Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

A Pose-Driven Relational Transformer Framework for Violence Detection in Surveillance Videos

P. Sreenivas, S. Pallavi (B22CS030), K. Deekshitha (B22CS056), G. Anvesh (B22CS035), J. Rahul (B22CS032)

Department of Computer Science and Engineering
KITS Warangal

***Abstract*—Violence detection in surveillance videos is a challenging task due to background clutter, camera motion, illumination variations, and occlusions. Conventional appearance-based deep learning methods often fail to generalize under such unconstrained conditions. This paper proposes a Pose-Driven Relational Transformer (RPAT) framework that emphasizes human body dynamics and interaction patterns rather than raw pixel appearance. The proposed system integrates RGB features, skeletal pose heatmaps, and motion-enhanced representations to form a robust multimodal feature space. A transformer-based temporal encoder models long-range dependencies, while a relational memory module captures sustained aggressive interactions across frames. Event prompt tokens and contrastive alignment further enhance classification reliability. Experimental evaluation on the Hockey Fight Dataset demonstrates that the proposed approach achieves improved robustness and accuracy in complex surveillance environments, making it suitable for real-world public safety applications.**

***Index Terms*—Violence Detection, Surveillance Videos, Pose Estimation, Relational Transformer, Multimodal Learning, Human Action Recognition**

I. Introduction

Violence detection has become an essential component of modern surveillance and public safety systems due to the increasing deployment of cameras in public spaces such as streets, campuses, transportation hubs, and commercial areas. Manual monitoring of surveillance feeds is both labor-intensive and error-prone, making automated violence recognition a critical requirement. However, detecting violent activity in unconstrained environments remains a difficult problem because violent actions often occur suddenly, involve complex human interactions, and are visually similar to normal activities such as sports or playful behavior.

Early computer vision approaches primarily relied on handcrafted features such as optical flow, spatio-temporal interest points, and motion histograms. While these methods achieved moderate success in controlled environments, their performance degraded significantly in real-world scenarios due to background clutter, camera motion, and varying lighting conditions. With the rise of deep learning, convolutional neural networks (CNNs) and 3D CNNs became popular for action recognition tasks. Despite their strong representation power, appearance-based CNN models are highly sensitive to irrelevant visual information and often misclassify non-violent actions that share similar visual patterns with violent events.

Human pose and skeletal motion provide a more reliable representation for understanding aggressive behavior, as violence is fundamentally characterized by body dynamics, forceful movements, and physical interactions. Pose-based representations are less affected by background noise and illumination changes, making them suitable for surveillance environments. Recent advancements in pose estimation and graph-based learning have enabled models to capture fine-grained human movement patterns and relational interactions between multiple individuals. However, many existing pose-based approaches struggle to model long-term temporal dependencies and sustained violent behavior.

Transformer architectures have shown remarkable success in modeling long-range temporal relationships through self-attention mechanisms. By treating video frames as sequences of tokens, transformers can learn contextual dependencies across time more effectively than recurrent models. Motivated by these advantages, this work proposes a pose-driven relational transformer framework that integrates pose dynamics, motion cues, and relational reasoning for violence detection. The system emphasizes human-centered motion understanding instead of raw appearance, allowing it to remain robust under occlusion, crowd density, and visual complexity.

Furthermore, violent behavior often involves repetitive or escalating actions rather than isolated movements. Capturing such patterns requires memory-aware temporal reasoning. To address this, a relational memory module is incorporated to maintain contextual information across frames. Combined with event prompt tokens and contrastive multimodal alignment, the proposed approach achieves improved discrimination between violent and non-violent activities. This makes the system suitable for real-time and large-scale surveillance deployments.

II. Literature Survey

A. Early Approaches and Handcrafted Features

Violence detection in videos has been extensively studied in the broader domain of human action recognition. Early works relied on handcrafted motion features such as optical flow magnitude, motion energy images, and spatio-temporal descriptors. Solmaz et al. [13] introduced the concept of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

stability analysis for identifying anomalous behaviors in crowd scenes, demonstrating that motion-based features could effectively capture crowd instability. Hassner et al. [14] proposed violent flows, a real-time detection method based on optical flow magnitude and improved dense trajectories. While these approaches were computationally efficient, they lacked robustness in real-world scenarios where camera motion and background clutter dominate the visual scene, severely limiting their applicability to real surveillance environments.

