Spotting the Aggressor: Pose-Based Violence Detection Through Spatial-Temporal Deep Learning Techniques

Spotting the Aggressor: Pose-Based Violence Detection Through Spatial-Temporal Deep Learning Techniques

Alessandro Rongoni
Dept. of Information Engineering (DII)
Polytechnic University of Marche
Ancona, Italy
s1114661@studenti.univpm.it

Lorenzo Longarini
Dept. of Information Engineering (DII)
Polytechnic University of Marche
Ancona, Italy
s1110740@studenti.univpm.it

Mariorosario Prist
Dept. of Information Engineering (DII)
Polytechnic University of Marche
Ancona, Italy
m.prist@staff.univpm.it

Geremia Pompei
Dept of Computer Science
University of Pisa
Pisa, Italy
geremia.pompei@di.unipi.it

Aldo F. Dragoni
Dept. of Information Engineering (DII)
Polytechnic University of Marche
Ancona, Italy
a.f.dragoni@univpm.it

***Abstract*—The automatic detection of violent behaviour in video sequences has emerged as a critical area of research in public safety and surveillance, as the early detection of aggressive actions enables rapid intervention and can significantly mitigate potential harm. This paper proposes a novel methodology for the automatic detection and identification of violent behaviours in video sequences, with particular emphasis on the recognition of the specific individual responsible for such actions. A significant innovation of our approach is the integrated extraction of human pose features using a You Only Look Once-based (YOLO) model that efficiently captures critical key points which serve as essential cues for the detection of violent interactions. The proposed approach integrates human pose estimation techniques used to extract spatial features with temporal analysis models designed to capture the dynamic nature of aggressive behaviour. To assess the effectiveness of the method, two temporal architectures, a Bidirectional Long-Short-Term Memory (BiLSTM) network and a transformer-based model, were evaluated on the AirtLab dataset. Experimental results demonstrate the robustness and reliability of the proposed approach, highlighting high accuracy alongside real-time applicability. Furthermore, by relying on pose-based representations that can be processed in distributed edge-cloud architectures, the methodology offers enhanced privacy preservation compared to raw video processing approaches.**

***Index Terms*—Violence detection, Pose estimation, You Only Look Once, Bidirectional Long Short-Term Memory, Transformer**

I. Introduction

In recent years, the automatic detection of violent behaviour in video sequences has become a key research topic in Computer Vision (CV) and Deep Learning (DL), driven by the need for enhanced security and prompt intervention in various environments. The challenge is particularly significant in surveillance applications, where systems must operate reliably in complex, crowded environments with varying lighting conditions and camera perspectives. Several studies have combined spatial analysis via Convolutional Neural Networks (CNNs) [1]–[3] with temporal models like Recurrent Neural Networks (RNNs) and Long-Short-Term Memory (LSTM) [4], [5] to identify violent actions. Notable contributions include CNN-based approaches for motion picture content rating [6], hybrid models integrating CNNs with recurrent networks [7] and systems that leverage pose estimation to capture human dynamics [8], [9].

Traditional approaches often struggle with real-world challenges, especially when relying on RGB data alone, which can be affected by environmental factors [10], [11]. CNN-based methods also require high computational resources, raise privacy concerns due to raw video processing, and offer limited interpretability for security applications. More recent methods have explored pose-based representations as a more robust alternative, focusing on human skeletal information rather than relying solely on appearance features [12], [13].

Human pose estimation has evolved significantly, from part affinity field and high-resolution representation methods [14] to streamlined, unified architectures such as YOLO-Pose, which fuse object detection and pose estimation in a single model [15]. Recent studies have demonstrated the effectiveness of transformer architectures for video understanding tasks [16], [17], offering advantages in modeling long-range dependencies compared to recurrent architectures.

However, many existing methods still face significant challenges:

- Poor generalization to real-world scenarios, particularly when trained on synthetic or highly controlled datasets [18], [19].
- High sensitivity to environmental variations such as lighting, camera angle, and occlusions.

