From Chaos to Detection: Accident Benchmarking in Surveillance Videos with a Curated Dataset and 3D CNNs

Muhammad Umer Danka
Sunway University, Petaling Jaya, Selangor, Malaysia
umerdanka2005@gmail.com

Ranjit Singh Sarban Singh
Research Centre for Human-Machine Collaboration (HUMAC), Faculty of Engineering and Technology, Sunway University, Bandar Sunway, Petaling Jaya, Selangor, Malaysia
ranjits@sunway.edu.my (corresponding author)

Muhammad Ayoub Danka
Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
ayoubdanka@gmail.com

Received: 18 December 2025 | Revised: 31 January 2026, 4 February 2026, and 7 February 2026 | Accepted: 8 February 2026

 DOI: https://doi.org/10.48084/etasr.17024

ABSTRACT

Automatic accident detection from surveillance videos is an important task for intelligent transportation and public safety, yet it remains underdeveloped compared to violence and anomaly detection. While existing methods often report high accuracy, these results are frequently based on contaminated datasets containing duplicates, overlapping scenes, or annotation artifacts, which inflate performance and limit real-world applicability. To address this gap, this paper introduces a curated benchmark dataset of 513 videos from multiple sources, which are segmented into 4 s clips, resulting in 1,122 clips (561 per class). Using this dataset, we evaluate a spectrum of approaches ranging from handcrafted Violent Flows (ViF) + Support Vector Machine (SVM) pipelines to Convolutional Neural Network–Recurrent Neural Network (CNN–RNN) hybrids and modern 3D convolutional networks. The conducted experiments confirm that traditional methods collapse under stricter evaluation, whereas lightweight architectures such as X3D achieve the best balance between accuracy and efficiency. The reconfigured X3D-S variant achieved 84.48% accuracy, establishing a strong baseline for accident detection under realistic conditions, while offering both a cleaner benchmark and a practical design for future deployment. The Surveillance Curated Accident Dataset (SCAD) and full implementation code are publicly available and can be cited through this paper.

Keywords-accident detection surveillance videos; anomaly detection; X3D; benchmark dataset; deep learning; computer vision; artificial intelligence

I. INTRODUCTION

Automatic accident detection from surveillance videos has emerged as a critical research domain due to its direct implications for public safety, intelligent transportation systems, and smart city development. Prior work on accident reporting systems, including a field survey in Dar es Salaam, suggests that improved reporting can reduce impacts by minimizing response time and supporting the mapping of accident-prone locations [1]. Early warning systems capable of identifying accidents in real time can significantly reduce emergency response times and improve situational awareness for relevant authorities. Despite rapid advancements in computer vision and deep learning, accident detection remains an open challenge, constrained by issues of data reliability, inflated performance claims, and limited computing capacity on edge devices for real-world feasibility.

Most existing studies report very high accuracy when accident detection is treated as an image-based classification problem [2]. In such approaches, Convolutional Neural Network (CNN) feature extractors combined with simple classifier heads frequently report accuracy above 95% on curated datasets [3]. Video-based methodologies, by contrast, generally report lower performance; for example, a recent You Only Look Once version 8 (YOLOv8) plus ByteTrack pipeline reported 94.8% under its evaluation setup [4]. A commonly

used video pipeline consists of vehicle detection using YOLO-based detectors [5], object tracking with correlation or association-based trackers such as MOSSE [6] or ByteTrack [7], crash estimation using track-compensated frame interpolation techniques [8, 9], followed by motion feature extraction with Violent Flows (ViF) descriptor [10] and final classification using a Support Vector Machine (SVM).

A. Accident Detection in Surveillance Systems

Accident detection in surveillance systems has evolved from modular pipelines to more advanced deep feature extraction and end-to-end learning. Most early studies used YOLO [5] for vehicle detection, followed by a tracker such as MOSSE [6] or ByteTrack [7] to track each vehicle. The result was then classified using ViF-style motion features and an SVM, as adopted in later real-time violence detection work [11] to confirm whether accidents have occurred. These pipelines initially achieved around 80% accuracy in 2018 [12], and refinements such as adding a Track Compensated Frame Interpolation (TCFI) increased performance to 92.88% in 2021 [8] and 94.8% in 2025 [4]. However, in our re-evaluation under more realistic conditions that included non-accident footage and full video streams, overall accuracy dropped below 60%, primarily due to many false positives that were not accounted for in earlier evaluations.

