; Wang, Y.; Shan, G.; Snoussi, H. Video object detection base on rgb and optical flow analysis. In Proceedings of the 2019 2nd China Symposium on Cognitive Computing and Hybrid Intelligence (CCHI), Xi'an, China, 21–22 September 2019; pp. 280–284.
41. Zhu, Y.; Lan, Z.; Newsam, S.; Hauptmann, A. Hidden two-stream convolutional networks for action recognition. In Proceedings of the Computer Vision—ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; Revised Selected Papers, Part III 14; Springer: Berlin/Heidelberg, Germany, 2019; pp. 363–378.
42. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
43. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*. [CrossRef]
44. Chong, Y.S.; Tay, Y.H. Abnormal event detection in videos using spatiotemporal autoencoder. In Proceedings of the Advances in Neural Networks-ISNN 2017: 14th International Symposium, ISNN 2017, Sapporo, Hakodate, and Muroran, Hokkaido, Japan, 21–26 June 2017; Proceedings, Part II 14; Springer: Berlin/Heidelberg, Germany, 2017; pp. 189–196.
45. Ravanbakhsh, M.; Nabi, M.; Sangineto, E.; Marcenaro, L.; Regazzoni, C.; Sebe, N. Abnormal event detection in videos using generative adversarial nets. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 1577–1581.