- Limited interpretability of features, making it difficult to understand model decisions.
- Computational inefficiency when processing multiple subjects in crowded scenes.
- Difficulties in capturing the nuanced temporal dynamics that distinguish violent from non-violent interactions.

To address these challenges, this paper proposes a novel methodology that integrates spatial features and temporal dynamics to effectively recognize violent behaviour and distinguish types of aggressive actions. Our approach differs fundamentally from previous work in its focus on structured pose-based representations and comparative analysis of advanced temporal modeling techniques.

The proposed pipeline begins with spatial feature extraction using YOLO-based pose estimation, providing efficient and accurate detection of human pose keypoints even in complex scenes. The extracted poses follow the COCO keypoint format [20], which defines 17 anatomical joints and enables structured representation of human motion. The proposed feature engineering component transforms raw pose data into structured representations that capture crucial indicators of violence, including joint velocities, relative body positioning, and dynamic postural changes.

For the temporal phase, we explore and compare two architectures: a Bidirectional Long Short-Term Memory (BiLSTM) network [21], which is well-suited for modeling sequential dependencies, and a Transformer-based model [22], known for its powerful capabilities in capturing long-range temporal relationships. This comparative analysis provides valuable insights into the trade-offs between different temporal modeling approaches for violence detection tasks.

The key contributions are:

- A tailored feature engineering pipeline that converts pose keypoints into informative spatio-temporal descriptors for violence detection.
- A novel strategy for capturing interpersonal dynamics via relative positioning and motion patterns between subjects.
- A comparative evaluation of BiLSTM and Transformer models for temporal reasoning in violent behaviour recognition.
- Experimental evidence showing superior performance over existing RGB-based and hybrid approaches while preserving privacy and computational efficiency.

Among the key advantages of our methodology are its robustness in complex, crowded scenes, its adaptability to different types of temporal models, and its reliance on pose-based descriptors, which enhance interpretability and reduce dependence on raw video input. Unlike CNN-based approaches that process raw video data, the proposed pose-based method offers significant improvements in privacy preservation, environmental robustness, and computational efficiency while maintaining comparable or superior detection accuracy.

The paper is structured as follows: Section 2 details the proposed methodology, with an overview of the AirtLab dataset, the YOLO-based pose extraction and labelling process, and the feature engineering pipeline. Section 3 presents the experimental evaluation of the BiLSTM and Transformer models, analyzing their performance and trade-offs. Finally, Section 4 offers conclusions and directions for future research, including potential applications in surveillance and behaviour analysis systems.

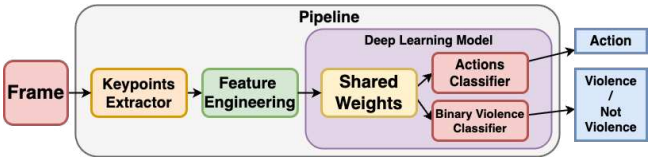


Fig. 1: Overview of the proposed pipeline.

II. Materials and Methods

This section outlines our privacy-preserving methodology for violence detection using skeletal pose data. We cover the AirtLab dataset, YOLO11m-based keypoint extraction, feature engineering for temporal representation, and neural network architectures (BiLSTM, Transformer). Figure 1 illustrates the complete pipeline from video input to classification output. In detail, from raw video input, human pose keypoints are extracted using YOLO11m. These are then processed into structured features capturing pose dynamics and interactions, which are fed into temporal models for violence and action recognition.

A. Dataset

This study employs the AirtLab *Dataset for Automatic Violence Detection in Videos*, containing 350 video sequences (230 violent, 120 non-violent) recorded at 1280×720 pixels and 30 FPS. Video duration ranges from 2-10 seconds (average 4.5s), captured in controlled indoor environments with varying lighting conditions and camera perspectives. The dataset was stratified into training (50%), validation (20%), and testing (30%) splits, maintaining proportional representation of behavior categories across partitions. Table I summarizes the types of behaviors included in the dataset, grouped into violent and non-violent categories.