Likewise, frame-level approaches also became extremely popular for accident detection. Instead of vehicle detection, tracking, and algorithm-based crash estimation, frames of videos immediately undergo feature extraction with ResNet, Vision Transformers (ViT), and Swin Transformers before being sent to a linear or temporal head for classification. Similarly, they may also be directly classified by newer YOLO versions. These methods reported accuracies that often exceed 95% and have reached 99% as well [13]. Much of this, however, was influenced by dataset design. For instance, the widely used Kaggle CCTV accident collection [14] includes many near-duplicate or overlapping clips; in our manual inspection, similar footage appeared across intended splits, enabling unusually high accuracy even with lightweight models.

Moreover, the CADP dataset [15], which is widely cited, was difficult to access and contained only positive accident clips. Researchers attempted to balance it with negatives from datasets such as UA-DETRAC [16], but this introduced shared backgrounds across classes, creating misleading cues. So-TAD [17] also presented complications, as many clips contained nearly identical backgrounds before and after collisions. In our own experiments, combining CADP and So-TAD produced perfect (100%) video-level accuracy, though this was clearly the result of overlapping scenes rather than genuine learning. TU-DAT [18], which contains non-accident clips that end seconds before its accident clips, highlighted the same weakness, leading models to misclassify based on context rather than accident or event onset.

These cases show that progress in accident detection has often reflected dataset quirks rather than methodological breakthroughs. While deeper architectures and stronger feature extractors appeared to yield rapid gains, their generalization collapsed when tested under stricter or more balanced conditions. The lesson is that without clean and standardized benchmarks, advances risk being overstated, leaving accident detection far from ready for reliable deployment.

B. Violence and Fight Detection as a Precursor

Accident detection shares many of its conceptual roots with research in violence and fight detection. The Hockey Fight dataset, which provided short clips of players engaging in fights on the ice, was one of the earliest and most widely used datasets in this domain. Early approaches such as the ViF descriptor with optical flow and SVM classifiers achieved an accuracy of 81.3% in 2016 [11], and the same method later became the first baseline for accident detection as well, where it is still occasionally used for comparison today. However, the dataset became saturated after a few years as CNN-based feature extractors quickly pushed performance above 98% consistently [19].

Therefore, more complex datasets such as ViF (Crowd Violence) [10] and the RLVS [20] dataset became popular. These benchmarks are more challenging because they moved from controlled fight scenes to crowd-level interactions. However, they too became saturated in the early 2020s [19], with deep convolutional and hybrid-based models coming into action.

Hence, the RWF-2000 [21] and SCFD [22] datasets remained important benchmarks because they better reflect real surveillance conditions. RWF-2000 with an X3D-S [23] fine-tuned on the dataset achieved 94% in 2022 [24], whereas the CUE-NET architecture built on the VideoMAE feature extraction backbone [25] reached 94% as well. SCFD, in comparison, achieved around 92% in 2021 using a 2D CNN with MSM and T-SE modules [26]. Later works in 2025, including a new prototype layer on top of the fine-tuned X3D-S backbone [27] and a system targeting temporal action localization for real-time fistfight detection [28], reported further improvements but have not yet been independently verified. Improvements largely came from implementing larger and computationally heavier networks, but recent progress has also focused in parallel on lighter 3D CNNs aiming to achieve state-of-the-art performance with far fewer FLOPs, making real-time deployment a more realistic possibility.

