

Spatiotemporal Action Recognition in Low-Altitude Aerial Monitoring: A Multi-Stage YOLOv8–ByteTrack–R3D Architecture

Sneha Tandon^{1*}, Dongbing Gu²

^{1,2} School of Computer Science and Electric Engineering, University of Essex, Colchester, United Kingdom

*Corresponding Author Email: 5688sneha8865@gmail.com

Abstract

Real-time anomaly and violent action detection from low-altitude aerial video feeds presents many computational challenges as a result of perspective distortion, camera motion, and the limited processing power available on-board. Frame-level 2D convolutional neural networks (CNNs) have a high false positive rate as a direct consequence of their inability to model the temporal dependencies that exist between consecutive frames of video. In an attempt to reduce these limitations, this research evaluates a decoupled, multi-stage spatiotemporal deep learning pipeline that was developed specifically to be executed on localised edge computing platforms. The initial step in the pipeline involves localizing humans in each frame of video using an ultra-lightweight YOLOv8-Nano detector; the relationship between humans in different frames is established through a multi-object tracking (MOT) ByteTrack algorithm. After establishing relationships between humans across multiple frames of video, action dynamics are classified using a spatiotemporal ResNet-3D (R3D-18) network which uses a sliding temporal window of frames. Evaluation of the proposed pipeline demonstrated that it achieved 91.30% classification accuracy and an F1-score of 0.91 when evaluated against real-world aerial surveillance and battle datasets. Additionally, prototype testing demonstrated an approximate processing throughput of 20 frames per second, while the spatiotemporal model achieved a 10.80 percentage-point improvement in classification accuracy compared with the frame-level 2D CNN baseline.

Keywords

Action Recognition, ByteTrack, Computer Vision, Edge Intelligence, Spatiotemporal Modeling, UAV Surveillance.

INTRODUCTION

Unmanned Aerial Vehicles (UAVs) offer important advantages in dynamic monitoring scenarios, including agile deployment and varied visual vantage points [1]. However, obtaining behavioural knowledge from aerial feeds presents major technical challenges not seen in fixed-camera terrestrial monitoring. These issues include high platform ego-motion, severe geometric viewpoint shifts, and small human targets that take up only a fraction of the pixel space [2].

While classic object detection networks excel at spatial localisation, detecting complex behaviors—such as interpersonal aggression or aggressive anomalies—requires a knowledge of temporal kinematics. Standard 2D Convolutional Neural Networks (CNNs) process frames separately. As a result, they fundamentally fail to record the motion vectors required to discriminate between a benign quick movement (such as a celebration gesture or athletic play) and a violent action.

To close this gap, this work presents an empirical assessment of a multi-stage spatiotemporal design. The method separates individual behavioural trajectories in real time by cascading a single-stage object detector (YOLOv8-Nano), a robust data association tracker (ByteTrack), and a 3D convolutional sequence model (R3D-18). This study

intends to quantify the performance improvements of spatiotemporal modelling over traditional frame-level inference under tight edge-computing resource limitations by assessing the pipeline on well-known open-source datasets such as VisDrone, UAV-Human, and RWF-2000 [3] [4] [5] [6].

This work does not introduce a new detector, tracker, or spatiotemporal network on its own. Instead, it focuses on integrating these components for use in aerial surveillance where resources are limited. While most studies treat human detection, multi-object tracking, and action recognition as separate problems, our approach brings together lightweight YOLOv8-Nano for person localisation, ByteTrack for tracking over time, and short-window R3D-18 for behaviour classification in one processing pipeline. The system is built to handle the specific challenges of aerial monitoring, such as small human targets, moving cameras, changing viewpoints, unclear timing, and limited computing power. The key experimental goal is to see if decoupling spatial detection and tracking from temporal behavior modeling enhances violence recognition over frame-level inference while retaining near-real-time processing.

MATERIALS & METHODS

The suggested architecture functions as a sequential, multi-threaded perception pipeline that separates spatial

localization and temporal action classification. To handle the high computational load of volumetric tensor processing, model training and optimization were performed during development in a Python-based, GPU-accelerated deep learning setting.

Datasets and Preprocessing

Model development made use of publicly available datasets that covered various stages of the computer vision pipeline. VisDrone and VisDrone-MOT were used to train and evaluate aerial object detection and multi-object tracking in high-altitude environments [3]. The Real-Life Violence Situations Dataset and RWF-2000 were used to establish a baseline for behavioural categorisation [4] [7]. The dataset was split 80/10/10 for training, validation, and testing. To enable spatiotemporal convolution, video streams were pre-processed, scaled, and normalised while maintaining aspect ratios to prevent geometric distortion before being partitioned into sequential short-window frame buffers.