B. Keypoint Extraction and Labeling

The YOLO 11m pose model (mAP: 0.87) was utilized for its balance between accuracy and computational efficiency. The processing pipeline consisted of two phases:

1) All videos were processed with YOLO 11m to generate JSON files containing frame-by-frame detections. Each detected person received a unique ID through an enhanced Hungarian tracking algorithm that integrated

TABLE I: Violent and Non-Violent Behaviours.

Behaviour Type	Examples
Violent	Choke, Club, Fight, Gunshot, Kick Punch, Push, Slap, Stab
Non-Violent	Greet, Hand gestures, Handshake, High five Hug, Jump, Walk, Friendly-Punch

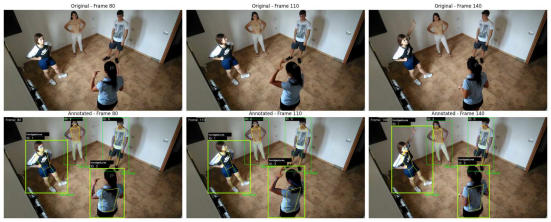


Fig. 2: Comparison between original frames (top) from non-violent videos and the same frames with keypoints overlaid (bottom).

Fig. 3: Comparison between original frames (top) from violent videos and the same frames with keypoints overlaid (bottom).

spatial proximity, pose similarity, and Intersection over Union (IoU) between consecutive frames.

2) A custom labeling tool enabled manual annotation of each tracked person with their behavioral status (violent/non-violent), role (aggressor/victim), and specific action performed, achieving a tracking consistency rate of 93.8%.

The model extracted 17 anatomical keypoints in COCO format: head region (5 points), upper body (6 points), and lower body (6 points). Each keypoint was represented by normalized (x,y) coordinates relative to bounding box dimensions and a confidence score, with points below 0.4 confidence handled through interpolation.

C. Feature Engineering and Data Preparation

A comprehensive feature extraction pipeline transforms raw pose keypoints into meaningful behavioral descriptors across three key aspects: individual body dynamics, interpersonal interactions, and temporal patterns.

1) Individual Pose Features:

- Joint velocities: Frame-to-frame movement speed of body parts (e.g., rapid hand movements during punching), computed over 5-frame windows.
- Joint accelerations: Sudden changes in movement speed indicating abrupt actions, calculated as second-order temporal derivatives.
- Postural angles: Body posture indicators such as arm angles and torso orientation, distinguishing aggressive from defensive stances.
- Normalized displacement: Movement patterns relative to center of mass, ensuring scale-invariant features.

2) Interaction Features:

- Inter-person distances: Physical proximity between people, critical for detecting contact-based violence.
- Approach velocities: Speed of movement toward or away from others, indicating aggressive approach or defensive retreat.
- Relative orientations: Whether people face each other (confrontational) or turn away (avoidance).
- Joint proximity alerts: Binary indicators when body parts come within striking distance of another person.

3) Temporal Context Features:

- Moving averages: Smoothed motion patterns over 10, 20, and 30-frame windows to distinguish sustained aggressive behavior from brief gestures.
- Peak detection: Identification of sudden spikes in velocity/acceleration during violent actions.
- State transitions: Detection of behavioral changes, such as calm-to-aggressive posture transitions.

The pipeline generates a 1390-dimensional vector per frame, combining spatial pose information with temporal dynamics and interpersonal relationships. The final dataset contains 1326 unique samples (video-person combinations), standardized to 150-frame sequences for consistent temporal modeling input.

D. Privacy-Preserving Deployment Architecture

While the complete pipeline processes raw video for evaluation purposes, the pose-based approach enables a distributed deployment architecture that enhances privacy preservation compared to CNN-based methods. In practical implementations, YOLO-based pose estimation runs locally on edge devices, extracting only skeletal keypoints without storing or transmitting raw video data. Only the extracted pose coordinates (34 numerical values per person per frame) are transmitted to cloud infrastructure for temporal analysis and violence detection. This architecture provides enhanced privacy compared to CNN-based approaches requiring full frame transmission, as pose coordinates contain significantly less identifiable information than raw video data. While not completely anonymous, the skeletal representation reduces identifiability compared to facial features, clothing, or environmental context present in full video frames.