B. Deep Learning and Convolutional Neural Networks

With the emergence of deep learning, CNN-based architectures became the dominant approach for violence detection. Simonyan and Zisserman [1] introduced the two-stream convolutional network architecture, which processes spatial frames and temporal optical flow separately before fusing their representations. This seminal work demonstrated that motion is complementary to appearance, forming the foundation for subsequent spatio-temporal models. Tran et al. [2] extended this concept with 3D CNNs, enabling direct learning of spatio-temporal patterns through volumetric convolutions. Although these models achieved higher accuracy than handcrafted methods, they primarily relied on appearance cues, making them sensitive to lighting variations and visually similar non-violent actions such as sports.

Feichtenhofer et al. [3] introduced SlowFast networks, which process videos at multiple temporal rates to capture both fine-grained spatial details and rapid temporal patterns. The SlowFast architecture achieved state-of-the-art results on several action recognition benchmarks. However, despite their strong representation power, all appearance-based CNN models suffer from sensitivity to illumination variations and struggle in crowded or occluded environments. Large model sizes and high computational costs also limited their deployment on resource-constrained systems.

C. Multimodal Learning Approaches

To overcome limitations of single-modality approaches, multimodal learning techniques were introduced, combining RGB frames with audio signals or optical flow. Ullah et al. [12] proposed a spatiotemporal convolutional long short-term memory (ConvLSTM) network for violence detection, demonstrating improved accuracy through temporal sequence modeling. While multimodal fusion improved detection accuracy, optical flow computation is computationally expensive and sensitive to noise. Additionally, such methods still depend heavily on pixel-level representations, which may fail in crowded or heavily occluded environments. Lightweight motion modeling techniques were later proposed to reduce computational overhead, but they often lacked semantic understanding of human interactions and struggled with complex scenarios.

D. Skeleton-Based and Pose-Driven Methods

Skeleton-based and pose-driven methods have gained increasing attention due to their robustness against background noise and computational efficiency. Yan et al. [5] introduced Spatial-Temporal Graph Convolutional Networks (ST-GCN), which represent human skeletons as graphs where nodes are joints and edges encode spatial connections. ST-GCN showed superior performance compared to traditional CNN-based approaches, particularly in crowded scenes. Shi et al. [6] extended this work with Two-Stream Adaptive Graph Convolutional Networks (2s-AGCN), enabling adaptive graph learning that dynamically adjusts edge weights based on the input data.

Liu et al. [7] proposed Disentangling and Unifying Graph Convolutions (MSG3D), which decompose graph convolutions into multiple scales to capture multi-scale spatial-temporal patterns. Duan et al. [8] revisited skeleton-based action recognition with improved graph attention mechanisms. While these graph-based models demonstrated improved violence recognition using skeleton-based features, particularly in crowded scenes, many GCN-based models struggle to capture long-range temporal dependencies and interaction patterns across multiple individuals.

Recent advances in pose estimation have made real-time skeletal keypoint extraction more accessible. Lugaresi et al. [9] introduced MediaPipe, a lightweight framework for building perception pipelines with real-time multi-person pose estimation. MediaPipe's efficiency and robustness have made it widely adopted in computer vision applications. The availability of efficient pose estimation has enabled the development of lightweight, real-time violence detection systems that operate on CPU-based devices.

E. Transformer-Based Approaches for Video Understanding

Transformer architectures have revolutionized video understanding through their ability to model long-range temporal dependencies via self-attention mechanisms. Dosovitskiy et al. [4] introduced Vision Transformer (ViT), which divides images into patches and applies transformer blocks directly to image sequences. ViT demonstrated that pure transformer-based approaches, without convolutions, could achieve competitive results on image classification tasks when trained on large datasets.

Recent works have successfully integrated transformer architectures with pose information. Duan et al. [8] applied transformers to skeleton-based action recognition, demonstrating improved temporal reasoning compared to recurrent models. The flexibility of transformer attention mechanisms enables modeling of complex temporal patterns and long-range dependencies that are crucial for recognizing sustained violent behavior. However, limited attention has been given to relational memory mechanisms and sustained aggression modeling, which are critical for accurately detecting violence in surveillance scenarios.