Object Detection and Localization

The initial operational stage leverages YOLOv8-Nano, which was chosen for its optimal balance of inference speed and parameter efficiency [8]. Traditional anchor-based detectors struggle with aerial perspectives since their bounding box priors are primarily designed for terrestrial views. YOLOv8 uses an anchor-free, center-based design that can improve the detection of small human targets. To reduce overfitting and replicate the significant viewpoint changes associated with UAV operation, numerous data augmentation procedures were used, such as horizontal flipping, dynamic brightness modification, random cropping, and scaling transformations. Training was done iteratively, with both automatically chosen optimization settings and deliberate SGD-based fine-tuning. The original YOLOv8-Nano trials had a learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005, while the latter recovery and fine-tuning stages employed SGD with a lower initial learning rate of 0.001. The input resolution and batch size were changed between experiments to evaluate the trade-off between small-target representation, training stability, and processing needs.

Multi-Object Tracking Data Association

To transition from spatial detection to temporal tracking, the ByteTrack algorithm was employed [9]. Most tracking-by-detection algorithms aggressively remove bounding boxes that fall below a predetermined confidence level. ByteTrack

mitigates this by linking all detection boxes, even those with low scores, and forecasting target trajectories across subsequent frames using a Kalman filter. The Hungarian assignment technique for bounding-box association has a worst-case computational cost of $O(n^3)$, where n is the number of candidate detections. Because the envisioned UAV surveillance situations typically involve a small number of simultaneously tracked subjects, the practical association overhead is manageable for near-real-time processing.

Spatiotemporal Sequence Modeling

Tracking sequences are extracted and concatenated into fixed-length temporal frame sequences. These sequential tensors are fed into an R3D-18 (3D ResNet) network [10] [11]. Unlike 2D CNNs, which only convolve over the X and Y spatial dimensions, the R3D-18 network employs 3D convolutional kernels across the time axis [10] [11]. As the tensor moves over the residual blocks, the network generates spatiotemporal feature maps that encode velocity, joint displacement, and the target's structural interaction over time. The final fully connected layer generates a binary classification distribution (Violent vs. Non-Violent) based on past motion rather than a static anomaly.

Model development was accomplished through a series of controlled training and fine-tuning tests rather than a single fixed training run. For the YOLOv8-based detection stage, trials were carried out with input resolutions of 512, 640, and 960 pixels, batch sizes ranging from 1 to 16, and training schedules ranging from 10 to 100 epochs. The main YOLOv8-Nano studies employed an initial learning rate of 0.01, momentum of 0.937, weight decay of 0.0005, and an early-stopping patience of 100 epochs. During development, both CPU and GPU execution were used, with GPU tests carried out in a CUDA-enabled environment running CUDA 12.2 and cuDNN 9.15.1.

Additional fine-tuning experiments were carried out with SGD at a lower initial learning rate of 0.001, notably during the recovery and higher-resolution optimization stages. To meet the computational requirements of the training settings, these experiments used 512- or 640-pixel inputs and batch sizes of 1-2. Model checkpoints were kept between selected stages, allowing training to resume from previously optimized weights rather than repeatedly initializing the detector with pretrained YOLOv8-Nano weights. This iterative process allowed us to assess the impact of input resolution, batch size, training period, and optimization settings while remaining practical in terms of computational requirements.

Table 1. Principal YOLOv8 training and optimisation configurations.

Training Stage	Model / checkpoint	Epochs	Batch	Image size	Optimizer	Initial LR	Momentum	Weight decay	Device
Initial YOLOv8n experiments	YOLOv8n pretrained	50	16	640	Auto	0.01	0.937	0.0005	CPU/GPU
Extended YOLOv8n training	YOLOv8n pretrained	100	16	640	Auto	0.01	0.937	0.0005	CPU

Training Stage	Model / checkpoint	Epochs	Batch	Image size	Optimizer	Initial LR	Momentum	Weight decay	Device
High-resolution experiment	YOLOv8n pretrained	50	8	960	Auto	0.01	0.937	0.0005	CPU
512-pixel aerial optimisation	YOLOv8n / resumed checkpoint	50	4	512	Auto	0.01	0.937	0.0005	GPU
Recovery/fine-tuning stage	Previous best checkpoint	21	1	512	SGD	0.001	0.937	0.0005	GPU
640-pixel fine-tuning	Previous best checkpoint	22	2	640	SGD	0.001	0.937	0.0005	GPU
Continued 640-pixel fine-tuning	Previous best checkpoint	10	2	640	SGD	0.001	0.937	0.0005	GPU

The GPU tests used CUDA 12.2 and cuDNN 9.15.1. The table shows example configurations from the iterative optimization process, rather than implying that all tests were part of a single continuous training run.