E. Models Architecture and Evaluation

Two complementary architectures were implemented:

1) BiLSTM Architecture: The BiLSTM architecture processes pose sequences of shape (150, 1390) through two stacked BiLSTM layers (128 units each), followed by fully connected layers (64 and 32 units) with dual output heads for violence detection and action classification. The model contains 1,970,067 trainable parameters.

2) Transformer Architecture: The Transformer architecture implements two transformer encoder blocks with 8 attention heads and an embedding dimension of 256, including positional encoding followed by global average pooling and dense layers (128 and 64 units). The model contains 1,453,075 trainable parameters.

TABLE II: Performance comparison between models.

Model	Acc. (%)	Prec. (%)	Rec. (%)	F1-score (%)
BiLSTM	97.05	98.31	92.06	95.08
Transformer	98.28	94.74	100.00	97.30

TABLE III: Confusion matrices for violence detection showing excellent classification performance with minimal misclassifications. **V** indicates Violence and **NV** indicates Non-Violence.

(a) BiLSTM model

True/Pred	NV	V
NV	279	2
V	10	116

(b) Transformer model

True/Pred	NV	V
NV	274	7
V	0	126

3) Training and Evaluation: Both models were trained using the Adam optimizer (initial learning rate: 0.001) with learning rate reduction and early stopping. Loss functions included binary cross-entropy for violence detection and categorical cross-entropy for action classification, weighted at 1.0 and 0.5 respectively [23]. The total loss function can be expressed as:

$$\text{Loss} = 1.0 \times \text{VDLoss} + 0.5 \times \text{ACLoss} \quad (1)$$

where VDLoss represents the violence detection loss and ACLoss represents the action classification loss. This weighting strategy prioritizes the primary task of violence detection while still incorporating action classification as a secondary objective. Performance evaluation utilized accuracy, precision, recall, and F1-score. Both models were evaluated on identical data splits, with testing performed on 407 held-out samples (281 non-violent, 126 violent sequences).

III. Results

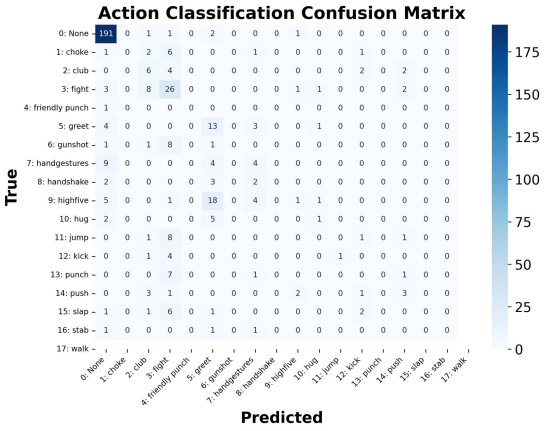
This section presents the experimental evaluation of BiLSTM and transformer-based models on the AirtLab dataset using a standard split of 50% training, 20% validation, and 30% testing.

A. Violence Detection Performance

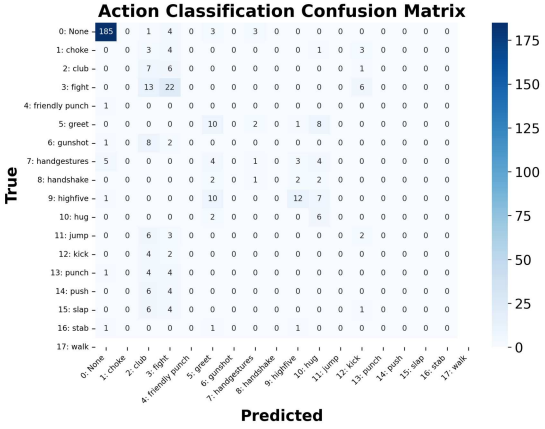
Both models achieved exceptional performance in binary violence detection (Table II). The Transformer model achieved 98.28% accuracy compared to BiLSTM's 97.05%, with perfect recall (100% vs 92.06%) but lower precision (94.74% vs 98.31%). The confusion matrices (Table III) show minimal misclassifications: BiLSTM produced 10 false negatives and 2 false positives, while Transformer had 0 false negatives but 7 false positives.