C. Anomaly Detection in Surveillance Videos

Anomaly detection in videos has also been a central research area for surveillance, with accident detection often considered a subset of this broader task. One of the most influential datasets is UCF-Crime [29], which introduced 14 anomaly categories, including accidents, robbery, and fighting. Early methods of weakly supervised anomaly detection struggled on this dataset, with steady progress being made throughout the years. In 2025, the combination of X3D [23] with the introduction of the Spatio-Temporal Efficient Anomaly Detection (STEAD) [30], a CNN–Transformer hybrid, set a strong baseline on UCF-Crime with 91.34% Area Under the Curve (AUC). The widely used Multi-Timescale Feature Learning (MTFL) [31] method reported 89.78% AUC. Moreover, recent progress has also been made with a Mixture-of-Experts method guided by Gaussian Splatting (GS-MoE)

[32], which assigns specialized models to different anomaly types and combines their outputs through a gating mechanism. This approach achieved 91.58% AUC, edging past earlier state-of-the-art results.

In addition to binary formulations, some studies have explored multi-class anomaly classification, such as a study that modeled six anomaly classes from UCF-Crime and reported overall performance above 90% [33].

Accident detection can also be viewed as a subset of anomaly detection in surveillance videos. UCF-Crime introduced multiple real-world anomaly categories, including road accidents, and remains a widely used benchmark for weakly supervised anomaly detection. However, models pretrained on generic human-action datasets may not capture accident dynamics well, leading to weaker performance on accident-related segments compared to headline metrics reported across the full benchmark.

D. Dataset Limitations and Contamination Issues

Despite massive progress in violence and anomaly detection, accident detection and datasets continue to suffer from flaws that undermine their reliability. Issues such as duplicated frames, overlapping scenes across training and testing, and over-reliance on background context instead of accident dynamics continue to distort results. In some cases, annotation artifacts such as overlays or red circles bias the learning process, whereas in others, like CADP or So-TAD, the lack of balanced non-accident data produces misleading results.

Beyond contamination, access and scale remain obstacles. CADP is difficult to obtain, TU-DAT is too small to yield statistically meaningful results, and many custom-built datasets are never released publicly. Hence, a clear gap exists, as current datasets fail to provide a clean, standardized, and reproducible foundation for accident detection research. This makes performance claims unreliable and impractical for real-world use. Closing this gap is a necessity if the field is to move beyond inflated numbers and toward solutions that are both practical and generalizable.

The main contributions of this work are as follows:

1. A critical analysis of existing accident detection datasets and benchmarks, highlighting problems of contamination, saturation, and lack of reproducibility.
2. The introduction of a curated dataset containing 513 clean, segmented, and annotated accident and non-accident videos to support fair and consistent evaluation.
3. A comprehensive study comparing handcrafted features, CNNs, transformer-based models, and 3D convolutional approaches, showcasing the limits of older techniques and identifying the models that currently achieve the strongest results.
4. Evidence that a lightweight X3D-S backbone offers the best trade-off between accuracy and efficiency, making real-time accident detection practical even on low-power hardware.

II. METHODOLOGY

A. Dataset Construction

The curated dataset for accident detection, the Surveillance Curated Accident Dataset (SCAD), was constructed by aggregating videos from four public sources: 1,416 from CADP [15], 282 from So-TAD [17], 60 from TU-DAT [18], and 150 from UCF-Crime [29]. This produced an initial pool of 1,908 videos. To ensure a clean and non-contaminated benchmark, we applied a multi-stage curation process to remove exact duplicates, near-duplicates, repeated scenes, and clips containing annotation artifacts such as text overlays, red circles, and animated highlights. After this filtering, 513 unique videos remained.

Each retained video was then segmented into shorter 4 s clips, inspired by the fixed-length clip design of RWF-2000 [21], while also addressing class imbalance to support effective supervised learning. Because splitting is performed by scene/group ID, all overlapping 4 s clips extracted with a 1 s stride from the same source video/scene are kept within the same split, preventing stride-induced leakage. Clips were extracted with a 1 s stride to capture accident onset. The final benchmark contains 1,122 clips, balanced at 561 accident and 561 non-accident clips. As an additional leakage check, we computed perceptual hashes on sampled frames and verified that no near-duplicate matches exceeded the similarity threshold across train and test splits. An illustration of the dataset construction and labeling pipeline is shown in Figure 1.