RESULTS AND DISCUSSION

Detection and Tracking Benchmarks

To establish a baseline of efficacy, the object detection backbone (YOLOv8-Nano) was tested utilising aerial-perspective data. The model had a mean Average Precision at 50% Intersection over Union (mAP@50) of 0.487 and a precision score of 0.691. The conservative recall measure of 0.435 emphasises the inherent difficulty of spotting heavily occluded or distant human targets from aerial perspectives. The detection and tracking stages, including video decoding, YOLO inference, and ByteTrack association, demonstrated an approximate processing throughput of 20 frames per second during prototype testing.

An incremental ablation study was carried out to optimise the YOLOv8-Nano detector for the extreme perspective distortions that occur during UAV surveillance. As seen in Figure 1, the model underwent multi-stage tuning at various resolution scales. The baseline model, initialised at a resolution of 512 pixels, achieved an mAP@50 of 0.437. After a recovery checkpoint was added and the input tensor resolution was raised to 640 pixels to capture minute pedestrian details, the model's accuracy considerably increased. The final adjusted configuration achieved a precision score of 0.691 and a peak mAP@50 of 0.487. The significance of high-resolution feature mapping for converting terrestrial anchor-free models to low-altitude aerial domains is highlighted by this optimization sequence.

An incremental optimization study was carried out to investigate the impact of training configuration and input resolution on aerial person detection. The initial YOLOv8-Nano tests were mostly done at 640-pixel resolution, although later experiments used 512- and 960-pixel inputs. During the final stages of recovery, previously established checkpoints were fine-tuned with SGD at a lower initial learning rate of

0.001 and in smaller batch sizes. The change from 512-pixel recovery to 640-pixel fine-tuning was designed to improve the representation of small and distant human targets while keeping the learnt detector weights. Figure 1 depicts the resulting optimization progression.

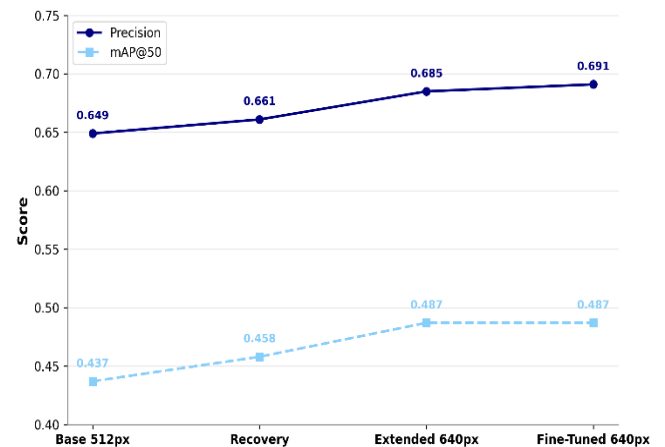


Figure 1. YOLOv8-Nano optimization progression tracking Precision and mAP@50 improvements across training stages

To ensure reproducibility, training configurations and model checkpoints were saved in experiment-specific configuration files. These data include the training period, batch size, image resolution, optimizer selection, learning-rate settings, momentum, weight decay, early-stopping patience, and execution device for each experiment.

Behavioural Classification Performance

To directly measure the impact of temporal modelling, the R3D-18 sequence model was compared to a normal frame-level 2D baseline network. The performance metrics are summarised in Figure 2.

The frame-level baseline showed a high false-negative rate of 0.3100. This suggests that static, single-frame feature extraction typically fails to identify aggressive interactions that occur dynamically across time. Because 2D networks lack temporal memory, they struggle to contextualise fast

kinematic changes.

In contrast, the R3D-18 spatiotemporal model obtained an optimised test accuracy of 91.30%. The evaluation confusion matrix for the R3D-18 model produced 20 True Positives, 22 True Negatives, 1 False Positive, and 3 False Negative.