B. Action Classification Performance

Action classification proved more challenging, with moderate performance (Table IV). BiLSTM achieved slightly higher accuracy (60.20% vs 59.71%), while Transformer demonstrated better precision (57.42% vs 51.34%) and F1-score (57.42% vs 54.32%). Both models performed well on the dominant "None" class (94-97% recall) but struggled with rare action classes. Misclassifications typically occurred within semantically similar action groups (Figure 4).



(a) BiLSTM model



(b) Transformer model

Fig. 4: Confusion matrices for action classification showing strong performance on the dominant class (None) and semantic grouping of errors.

C. Comparison with State-of-the-Art Methods

Table V compares the pose-based approach with existing CNN methods, demonstrating superior performance with potential privacy advantages through distributed deployment architectures.

D. Training Dynamics

Training analysis (Figure 5) shows distinct convergence patterns: Transformer exhibits rapid initial improvement then stabilization, while BiLSTM shows gradual, consistent convergence. Violence detection achieves higher validation accuracy (95-98%) than action classification (55-60%), highlighting the multi-class task complexity. Both models show overfitting in action classification after 7 epochs.

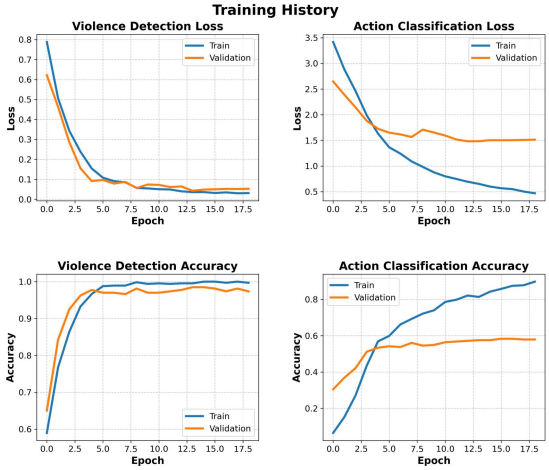
TABLE IV: Performance comparison for action classification.

Model	Acc. (%)	Prec. (%)	Rec. (%)	F1-score (%)
BiLSTM	60.20	51.34	60.20	54.32
Transformer	59.71	57.42	59.71	57.42

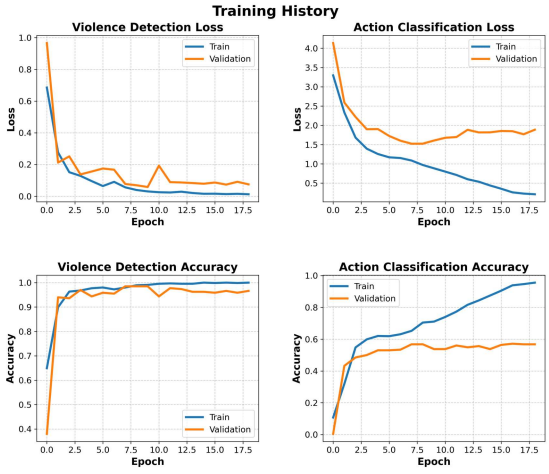
TABLE V: Performance comparison with existing violence detection methods (Priv stands for Privacy)

Method	Approach	Dataset	Acc.%	Priv.
Gruosso et al.	CNN	Custom movies	90.9	No
Senst et al.	Lagr.+SVM	Violent Crowd	94.4	No
Sudhakaran et al.	Conv-LSTM	Hockey Fight	97.1	No
Our BiLSTM	**Pose-based**	**AirtLab**	**97.05**	**Yes***
Our Transformer	**Pose-based**	**AirtLab**	**98.28**	**Yes***

*Enhanced privacy through distributed edge-cloud deployment



(a) BiLSTM model



(b) Transformer model

Fig. 5: Training and validation history showing convergence patterns for both models across epochs.