F. Self-Supervised and Contrastive Learning

Self-supervised and contrastive learning techniques have emerged as powerful methods for learning robust representations. Chen et al. [10] introduced SimCLR, a simple

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

framework for contrastive learning of visual representations. SimCLR demonstrates that contrastive learning can produce learned representations comparable to supervised learning, even without labeled data. Grill et al. [11] proposed Bootstrap Your Own Latent (BYOL), which achieves strong performance through self-supervised learning without requiring negative pairs in the contrastive loss.

These contrastive approaches have proven effective for multi-modal learning tasks, where aligning representations across different modalities improves overall system robustness and generalization. In the context of violence detection, contrastive learning can enforce consistency across RGB appearance, pose structure, and motion dynamics.

G. Anomaly and Violence Detection in Surveillance

Specialized work on violence detection in surveillance has focused on developing robust systems for real-world deployment. Solmaz et al. [13] introduced stability-based anomaly detection for crowd analysis, showing that analyzing optical flow divergence can identify abnormal crowd behaviors. Hassner et al. [14] developed violent flows for real-time detection in surveillance videos, pioneering the application of dense optical flow to violence detection.

Niebles et al. [15] proposed unsupervised learning of visual representations using videos, demonstrating the value of temporal information for learning action-specific features. These foundational works established that temporal dynamics and motion patterns are crucial for distinguishing violent behavior from normal activities.

H. Motivation for Proposed Work

The proposed RPAT framework builds upon these advancements by introducing a novel pose-driven relational transformer that combines the strengths of multiple approaches:

- **Pose-Centric Processing:** Unlike CNN-only approaches that overemphasize appearance, the framework explicitly models human skeletal dynamics as the primary signal for violence detection.
- **Relational Memory:** Beyond standard transformer attention, learnable memory slots maintain contextual information across frames, enabling recognition of sustained and escalating aggressive patterns.
- **Multimodal Fusion:** RGB features provide scene context, pose heatmaps capture body dynamics, and motion maps reveal movement intensity. This combination addresses complementary aspects of violent behavior.
- **Computational Efficiency:** By leveraging lightweight ViT-tiny and real-time MediaPipe pose estimation, the framework remains suitable for practical surveillance deployment on resource-constrained systems.
- **Contrastive Alignment:** Novel integration of contrastive learning across modalities improves robustness to noise in individual feature extractors and enhances generalization.

The proposed work directly addresses gaps identified in existing literature by providing a unified framework that effectively combines pose dynamics, transformer temporal reasoning, and multimodal consistency for robust violence detection in unconstrained surveillance environments.

III. Problem Statement

Despite significant advancements in video-based violence detection, existing systems often fail in unconstrained surveillance environments due to background clutter, occlusions, camera motion, and illumination variations. Appearance-based models struggle to accurately infer violent intent and frequently misclassify visually similar non-violent activities. Therefore, there is a need for a robust violence detection system that focuses on human motion dynamics and interaction patterns rather than unreliable pixel-level features.

IV. Objectives

The primary objective of this project is to develop a pose-driven relational transformer-based violence detection system that accurately identifies violent activity in surveillance videos. The system aims to integrate human pose dynamics, motion cues, and relational reasoning to improve robustness and reliability. Additionally, the model seeks to achieve efficient temporal modeling and practical feasibility for real-world deployment.

V. Dataset Description

The proposed system is evaluated using the Hockey Fight dataset, which contains video clips of violent (fight) and non-violent (non-fight) scenarios captured during hockey matches. The dataset presents challenges such as fast motion, crowd presence, and visually similar non-violent actions. Videos are preprocessed into frame sequences, and pose information is extracted for multimodal feature learning.

The dataset comprises 1,000 videos with a balanced distribution of 500 fight clips and 500 non-fight clips. Videos are of variable duration (2-20 seconds) at 25 fps original frame rate. Frame extraction is performed at 8 fps target rate, resulting in 14-30 frames per video on average. Using sliding-window sampling with sequence length $T = 8$ frames and stride of 1 frame, the dataset yields approximately 807 training sequences and 200 validation sequences after stratified 80-20 splitting.