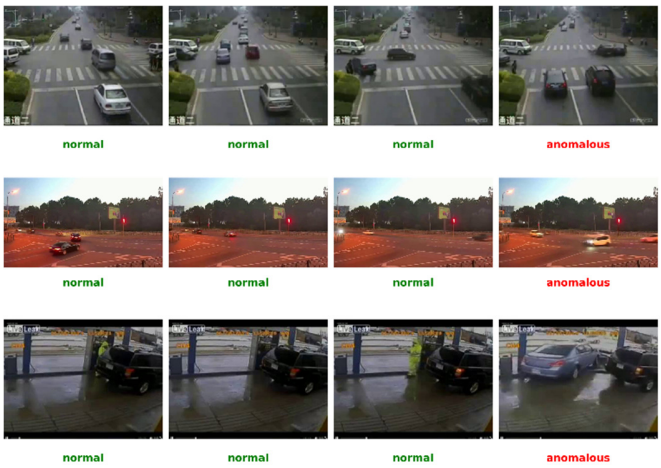

Fig. 1. Examples of video segmentation and labeling process.

All supervised training and evaluation are performed using clip-level labels (normal vs anomalous). Train/test partitioning is performed at the source video/scene level so that all clips derived from the same scene remain within a single split. Evaluation used a scene-disjoint split defined by group ID, yielding approximately 890 training clips and 232 testing clips, with zero cross-split scene overlap. In this work, a scene is operationalized as a unique camera viewpoint and location context within a source video (same camera angle/background layout), and all clips extracted from that scene are assigned the same group ID. This corresponds to approximately 890 training clips (445 per class) and 232 testing clips (116 per class).

SCAD and the full preprocessing, training, and evaluation code are released at [34], including per-clip metadata with scene/group IDs and the official train/test split file by group ID to enable reproducible zero-overlap evaluation; the dataset should be cited through this paper.

B. Proposed Methodology

We employed X3D [23], a lightweight 3D convolutional network introduced by Facebook AI for efficient spatiotemporal modeling. X3D begins as a compact base network and gradually expands its depth, width, temporal span, and spatial resolution during the feature extraction process. This design makes the model ideal for capturing spatiotemporal patterns while keeping the computational cost much lower than older 3D CNNs like I3D [35] or SlowFast [36].

X3D was subsequently fine-tuned in 2022 for a violence detection task [24], which effectively created the state-of-the-art method for RWF-2000 at 94%, before being surpassed by a similar model with a prototype head expansion in 2025 [27]. Their implementations generally used higher spatial resolutions to optimize performance on large-scale datasets.

In this study, the X3D-S variant adopts the architecture shown in Table I. The original model, which is initialized with weights pretrained on Kinetics-400, was then fine-tuned on the new accident dataset. Consequently, the input resolution in the experiment was reduced to 182 × 182, while 15 frames were extracted instead of the usual 13. Therefore, a reliable yet high-performance system is achieved with low computational load, making the approach suitable for practical accident detection.

TABLE I. PROPOSED X3D-S VARIANT ARCHITECTURE

Stage	**Layer**	**×**	**Output *size* (T × H × W)**
Input	Data layer	—	15 × 182 × 182
Conv1	1×3^2, 3 × 1, 24	1	15 × 91 × 91
Res2	1×1^2, 54 3×3^2, 54 1×1^2, 24	3	15 × 46 × 46
Res3	1×1^2, 108 3×3^2, 108 1×1^2, 48	5	15 × 23 × 23
Res4	1×1^2, 216 3×3^2, 216 1×1^2, 96	11	15 × 12 × 12
Res5	1×1^2, 432 3×3^2, 432 1×1^2, 192	7	15 × 6 × 6
Conv5	1×1^2, 432	1	15 × 6 × 6
Pool5	15 × 6 × 6	—	1 × 1 × 1
FC1	1×1^2, 2,048	1	1 × 1 × 1
FC2	1×1^2, #classes	1	1 × 1 × 1

III. RESULTS

Table II presents the comparative performance of representative accident detection approaches evaluated on SCAD, a balanced benchmark of 1,122 manually labeled 4 seconds surveillance clips (561 accident and 561 non-accident) curated from CADP, So-TAD, TU-DAT, and UCF-Crime, as described in Section II.A. Additional experiments on the original benchmarks exhibited inflated performance consistent with the contamination effects discussed in Section I.D; therefore, SCAD is used as the main evaluation benchmark.