E. Feature Ablation Study

To assess the contribution of different feature groups, a systematic ablation study was conducted by removing specific feature categories and evaluating performance impact. Table VI presents the key findings. The ablation study reveals a clear feature hierarchy for both architectures. **Individual Pose Features** (**IP-Only**) emerged as most critical, with their removal (**No Individual Pose**, **No-IP**) causing dramatic performance degradation (from 97% to 46-53% F1 score), confirming that basic keypoint coordinates form the foundation of violence detection. Remarkably, **IP-Only** with just 306 dimensions achieves superior performance compared to the full feature set, demonstrating that core skeletal information is sufficient for effective detection. Removing **Interaction Features** (**No-IF**) slightly improved performance, suggesting redundancy in interpersonal calculations. **Temporal Context Features** (**No-TC**) showed moderate importance while **Motion Dynamics** (**No-MD**) appeared largely redundant. This analysis demonstrates that violence detection relies primarily on individual body configuration rather than complex interpersonal relationships, providing clear guidance for computational optimization in resource-constrained environments.

TABLE VI: Ablation Study Results: Impact of Feature Groups on Performance

Config	Features	Violence F1 (%)		Action F1 (%)	
		BiLSTM	Trans.	BiLSTM	Trans.
Baseline	1390	96.83	96.83	55.58	59.27
IP-Only	306	98.40	98.80	55.08	62.04
No-IP	1084	46.38	53.28	36.58	36.37
No-IF	896	98.80	98.81	55.11	58.13
No-MD	1200	96.83	98.80	53.15	57.73
No-TC	990	95.55	98.81	51.43	57.36

IV. Discussions

The experimental results confirm the effectiveness of the pose-based approach, demonstrating superior performance and practical advantages:

Performance and Comparison: Skeletal pose data achieves high accuracy (97–98%) for violence detection, demonstrating superior performance compared to CNN-based methods like Gruosso et al. (90.9%) and Sudhakaran et al. (94.6%) while enabling enhanced privacy preservation through distributed deployment architectures. The complementary architectures provide deployment flexibility: Transformer models achieve perfect recall (100%) for high-security scenarios, while BiLSTM offers higher precision (98.31%) when false alarms must be minimized. **Interpretability and Optimization:** The ablation study reveals that IP-Only are most critical, with their removal (No-IP) causing dramatic performance degradation. Using IP-Only with just 306 dimensions achieves near-optimal performance, enabling significant computational optimization. No-IF proved redundant, while No-TC showed moderate importance and No-MD appeared largely unnecessary. This demonstrates that violence detection relies primarily on individual body configuration, enhancing interpretability compared to black-box CNN methods. **Limitations and Privacy Considerations:** The controlled indoor evaluation (350 videos) may limit real-world generalizability, and the method depends on pose estimation accuracy. The pose-based architecture enables distributed deployment where pose estimation occurs locally and only skeletal coordinates are transmitted, reducing data sensitivity compared to full

video processing. However, privacy enhancement is relative-skeletal representations reduce but do not eliminate identifying information. Moderate action classification performance (60%) reflects class imbalance issues, highlighting the need for specialized augmentation strategies.

Overall, this keypoint-based framework balances accuracy, efficiency, and enhanced privacy preservation potential, positioning it as a viable solution for surveillance deployment while establishing clear directions for future enhancement.