VI. Proposed Methodology

The proposed system introduces a multimodal *Relational Pose-driven Action Transformer (RPAT)* framework for effective violence detection in surveillance videos. The core motivation behind this methodology is to move beyond appearance-based representations and instead focus on human-centric motion dynamics, pose relationships, and long-term temporal context. By jointly modeling RGB information, skeletal pose, and motion cues, the system achieves robust performance even in complex and unconstrained environments such as crowded public spaces, low-light conditions, and scenes with camera motion.

The overall pipeline consists of video preprocessing, pose heatmap generation, motion enhancement, multimodal feature encoding, transformer-based temporal modeling, relational

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

memory integration, and final classification. Each stage is carefully designed to capture complementary aspects of violent behavior while suppressing irrelevant background information.

A. Video Preprocessing and Frame Extraction

Given an input surveillance video V, the video is first decomposed into a sequence of RGB frames:

$$V = \{F_1, F_2, \ldots, F_T\} \tag{1}$$

where T represents the total number of frames in the video. Frame extraction is performed at a fixed frame rate to preserve temporal consistency. Each frame is resized to a standard resolution of 224×224 and normalized using ImageNet statistics to reduce the impact of illumination variations, camera noise, and compression artifacts. This preprocessing stage ensures uniform input quality and improves the reliability of feature extraction in subsequent modules.

B. Pose Heatmap Generation

For each preprocessed RGB frame, a human pose estimation model is applied to identify key skeletal joints such as the head, shoulders, elbows, wrists, hips, knees, and ankles. MediaPipe Pose, a lightweight real-time pose estimation framework [9], is employed due to its computational efficiency and robustness. The detected keypoints are transformed into pose heatmaps that encode spatial body structure and posture:

$$H^{(t)}(u, v) = \sum_{i=1}^{K} \exp\left(-\frac{(u - x_i W)^2 + (v - y_i H)^2}{2\sigma^2}\right) \tag{2}$$

where $K = 17$ represents the number of joints, (x_i, y_i) are normalized keypoint coordinates, (u, v) are heatmap spatial coordinates, $H = W = 56$ is the heatmap resolution, and $\sigma = 1.5$ is the Gaussian width. Pose heatmaps provide a compact and structured representation of human movement, allowing the system to focus on body dynamics rather than background appearance. This abstraction is particularly beneficial in surveillance scenarios where occlusions, crowd density, and background clutter are common.

C. Motion Enhancement

Violent actions are often characterized by sudden, irregular, and high-intensity movements. To capture such dynamics, motion features are extracted using two complementary approaches:

Frame Differencing:

$$D_t = \frac{|G_t - G_{t-1}|}{255} \tag{3}$$

where G_t denotes the grayscale version of frame F_t. Frame differencing highlights regions of rapid movement and abrupt changes, which are strong indicators of aggressive behavior.

Optical Flow:

$$M_t = \frac{\sqrt{u_t^2 + v_t^2}}{\max(M) + \epsilon} \tag{4}$$

where (u_t, v_t) are horizontal and vertical flow components computed via Farneback's algorithm. The magnitude is normalized to $[0, 1]$ range. Both motion features are downsampled to 56×56 resolution and stacked into a two-channel motion map $M^{(t)} \in \mathbb{R}^{2\times56\times56}$.

D. ultimodal Feature Encoding

The RGB frames, pose heatmaps, and motion maps are independently encoded using modality-specific embedding layers to extract discriminative feature representations:

RGB Feature Extraction: A pretrained Vision Transformer (ViT-tiny) [4] backbone processes each frame to extract global appearance features:

$$E_t^{rgb} = \text{ViT}(F_t) \in \mathbb{R}^{D_{rgb}} \tag{5}$$

where $D_{rgb} = 192$ is the output feature dimension.

Pose Feature Extraction: Spatial-channel attention and convolutional layers process pose heatmaps:

$$E_t^{pose} = \text{Conv2D+Pool+Linear}(H^{(t)}) \in \mathbb{R}^{D_{pose}} \tag{6}$$

Motion Feature Extraction: Similar convolutional processing extracts motion-specific features:

$$E_t^{motion} = \text{Conv2D+Pool+Linear}(M^{(t)}) \in \mathbb{R}^{D_{motion}} \tag{7}$$

These embeddings are projected to a common dimension (256-d) and fused via averaging to form a unified multimodal representation:

$$E_t = \text{Proj}_{rgb}(E_t^{rgb}) + \text{Proj}_{pose}(E_t^{pose}) + \text{Proj}_{motion}(E_t^{motion}) \tag{8}$$

This fusion enables comprehensive action understanding by integrating complementary information from multiple modalities.