TABLE II. MODEL PERFORMANCE

Method	**Accuracy**	**GFLOPs**	**Notes**
ViT + classifier head	50.0	~140	Ineffective
YOLO & filter & ViF + SVM [8]	60.2	~132	Outdated
ResNet-50 + GRU	78.3	~30	Unstable
ResNet-50 + 2 Transformer heads	79.8	~35	Still unstable
X3D-L + STEAD (30 frames) [30]	80.6	~36	STEAD baseline
X3D-L + STEAD (60 frames) [30]	81.7	~72	High GFLOPs
X3D-S (30 frames, 128 × 128)	81.9	3.63	X3D-S baseline
X3D-S (15 frames, 182 × 182)	84.48	3.36	Strong baseline

The evaluated methods span classical detection-and-motion pipelines, 2D CNN feature extractors combined with temporal modeling, and end-to-end 3D convolutional networks for spatiotemporal learning. All results are obtained using implementations reproduced from the authors publicly released repositories where available or re-implemented to closely follow the original architectures and training protocols.

In addition to overall accuracy, we report the confusion matrix and class-wise precision, recall, and F1-score to reflect false-alarm and miss behavior under balanced testing. On the SCAD test set (232 clips: 116 normal, 116 accident), the best-performing X3D-S model achieves 84.48% accuracy with confusion matrix (TN = 97, FP = 19, FN = 17, TP = 99), as shown in Figure 2. The corresponding scores are normal: precision = 0.85, recall = 0.84, F1-score = 0.84, and accident: precision = 0.84, recall = 0.85, F1-score = 0.85.

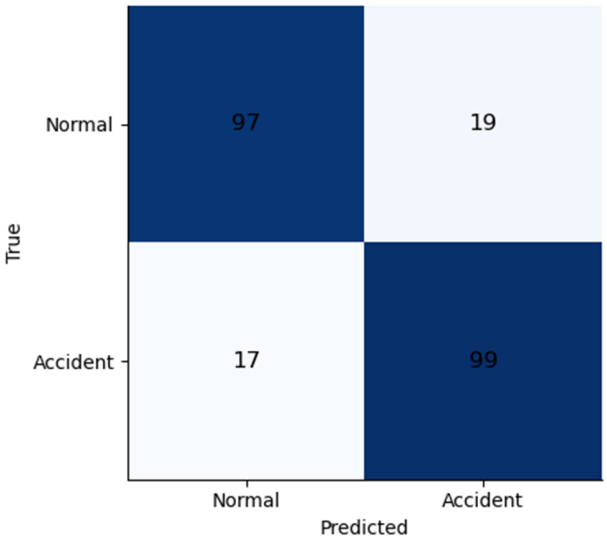

Fig. 2. Confusion matrix for X3D-S model on the SCAD test set.

These results show that classical detection-and-motion pipelines degrade sharply under the scene-disjoint and

contamination-controlled SCAD evaluation, whereas CNN-based backbones with temporal heads peak around 78–80% accuracy. Heavier 3D CNNs such as X3D-L improve performance but incur substantially higher computational costs. By contrast, the proposed X3D-S configuration achieves the strongest balance between accuracy and efficiency, establishing a practical baseline for surveillance accident detection.

IV. DISCUSSION

Accident detection in surveillance videos is still behind related fields such as violence or anomaly detection. The lack of a large, standardized, and clean dataset remains the main issue, due to existing collections containing duplicates or overlapping scenes that inflate reported results. When past methodologies are tested under cleaner conditions, they collapse, whereas CNN-based methods struggle to generalize beyond 80% accuracy.

In response to benchmarking concerns, we also trained and evaluated X3D-S using the original datasets and commonly used protocols reported in prior work. However, performance on these benchmarks was strongly influenced by the same issues identified in Section I.D, including scene overlap, near-duplicate clips, and background leakage, which led to inflated results that did not reflect true accident understanding. For this reason, we report SCAD as the primary benchmark, as it enforces scene-disjoint splits and contamination-controlled evaluation, enabling more reliable comparisons for future studies.