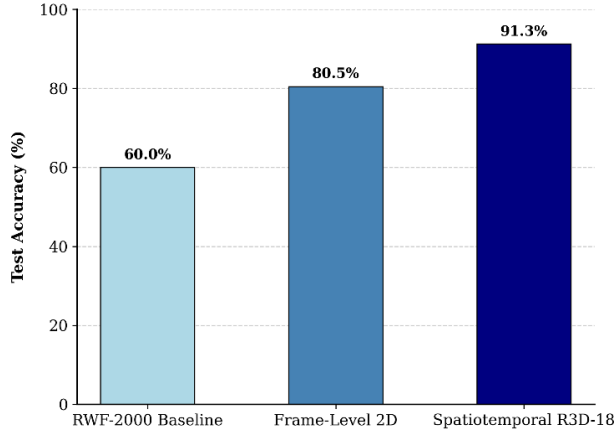


Figure 2. Comparison of reported classification accuracies for the frame-level 2D CNN baseline, R3D-18 spatiotemporal model, and RWF-2000 reference benchmark

The quantitative comparison highlights the advantages of temporal modeling. The frame-level violence classifier had a test accuracy of 80.50% and an F1-score of 0.7797. Although its violence-class precision was 0.8961, recall was lower at 0.6900, resulting in a false-negative rate of 0.3100. The confusion matrix had 69 true positives, 92 true negatives, 8 false positives, and 31 false negatives, demonstrating that frame-level inference was moderately conservative but frequently overlooked violent action when discriminative motion was not seen inside an individual frame.

On the other hand, R3D-18 obtained a test accuracy of 91.30%, with macro-averaged precision, recall, and F1-scores of around 0.91. In the battle class, precision was 0.95, recall was 0.87, and the F1-score was 0.91. The R3D-18 confusion matrix consisted of 20 true positives, 22 true negatives, one false positive, and three false negatives.

Although the two evaluations have different test-set sizes and should not be compared solely based on raw confusion-matrix counts, the accuracy and class-balanced metrics show that including temporal information produced more reliable behavioral discrimination than frame-level reasoning.

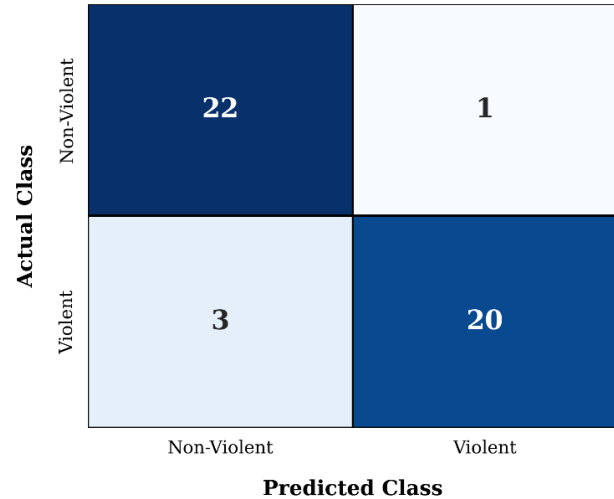


Figure 3. Confusion matrix for the R3D-18 spatiotemporal model on the test split, illustrating high true-positive retention and minimal false alarms

COMPARISON WITH CONTEMPORARY SPATIOTEMPORAL ARCHITECTURES

The following literature review contextualizes the use of R3D-18 as a practical compromise between temporal modeling capability and computing expense. Representative modern architectures include the efficient X3D model, the dual-pathway SlowFast architecture, the attention-based TimeSformer, and the Video Swin Transformer [12] [13] [14] [15]. These methods are not considered direct experimental baselines because they were tested using distinct datasets and experimental protocols.

Table 2. A qualitative assessment of representative spatiotemporal action recognition architectures and their applicability to resource-constrained UAV deployment

Model	Temporal modelling approach	Representative reported result	Relevance to UAV/edge deployment
R3D-18 (this work)	Short-window 3D CNN	91.30% accuracy; macro F1 $\approx$ 0.91	Chosen for the proposed pipeline
X3D	Efficient expanded 3D CNN	Efficiency-oriented video model	Promising lightweight benchmark
SlowFast	Dual-pathway slow/fast streams	Explicit motion + semantics	Effective but heavier than compact 3D CNNs
TimeSformer	Spatial-temporal self-attention	Models longer-range relationships	Useful comparator, more demanding
Video Swin Transformer	Hierarchical local attention	Strong modern video recognition	High-capacity alternative for future testing

Algorithmic Trade-Offs

The 10.80 percentage-point increase in test accuracy suggests that temporal context has a significant discriminative value for surveillance-style behavior classification. The accompanying macro F1-score of roughly 0.91 indicates that this improvement was not achieved exclusively by favoring one behavioral class. However, extra temporal modelling requires more computing and memory resources than frame-level inference, resulting in a significant accuracy-efficiency trade-off for resource-constrained UAV deployment.