E. Transformer-based Temporal Modeling

The fused multimodal embeddings are converted into token sequences and passed to a vision transformer encoder. The transformer leverages self-attention mechanisms to capture long-range temporal dependencies across frames [4]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{9}$$

where Q, K, and V are query, key, and value projections, and d is the feature dimension. Multi-head attention with h heads enables the model to attend to multiple aspects of temporal dynamics simultaneously:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{10}$$

This temporal modeling capability allows the system to analyze the evolution of actions over time and distinguish short, non-violent movements from sustained aggressive behavior, which is a key characteristic of violent events.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

F. Relational Memory Module

To explicitly model long-term interactions and maintain contextual continuity, a relational memory module is integrated with the transformer encoder. The memory module maintains a set of learnable memory slots $M \in \mathbb{R}^{M \times D}$ where $M = 4$ is the number of slots and $D = 256$ is the slot dimension. At each time step, the memory is updated through attention-based interaction with the token sequence:

$$M_t = \text{LayerNorm}(M_{t-1} + \text{MultiHeadAttention}(M_{t-1}, E_t, E_t)) \tag{11}$$

followed by feed-forward refinement:

$$M_t = M_t + \text{FFN}(M_t) \tag{12}$$

This module enables the system to recognize repetitive and escalating patterns of aggression, such as continuous hitting or chasing, rather than relying solely on isolated motion cues. The memory acts as a bottleneck that learns to accumulate violence-relevant context across frames.

G. Event Prompt Tokens

To further enhance discrimination between violent and non-violent activities, learnable event prompt tokens are introduced:

$$P = \{p_{violent}, p_{non-violent}\} \in \mathbb{R}^{2 \times D} \tag{13}$$

These tokens are learned during training and interact with the transformer attention mechanism. They serve as anchors that guide the model toward violence-relevant motion patterns and interactions. Event prompt tokens improve classification confidence by biasing the attention mechanism toward action-specific features. The tokens are concatenated with the main token sequence before feeding to the transformer:

$$E_t' = [P\,;\, E_t] \tag{14}$$

H. Contrastive Multimodal Alignment

To ensure consistency across RGB, pose, and motion modalities, contrastive learning is employed [10] to align their feature representations. Per-modality embeddings are extracted by pooling the temporal dimension:

$$e^{rgb} = \text{AvgPool}(\text{Proj}^{rgb}_{contrast}(E^{rgb})) \in \mathbb{R}^{D_c} \tag{15}$$

where $D_c = 128$ is the contrastive embedding dimension. Similar embeddings are obtained for pose and motion. Normalized Temperature-scaled Cross-Entropy (NT-Xent) loss enforces modality alignment [10]:

$$L_{contrast} = L_{NT-Xent}(e^{rgb}, e^{pose}) + L_{NT-Xent}(e^{rgb}, e^{motion}) \tag{16}$$

where:

$$L_{NT-Xent}(z_i, z_j) = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_k \exp(\text{sim}(z_i, z_k)/\tau)} \tag{17}$$

with temperature $\tau = 0.1$. This alignment enforces modality-invariant feature learning and improves robustness under challenging conditions such as low illumination, occlusion, and noisy pose estimation.

I. Classification Head

The final feature representation obtained after transformer processing, relational memory integration, and multimodal alignment is passed through a multilayer perceptron classification head. Memory slots are flattened and processed:

$$z = \text{Flatten}(M_{final}) \in \mathbb{R}^{M \times D} \tag{18}$$

$$h = \text{ReLU}(\text{LayerNorm}(z)W_1 + b_1) \in \mathbb{R}^{128} \tag{19}$$

$$\hat{y} = \text{softmax}(W_2 h + b_2) \in \mathbb{R}^2 \tag{20}$$

The classifier outputs probability scores for violent and non-violent classes, and the class with the highest probability is selected as the final prediction.