On the other hand, newer architectures that incorporate motion features alongside visual features such as X3D have consistently topped benchmarks including RWF-2000, SCFD, and UCF-Crime. Our findings confirm that X3D also continues to remain the strongest available option under real-time resource constraints.

V. FUTURE WORK

The most immediate step forward is the development of a larger and cleaner dataset, ideally containing at least 1,000 accident and 1,000 non-accident clips per class, to improve generalization and reliability. Beyond dataset scale, specialized deployment-oriented testing is required: for example, placing cameras in fixed locations and collecting accident footage over time until sufficient real-world data are obtained, which would provide a deployable model.

On the modeling side, exploring refinements to X3D features can potentially improve performance further while maintaining low GFLOPs. X3D already leads results on RWF-2000, SCFD, UCF-Crime, and now our curated dataset. Further refinements tailored to accident-specific dynamics and real-time inference could push accuracy slightly higher without sacrificing efficiency. These steps would help move accident detection from experimental studies toward practical deployment.

VI. CONCLUSION

This study addressed the persistent gap between reported benchmark performance and real-world feasibility in accident detection. We demonstrated that existing datasets often suffer from contamination and duplication, leading to inflated results that do not generalize. By curating a dataset of 513 unique accident and non-accident clips, segmented and annotated at the scene level, we provide a cleaner and more reliable testbed for the community.

Through a comparative evaluation of classical pipelines, Convolutional Neural Network (CNN)–based backbones, and modern 3D convolutional approaches, we showed that traditional methods collapse under realistic conditions, whereas lightweight 3D CNNs achieve the strongest balance between accuracy and efficiency. The reconfigured X3D-S variant, operating with reduced spatial resolution and shorter clip lengths, achieved competitive performance at a fraction of the computational cost.

Together, these contributions deliver both a reproducible dataset and a practical modeling baseline. They help shift the field from inflated benchmark claims toward methods that are reproducible, efficient, and closer to real-world deployment. Future progress will depend on building larger datasets, exploring multimodal signals, and refining architectures like X3D for accident-specific dynamics.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no competing interests related to this work.

ACKNOWLEDGMENT

This publication was funded by Faculty of Engineering and Technology (FET), Sunway University. The authors would like to also thank the Research Centre for Human–Machine Collaboration (HUMAC) for their guidance and research support. This work was conducted as part of an academic research initiative with no external commercial funding. The authors also acknowledge the developers and maintainers of the publicly available datasets used in this study, whose contributions enabled rigorous benchmarking and evaluation.

DATA AVAILABILITY

The dataset used in this study was developed by the authors through the integration of the publicly available datasets, which can be found in [15, 17, 18, 29]. All original data sources are properly cited in the manuscript and remain publicly accessible through their respective repositories.

DECLARATION OF GENERATIVE AI USE

During the preparation of this work, the authors used ChatGPT for language refinement, readability improvement, and brainstorming support after the manuscript had been drafted by the authors. The authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

[1] I. J. Mrema and M. A. Dida, "A Survey of Road Accident Reporting and Driver's Behavior Awareness Systems: The Case of Tanzania," *Engineering, Technology & Applied Science Research*, vol. 10, no. 4, pp. 6009–6015, Aug. 2020, https://doi.org/10.48084/etasr.3449.

[2] M. S. Arefin, M. I. S. Mahin, and F. A. Mily, "Real-time rapid accident detection for optimizing road safety in Bangladesh," *Heliyon*, vol. 11,

no. 4, Feb. 2025, Art. no. e42432, https://doi.org/10.1016/j.heliyon.2025.e42432.

[3] B. Pérez, M. Resio, T. Seco, F. García, and A. Al-Kaff, "Innovative Approaches to Traffic Anomaly Detection and Classification Using AI," *Applied Sciences*, vol. 15, no. 10, May 2025, Art. no. 5520, https://doi.org/10.3390/app15105520.

[4] P. Kalpana, G. Sowmiya, C. R. S. Sri, and S. Sivapriya, "Road traffic accident detection based on Yolov8 and Byte Track," *AIP Conference Proceedings*, vol. 3204, no. 1, Feb. 2025, Art. no. 040013, https://doi.org/10.1063/5.0248651.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 779–788, https://doi.org/10.1109/CVPR.2016.91.