Several restrictions should be addressed when interpreting these findings. Tracking performance is still susceptible to thick crowds, rapid UAV motion, and very small targets, which might result in momentary identification fragmentation or track loss. Furthermore, R3D-18 simulates behavior using narrow temporal windows, which may make it less effective for interactions that take place over much longer time periods. Runtime measurements should also be viewed as prototypical near-real-time observations rather than deterministic hardware-level latency benchmarks. Finally, the disparities between aerial detection datasets and surveillance violence datasets create a domain gap that requires further investigation utilizing aerial-specific behavioral data.

CONCLUSION

This study investigated a multi-stage spatiotemporal architecture for detecting aggression in UAV-based surveillance. The suggested pipeline offers a modular way to integrate temporal reasoning into resource-constrained aerial surveillance by separating spatial person localization with YOLOv8-Nano, temporal identity association with ByteTrack, and behavioral categorization with R3D-18. R3D-18 scored 91.30% test accuracy and a macro F1-score of about 0.91, compared to 80.50% accuracy for the frame-level baseline, resulting in a 10.80 percentage-point increase in test accuracy. This work's contribution is essentially deployment-oriented, rather than introducing a novel individual neural architecture. The findings show the practical utility of combining lightweight detection, multi-object tracking, and spatiotemporal classification in a near-real-time aerial monitoring pipeline. Nevertheless, difficulties such as aerial domain variation, dense-scene tracking, small temporal windows, and processing needs persist. Future research should look into efficient spatiotemporal architectures, model quantization, and further evaluation with aerial-specific behavioral datasets and embedded processing platforms.

Acknowledgements

The authors would like to acknowledge the CSEE Laboratory Technical Team at the University of Essex for their technical insights and infrastructure support during the model training and evaluation phases.

REFERENCES

- [1] N. Masmoudi, W. Jaafar, S. Cherif, J. Ben Abderrazak, and H. Yanikomeroglu, "UAV-based crowd surveillance in post COVID-19 era," *IEEE Access*, vol. 9, pp. 162276–162290, 2021. doi: 10.1109/ACCESS.2021.3131751.
- [2] D. P. Simões et al., "Human detection in UAV imagery using deep learning: A review," *Neural Computing and Applications*, vol. 37, no. 22, pp. 18109–18150, 2025. doi: 10.1007/s00521-025-10874-y.
- [3] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," *arXiv preprint arXiv:1804.07437*, 2018. [Online]. Available: <https://arxiv.org/abs/1804.07437>.
- [4] J. Cheng et al., "RWF-2000: An open large scale video database for violence detection," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2021, pp. 4183–4190. doi: 10.1109/ICPR48806.2021.9412502.
- [5] J. Li et al., "UAV-Human: A large benchmark for human behavior understanding with UAV," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16263–16272. doi: 10.1109/CVPR46437.2021.01599.
- [6] T. Ahmad, M. U. Hadi, V. Vassiliou, L. Dimitriou, A. Anwar, and T. A. Tran, "Real-time anomaly detection in smart vehicle-to-uav networks for disaster management," *Transactions on Emerging Telecommunications Technologies*, vol. 36, no. 5, p. e70162, 2025. doi: 10.1002/ett.70162.
- [7] Soliman et al., "Real Life Violence Situations Dataset," *Kaggle*, 2020. [Online]. Available: <https://www.kaggle.com/datasets/mohamedmustafa/real-life-violence-situations-dataset>.
- [8] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>.
- [9] Y. Zhang et al., "ByteTrack: Multi-object tracking by associating every detection box," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 1–21. doi: 10.1007/978-3-031-20047-2_1.
- [10] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6546–6555. doi: 10.1109/CVPR.2018.00685.
- [11] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497. doi: 10.1109/ICCV.2015.510.
- [12] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 200–210. doi: 10.1109/CVPR42600.2020.00028.
- [13] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6202–6211.
- [14] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, vol. 139, 2021, pp. 813–824.

- [15] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video Swin Transformer," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3202–3211.