VII. Experimental Setup and Results

A. Implementation Details

The proposed RPAT framework is implemented in PyTorch 2.8+ with CUDA acceleration. The system is trained on a GPU (NVIDIA V100 or Google TPU) with the following hyperparameters:

- Learning rate: 3×10^{-4} (Adam optimizer)
- Weight decay: 10^{-5} (L2 regularization)
- Batch size: 8
- Number of epochs: 50
- CE Loss weight (w_{CE}): 1.0
- Contrastive Loss weight ($w_{contrast}$): 0.5
- Total loss: $L_{total} = w_{CE} \cdot L_{CE} + w_{contrast} \cdot L_{contrast}$

ViT-tiny backbone is initialized with ImageNet pretrained weights and optionally fine-tuned during training. MediaPipe Pose [9] operates in real-time on CPU and provides 17 keypoint detections per frame.

B. Performance Metrics

Evaluation is performed using standard metrics:

- **Accuracy**: $\frac{TP+TN}{TP+TN+FP+FN}$
- **Precision**: $\frac{TP}{TP+FP}$ (violence class)
- **Recall**: $\frac{TP}{TP+FN}$ (violence class)
- **F1-Score**: $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$
- **Area Under ROC Curve (AUROC)**: Measures classifier performance across thresholds

Additionally, per-sequence predictions are aggregated at the video level by averaging softmax probabilities across all sequences:

$$P_{video} = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i \tag{21}$$

where N is the number of sequences in the video.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

C. Experimental Results

Training Dynamics: On the Hockey Fight dataset, the model demonstrates effective learning with loss convergence across both classification and contrastive objectives. Per-epoch metrics show:

TABLE I
TRAINING METRICS OVER EPOCHS

Epoch	Train Loss	Train Acc	Val Acc
1	12.99	0.00%	52.5%
10	8.47	68.5%	71.3%
20	5.23	81.2%	78.9%
30	3.12	88.7%	82.4%
40	1.94	93.1%	84.2%
50	0.87	96.5%	85.6%

Final Validation Performance: Comprehensive evaluation on the validation set yields:

TABLE II
FINAL PERFORMANCE METRICS ON VALIDATION SET

Metric	Value
Accuracy	85.6%
Precision (Violence)	0.852
Recall (Violence)	0.848
F1-Score (Violence)	0.850
AUROC	0.912

Inference Results: On representative test videos:

- Fight video (fi100): Per-sequence confidence: [0.998, 0.997, 0.996, . . .], Video-level prediction: **Violence** ($P_{violence} = 0.9984$)
- Non-fight video (no100): Per-sequence confidence: [0.167, 0.171, 0.153, . . .], Video-level prediction: **Non-violence** ($P_{violence} = 0.1629$)

Modality Contribution Analysis: Ablation studies quantifying individual modality contributions:

TABLE III
ABLATION STUDY RESULTS

Model Variant	Accuracy
Vision Only (ViT)	78.2%
Pose Only (Heatmaps)	72.5%
Motion Only (OptFlow+Diff)	71.3%
Vision + Pose	82.7%
Vision + Motion	80.9%
Pose + Motion	79.4%
Full Model (RPAT)	85.6%

The results demonstrate that vision features contribute the highest individual accuracy (78.2%), while pose and motion provide complementary information. The full multimodal fusion achieves 7.4 percentage points improvement over vision-only baseline, validating the effectiveness of the proposed architecture.

Computational Efficiency: The proposed system achieves practical performance characteristics:

- **Latency per frame**: 50-100 ms (GPU-accelerated)
- **Throughput**: 10-20 fps (single GPU)
- **Memory footprint**: 4-8 GB (GPU)
- **Model parameters**: 24M (ViT) + 10M (remaining modules) = 34M total

VIII. DISCUSSION

A. Key Findings

The experimental results demonstrate several important insights:

Multimodal Complementarity: The 7.4 percentage point improvement of the full model over vision-only baseline indicates that pose and motion features provide significant complementary information. Pose heatmaps capture skeletal dynamics directly relevant to violent behavior [5], while optical flow reveals movement intensity and patterns that RGB features alone cannot capture.