[6] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010, pp. 2544–2550, https://doi.org/10.1109/CVPR.2010.5539960.

[7] Y. Zhang *et al.*, "ByteTrack: Multi-object Tracking by Associating Every Detection Box," in *17th European Conference on Computer Vision*, Tel Aviv, Israel, 2022, pp. 1–21, https://doi.org/10.1007/978-3-031-20047-2_1.

[8] K. Sabry and M. Emad, "Road Traffic Accidents Detection Based On Crash Estimation," in *2021 17th International Computer Engineering Conference*, Cairo, Egypt, 2021, pp. 63–68, https://doi.org/10.1109/ICENCO49852.2021.9698968.

[9] S. Dikbas and Y. Altunbasak, "Novel True-Motion Estimation Algorithm and Its Application to Motion-Compensated Temporal Frame Interpolation," *IEEE Transactions on Image Processing*, vol. 22, no. 8, pp. 2931–2945, Aug. 2013, https://doi.org/10.1109/TIP.2012.2222893.

[10] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Providence, RI, USA, 2012, pp. 1–6, https://doi.org/10.1109/CVPRW.2012.6239348.

[11] V. Machaca Arceda, K. Fernández Fabián, and J. C. Gutíerrez, "Real time violence detection in video," in *International Conference on Pattern Recognition Systems (ICPRS-16)*, Talca, Chile, 2016, pp. 1–7, https://doi.org/10.1049/ic.2016.0030.

[12] V. Machaca Arceda and E. Laura Riveros, "Fast car Crash Detection in Video," in *2018 XLIV Latin American Computer Conference*, Sao Paulo, Brazil, 2018, pp. 632–637, https://doi.org/10.1109/CLEI.2018.00081.

[13] F. Bukhari, B. Gul, J. H. Shah, and A. Ali, "Attention-Guided Transformer-CNN Hybrid for Real-Time Road Accident Classification with Interpretability." Social Science Research Network, Rochester, NY, Sept. 05, 2025, https://doi.org/10.2139/ssrn.5445923.

[14] C. Charan Kumar, "Accident Detection From CCTV Footage." Kaggle, https://doi.org/10.34740/kaggle/dsv/1379553.

[15] A. P. Shah, J.-B. Lamare, T. Nguyen-Anh, and A. Hauptmann, "CADP: A Novel Dataset for CCTV Traffic Camera based Accident Analysis," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance*, Auckland, New Zealand, 2018, pp. 1–9, https://doi.org/10.1109/AVSS.2018.8639160.

[16] L. Wen *et al.*, "UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking," *Computer Vision and Image Understanding*, vol. 193, Apr. 2020, Art. no. 102907, https://doi.org/10.1016/j.cviu.2020.102907.

[17] X. Chen, H. Xu, M. Ruan, M. Bian, Q. Chen, and Y. Huang, "SO-TAD: A surveillance-oriented benchmark for traffic accident detection," *Neurocomputing*, vol. 618, Feb. 2025, Art. no. 129061, https://doi.org/10.1016/j.neucom.2024.129061.

[18] P. P. Kumar and K. Kant, "TU-DAT: A Computer Vision Dataset on Road Traffic Anomalies," *Sensors*, vol. 25, no. 11, May 2025, Art. no. 3259, https://doi.org/10.3390/s25113259.

[19] M. S. M. Shubber and Z. T. M. Al-Ta'i, "A review on video violence detection approaches," *International Journal of Nonlinear Analysis and Applications*, vol. 13, no. 2, pp. 1117–1130, July 2022, https://doi.org/10.22075/ijnaa.2022.6369.

[20] M. M. Soliman, M. H. Kamal, M. A. El-Massih Nashed, Y. M. Mostafa, B. S. Chawky, and D. Khattab, "Violence Recognition from Videos using Deep Learning Techniques," in *2019 Ninth International Conference on Intelligent Computing and Information Systems*, Cairo, Egypt, 2019, pp. 80–85, https://doi.org/10.1109/ICICIS46948.2019.9014714.