V. Conclusions

This study establishes pose-based violence detection as a viable and effective alternative to CNN approaches, demonstrating competitive performance while providing privacy preservation potential through distributed deployment architectures that process skeletal data locally. The comprehensive ablation study reveals that individual pose features are fundamental to violence detection, while complex interpersonal and motion dynamics prove largely redundant, enabling computational optimization. The systematic evaluation demonstrates strong performance compared to state-of-the-art methods. Unlike black-box CNN methods, the proposed approach reveals biologically plausible patterns where individual body configuration emerges as the primary violence indicator, enhancing system interpretability and trustworthiness for security applications. The complementary architectures offer flexible deployment options with different precision-recall trade-offs. However, important limitations must be acknowledged. The evaluation remains constrained to controlled indoor environments with a relatively small dataset, raising concerns about real-world generalizability. Action classification performance remains modest due to class imbalance issues, and the approach's dependence on accurate pose estimation may introduce vulnerabilities in crowded or occluded scenes. Additionally, privacy enhancement is relative-skeletal coordinates reduce but do not eliminate identifying information compared to raw video processing. Despite these limitations, this pose-based framework establishes a new paradigm for surveillance deployment that balances detection performance with enhanced privacy preservation potential. Future research should prioritize validation across diverse outdoor datasets, targeted data augmentation strategies for class imbalance, automated annotation pipelines for scalability, and robustness evaluation with uncertainty quantification for degraded pose estimation scenarios.

References

[1] L. R. Medsker, L. Jain *et al.*, "Recurrent neural networks," *Design and Applications*, vol. 5, no. 64-67, p. 2, 2001.

[2] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[3] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 12, pp. 6999–7019, 2021.

[4] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo, and A. F. Dragoni, "Deep learning for automatic violence detection: Tests on the airtlab dataset," *IEEE Access*, vol. 9, pp. 160 580–160 595, 2021.

[5] P. Zhang, L. Dong, X. Zhao, W. Lei, and W. Zhang, "An end-to-end framework for real-time violent behavior detection based on 2d cnns," *Journal of Real-Time Image Processing*, vol. 21, no. 2, p. 57, 2024. [Online]. Available: https://doi.org/10.1007/s11554-024-01443-7

[6] M. Gruosso, N. Capece, U. Erra, and N. Lopardo, "A deep learning approach for the motion picture content rating," in *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2019, pp. 137–142.

[7] N. Jain, V. Gupta, U. Tariq, and D. J. Hemanth, "Fast violence recognition in video surveillance by integrating object detection and conv-lstm," *International Journal on Artificial Intelligence Tools*, vol. 32, no. 03, p. 2340018, 2023. [Online]. Available: https://doi.org/10.1142/S0218213023400183

[8] S. Deshmukh, D. Mistry, S. Joshi, and C. Bhole, "Sentinel eyes violence detection system," in *Proceedings of International Conference on Communication and Computational Technologies*, S. Kumar, S. Hiranwal, R. Garg, and S. D. Purohit, Eds. Singapore: Springer Nature Singapore, 2024, pp. 321–333.

[9] Üstek, J. Desai, I. López Torrecillas, S. Abadou, J. Wang, Q. Fever, S. R. Kasthuri, Y. Xing, W. Guo, and A. Tsourdos, "Two-stage violence detection using vitpose and classification models at smart airports," 2023. [Online]. Available: https://arxiv.org/abs/2308.16325

[10] T. Senst, V. Eiselein, A. Kuhn, and T. Sikora, "Crowd violence detection using global motion-compensated lagrangian features and scale-sensitive video-level representation," *IEEE transactions on information forensics and security*, vol. 12, no. 12, pp. 2945–2956, 2017.

[11] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2017, pp. 1–6.

[12] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[13] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2334–2343.

[14] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[15] D. Maji, S. Nagori, M. Mathew, and D. Poddar, "Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2637–2646.

[16] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 244–253.

[17] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6836–6846.

[18] P. Kumar, G.-L. Shih, B.-L. Guo, S. K. Nagi, Y. C. Manie, C.-K. Yao, M. A. Arockiyadoss, and P.-C. Peng, "Enhancing smart city safety and utilizing ai expert systems for violence detection," *Future Internet*, vol. 16, no. 2, 2024. [Online]. Available: https://www.mdpi.com/1999-5903/16/2/50

[19] M. Bianculli, N. Falcionelli, P. Sernani, S. Tomassini, P. Contardo, M. Lombardi, and A. F. Dragoni, "A dataset for automatic violence detection in videos," *Data in Brief*, vol. 33, p. 106587, 2020.

[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," *ECCV*, 2014.

[21] A. Graves and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," *Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005*, pp. 799–804, 2005.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[23] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, 2020.