Relational Memory Effectiveness: The relational memory module successfully captures long-term temporal dependencies. By maintaining learnable memory slots updated through attention, the system can accumulate context about escalating or repetitive aggressive behavior—a characteristic of real violence that isolated frame-level features miss.

Event Prompt Tokens: The incorporation of learnable event prompt tokens guides the transformer attention toward action-relevant features. This explicit biasing mechanism improves classification confidence in ambiguous scenarios where violent and non-violent actions exhibit similar visual characteristics (e.g., sports vs. assault).

Contrastive Alignment Benefits: The 0.5 weight on contrastive loss [10] balances classification accuracy with modality consistency. This alignment mechanism reduces spurious correlations between modalities and enforces learning of shared underlying patterns, improving model robustness to noise in individual modality extractors (e.g., pose estimation errors under occlusion).

B. Comparison with Baselines

While direct comparison with published benchmarks is limited due to dataset differences, the proposed approach addresses known limitations of existing methods:

vs. Appearance-Based CNN (2-Stream): Traditional two-stream networks [1] fuse spatial and optical flow features but remain sensitive to background clutter and illumination changes. The RPAT framework achieves superior robustness by centering on pose dynamics, which are inherently less affected by scene context.

vs. Graph-Based Skeleton Methods (ST-GCN): Skeleton-based GCN approaches [5] model spatial relationships between joints but struggle with long-term temporal dependencies. RPAT's transformer backbone and relational memory module better capture how skeletal configurations evolve and escalate over sustained violent episodes.

vs. Generic Vision Transformers: Generic ViT [4] applied to violence detection may overemphasize appearance cues. The proposed multimodal fusion with explicit pose heatmaps

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

and motion enhancement provides stronger inductive bias toward violence-relevant features.

C. Robustness Analysis

The system demonstrates improved robustness in challenging scenarios:

Occlusion Handling: Pose-driven representations remain functional even when parts of the body are occluded, as skeletal keypoints are typically robust to partial visibility. Motion features complement this by capturing movement despite occlusion.

Lighting Variations: Pose heatmaps and motion maps are largely invariant to illumination changes since they operate on grayscale or normalized features. This contrasts with appearance-based methods that may struggle in low-light surveillance footage.

Background Clutter: Focus on human pose and motion reduces sensitivity to background objects and scene changes. This is particularly valuable in crowded public spaces where multiple people and complex backgrounds are present.

IX. Limitations and Future Work

A. Current Limitations

Single-Person Assumption: The current heatmap aggregation treats all detected keypoints equally, without distinguishing between multiple individuals or modeling inter-person relationships. In crowded violence scenarios (e.g., group fights, brawls), this limits performance.

Frame Extraction Trade-off: Sampling at 8 fps reduces redundancy but may miss very rapid, short-duration violent events. The optimal frame rate is dataset and domain-dependent.

Domain Specificity: The hockey dataset, while balanced, represents a specific violence type. Generalization to street violence, domestic assault, or other contexts remains unvalidated.

Pose Estimation Sensitivity: MediaPipe Pose [9] provides lightweight, real-time detection but may degrade under extreme viewpoints, heavy occlusion, or unusual body configurations. 3D pose information would improve robustness.

Audio Information: The framework ignores audio modality. Impact sounds, screams, and verbal aggression provide important contextual cues that could improve detection.

B. Future Research Directions

Multi-Person Graph Modeling: Extend the framework to explicitly model interactions between multiple individuals using graph neural networks. Each person's skeleton would form a node, with edges capturing spatial proximity and relative motion patterns.

3D Pose Integration: Incorporate 3D skeleton estimation from depth sensors or multi-view cameras. 3D pose is more robust to viewpoint changes and occlusions, enabling better performance in crowded scenes.

Audio-Visual Fusion: Integrate audio features (spectrograms, acoustic embeddings) with video. Joint modeling of sound and motion would improve detection of verbal aggression and impact sounds.

Domain Adaptation: Develop transfer learning and domain adversarial training techniques to enable the model to generalize across different violence types and camera configurations. Meta-learning approaches could enable rapid adaptation to new domains.

Temporal Localization: Extend beyond binary classification to identify the exact temporal boundaries of violent events within videos. This requires frame-level or segment-level predictions rather than video-level aggregation.

Explainability: Develop attention visualization and saliency map techniques to highlight which body parts and motion patterns drive violence predictions. This interpretability is critical for human operators reviewing automated alerts.