[21] M. Cheng, K. Cai, and M. Li, "RWF-2000: An Open Large Scale Video Database for Violence Detection," in *2020 25th International Conference on Pattern Recognition*, Milan, Italy, 2021, pp. 4183–4190, https://doi.org/10.1109/ICPR48806.2021.9412502.

[22] Ş. Aktı, G. A. Tataroğlu, and H. K. Ekenel, "Vision-based Fight Detection from Surveillance Cameras," in *2019 Ninth International Conference on Image Processing Theory, Tools and Applications*, Istanbul, Turkey, 2019, pp. 1–6, https://doi.org/10.1109/IPTA.2019.8936070.

[23] C. Feichtenhofer, "X3D: Expanding Architectures for Efficient Video Recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 200–210, https://doi.org/10.1109/CVPR42600.2020.00028.

[24] J. Su, P. Her, E. Clemens, E. Yaz, S. Schneider, and H. Medeiros, "Violence Detection using 3D Convolutional Neural Networks," in *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance*, Madrid, Spain, 2022, pp. 1–8, https://doi.org/10.1109/AVSS56176.2022.9959393.

[25] D. C. Senadeera, X. Yang, D. Kollias, and G. Slabaugh, "CUE-Net: Violence Detection Video Analytics with Spatial Cropping, Enhanced UniformerV2 and Modified Efficient Additive Attention," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Seattle, WA, USA, 2024, pp. 4888–4897, https://doi.org/10.1109/CVPRW63382.2024.00493.

[26] M.-S. Kang, R.-H. Park, and H.-M. Park, "Efficient Spatio-Temporal Modeling Methods for Real-Time Violence Recognition," *IEEE Access*, vol. 9, pp. 76270–76285, 2021, https://doi.org/10.1109/ACCESS.2021.3083273.

[27] P. Her, E. Yaz, and S. Schneider, "Interpretable Convolutional Neural Network for Violence Recognition," in *11th International Conference on Computational Science and Computational Intelligence*, Las Vegas, NV, USA, 2024, pp. 40–53, https://doi.org/10.1007/978-3-031-94962-3_4.

[28] B. Qi, B. Wu, and B. Sun, "Automated violence monitoring system for real-time fistfight detection using deep learning-based temporal action localization," *Scientific Reports*, vol. 15, no. 1, Aug. 2025, Art. no. 29497, https://doi.org/10.1038/s41598-025-12531-4.

[29] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 6479–6488, https://doi.org/10.1109/CVPR.2018.00678.

[30] A. Gao and J. Liu, "STEAD: Spatio-Temporal Efficient Anomaly Detection for Time and Compute Sensitive Applications," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Hangzhou, China, 2025, pp. 6256–6263, https://doi.org/10.1109/IROS60139.2025.11246678.

[31] W. Sun, L. Cao, Y. Guo, and K. Du, "Multimodal and multiscale feature fusion for weakly supervised video anomaly detection," *Scientific Reports*, vol. 14, no. 1, Oct. 2024, Art. no. 22835, https://doi.org/10.1038/s41598-024-73462-0.

[32] G. D'Amicantonio *et al.*, "Mixture of Experts Guided by Gaussian Splatters Matters: A new Approach to Weakly-Supervised Video Anomaly Detection." arXiv, Aug. 08, 2025, https://doi.org/10.48550/arXiv.2508.06318.

[33] A. Phapale and S. Bhingarkar, "Deep Context-Aware Feature Extraction for Anomaly Detection in Surveillance Videos," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 21633–21638, Apr. 2025, https://doi.org/10.48084/etasr.9810.

[34] WhoCares258, "WhoCares258/SCAD: Surveillance Curated Accident Dataset." Zenodo, Feb. 04, 2026, https://doi.org/10.5281/zenodo.18483633.

[35] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 4724–4733, https://doi.org/10.1109/CVPR.2017.502.

[36] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast Networks for Video Recognition," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, South Korea, 2019, pp. 6201–6210, https://doi.org/10.1109/ICCV.2019.00630.