Efficiency Optimization: Apply model quantization, pruning, and knowledge distillation to reduce model size and inference latency for edge deployment on resource-constrained devices (mobile phones, embedded systems).

Real-World Evaluation: Test the framework on real surveillance footage from public spaces, transportation hubs, and security-sensitive environments. This evaluation would reveal failure modes and necessary robustness improvements.

X. Conclusion

This paper proposes a Pose-Driven Relational Transformer (RPAT) framework for robust violence detection in surveillance videos. By centering on human body dynamics rather than raw appearance, the system achieves improved robustness to background clutter, occlusion, and illumination variations. The integration of pose heatmaps, motion features, and appearance cues through a multimodal fusion mechanism provides complementary information for discriminating violent behavior.

The relational memory module enables effective temporal reasoning by maintaining learnable memory slots that accumulate context across frames. This mechanism is particularly valuable for recognizing sustained or escalating aggressive patterns. Event prompt tokens and contrastive multimodal alignment further enhance classification reliability and modality consistency.

Experimental evaluation on the Hockey Fight dataset demonstrates the framework's effectiveness, achieving 85.6% validation accuracy with 0.912 AUROC. Ablation studies confirm that multimodal fusion substantially outperforms single-modality baselines. The system maintains practical computational efficiency suitable for real-time surveillance deployment.

While the current framework addresses key limitations of existing approaches [1], [4], [5], several directions warrant future investigation: multi-person interaction modeling, 3D pose integration, audio-visual fusion, and domain adaptation across different violence types. These extensions would further enhance robustness and generalizability for large-scale real-world surveillance systems.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The pose-driven approach proposed in this work provides a principled foundation for human-centered action understanding and offers promising potential for practical deployment in public safety applications. As surveillance systems increasingly integrate artificial intelligence, the ability to accurately and robustly detect violent behavior remains an important research challenge and societal need.

ACKNOWLEDGMENT

We acknowledge the contributions of all authors to this research project. This work was conducted as a part of the Computer Science and Engineering curriculum at Kakatiya Institute of Technology (KITSW), Warangal. We extend our gratitude to the Department of Computer Science and Engineering for providing the necessary computational resources and lab facilities. We thank our faculty advisor for invaluable guidance and feedback throughout the project development. We also acknowledge the Hockey Fight Dataset creators for providing the dataset used in this research and the open-source community for tools and libraries including PyTorch, MediaPipe, and TIMM that enabled this implementation. Finally, we recognize the contributions of the broader computer vision research community whose work in video understanding, pose estimation, and transformer architectures formed the foundation for this research.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 568–576.

[2] D. Tran, L. Bourdev, R. Fergus, L. Torrance, and M. Paluri, "Learning spatio-temporal features with 3D convolutional networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4694–4702.

[3] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 6202–6211.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[5] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI Conference on Artificial Intelligence*, 2018, pp. 7444–7452.

[6] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12026–12035.

[7] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13655–13664.

[8] H. Duan, Y. Zhao, K. Xiong, D. Liu, E. P. Lafortune, and Y. Wang, "Revisiting skeleton-based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 3616–3627.

[9] C. Lugaresi, Y. Tang, H. Nash, et al., "MediaPipe: A framework for building perception pipelines," arXiv preprint arXiv:1906.08172, 2019.

[10] T. Chen, S. Kornblith, M. Noroulli, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*, 2020, pp. 1597–1607.

[11] J.-B. Grill, F. Strub, A. Altche´, et al., "Bootstrap your own latent: A new approach to self-supervised learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 21929–21940.

[12] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Violence detection using spatiotemporal convolutional long short-term memory," in *IEEE International Conference on Computer and Information Technology (ICCIT)*, 2020, pp. 1–6.

[13] B. Solmaz, B. E. Moore, and M. Shah, "Identifying behaviors in crowd scenes using stability analysis," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, vol. 34, no. 10, pp. 2064–2070.

[14] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 1–6.

[15] J. C. Niebles, C. W. Chen, and L. Fei-Fei, "Unsupervised learning of visual representations using videos," in *IEEE International Conference on Computer Vision (ICCV)*, 2010, pp. 1622–1629.