ISSN 2355-0082

Performance Analysis of X3D-XS Architecture for Cross-Domain Real-World Violence Detection

Beril Berekhya Mutia Hasibuan[1], Irma Amelia Dewi[2*], Muhammad Ichwan[3]
[1,2,3]Department of Informatics, Institut Teknologi Nasional Bandung, Indonesia
[1]beril.berekhya@mhs.itenas.ac.id, [2*]irma_amelia@itenas.ac.id, [3]michwan86@gmail.com

Accepted 26 November 2026
Approved 26 May 2026

***Abstract*— Real-time violence detection systems need models that are efficient and can adapt to different environments. This study looks at the performance of the X3D-XS architecture and focuses on the issue of generalizing across various domains. We trained the model on three controlled source domains: AVD, HockeyFight, and MovieFight. We then tested its performance on the mixed Real-Life Violence Situations (RLVS) dataset. The experimental results show that X3D-XS is highly efficient, achieving inference speeds of up to 191 FPS, which makes it suitable for edge deployment. However, the model faces significant challenges due to domain shift; training on a single domain resulted in varying accuracy between 49.5% and 61.1% on real-world data. This indicates that staged and cinematic violence differ quite a bit from real-life situations. Importantly, combining different source domains improved the model's sensitivity, leading to a Recall of 94.28%. These findings demonstrate that while X3D offers the speed needed for real-time monitoring, relying solely on staged training data is not enough for real-world effectiveness, highlighting the essential need for data diversity in surveillance applications.**

***Index Terms*— Cross-Domain Generalization; Real-Time Surveillance; Video Violence Detection; X3D Architecture; Deep Learning; Computer Vision.**

I. INTRODUCTION

The widespread use of video surveillance has changed how public safety systems operate. Millions of Closed-Circuit Television (CCTV) cameras are set up worldwide in smart cities, schools, and industrial sites. This has led to a huge amount of video data that is too much for people to handle. Research shows that human attention drops significantly after just 20 minutes of watching video feeds. This results in security incidents being overlooked [1]. As a result, there is a pressing need for smart video analytics that can spot violent activities, like assaults, riots, and physical fights, in real-time.

In recent years, the field of action recognition has moved from manual feature extraction methods like Optical Flow and HOG to Deep Learning techniques. Early Deep Learning models used 2D Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) units to capture temporal dynamics [2]. However, the arrival of 3D CNNs, such as C3D [3] and I3D [4], changed the game. These models learn spatiotemporal features at the same time, setting new performance standards. Despite their precision, traditional 3D CNNs are costly in terms of computation. They require substantial GPU resources, which are often not available in edge computing environments where surveillance systems run. This computational challenge has historically limited the use of high-accuracy 3D models in real-world security systems, forcing practitioners to settle for lower detection accuracy.

To solve the efficiency problem, Feichtenhofer suggested the X3D architecture [5]. This approach takes 2D ConvNets and extends them into 3D by improving on several factors: frame rate, duration, resolution, and network width. X3D delivers high accuracy while using much less computing power than earlier models. Although being efficient is necessary for deployment, the main issue with today's violence detection systems is not how fast they operate, but how well they generalize.

Most existing studies evaluate models using an "Intra-domain" protocol. In this setup, the training and testing data come from the same source. For example, training might happen on movie clips and testing on separate movie clips [6]. This method hides the "Domain Gap." In real-life situations, a system trained on clean, bright, or staged datasets has to work in uncontrolled environments. These environments often suffer from poor lighting, occlusion, and camera noise. This study tackles this gap by doing a "Cross-Domain" evaluation. We look at how the lightweight X3D-XS model performs when moved from controlled areas like Movies, Sports, and Surveillance to real-life surveillance footage. This transfer learning analysis sets the performance baseline needed for future research on domain adaptation techniques[18].

This study contributes to the field of intelligent surveillance by primarily validating the computational efficiency of the X3D-XS architecture, confirming its

suitability for low-latency, real-time usage on resource-constrained devices. Furthermore, the analysis quantifies the "generalization gap" by measuring the performance degradation that occurs when models trained on synthetic or staged violence are deployed in real-world CCTV environments. Finally, the research assesses the impact of data aggregation, demonstrating that combining multiple data sources significantly enhances model sensitivity (Recall), thereby ensuring that critical threats are detected despite environmental variations.

II. Related Work

The progress in computer vision has greatly sped up the creation of automated surveillance systems. Early methods mainly used hand-crafted features to model human movement, but these often had difficulty capturing the complexities of crowd behavior. The shift to Deep Learning has helped overcome many of these issues by allowing for complete feature learning. However, despite these improvements, two major challenges still appear in recent studies: the high computational cost of 3D architectures and the inability to generalize across different surveillance settings. This section reviews current research on deep learning-based violence detection, efficient video architectures, and the issues of domain adaptation.

A. Deep Learning for Violence Detection

The evolution of violence detection reflects broader advancements in video understanding. Early efforts, like the Violent Flows (ViF) descriptor proposed by Hassner et al. [7], estimated crowd violence by looking at statistical changes in optical flow magnitude. With the growing popularity of deep learning two-Stream networks is introduced, which separated spatial and temporal processing[8]. Later, Tran et al. [3] showed that 3D convolutions (C3D) could effectively capture motion patterns directly from video volumes. More recently, the SlowFast network [9] used dual pathways that operate at different frame rates to capture both static semantics and fast motion. However, these models often focus on accuracy, which can come at the expense of inference speed and parameter efficiency, limiting their usefulness in large-scale surveillance systems.

Recent studies have therefore begun to emphasize efficiency-aware violence detection models that balance recognition performance with computational cost. Lightweight 3D CNN architectures and optimized spatiotemporal representations have been shown to enable real-time inference without significant accuracy degradation, making them more suitable for continuous surveillance scenarios [20], [21]. Moreover, survey-based analyses highlight that deployment feasibility, including latency and scalability, has become a critical consideration in modern violence detection research [22], [24]. In addition, comprehensive reviews on abnormal behavior and violence detection in intelligent surveillance systems further underline the challenges of balancing accuracy and efficiency in real-world deployments [23].

B. Efficient Video Architectures

Efficiency is a key requirement for scalable video surveillance systems, especially under limited computational and energy resources. The X3D family of networks [5] was designed to balance recognition accuracy and computational efficiency through a coordinate descent strategy that expands the network along the most effective architectural dimensions, rather than uniformly increasing depth or width. This approach allows X3D to maintain strong performance while minimizing redundant computation.

Architecturally, X3D employs channel-wise separable convolutions and a multigrid expansion strategy, which significantly reduce Floating Point Operations (FLOPs) compared to conventional 3D CNN backbones such as I3D or ResNet-3D. While transformer-based models like ViViT and TimeSformer often require hundreds of GigaFLOPs, X3D-XS operates with fewer than 5 GFLOPs, making it well-suited for real-time and edge-based surveillance applications.

The growing focus on lightweight video recognition models highlights the importance of deployment feasibility alongside accuracy. Recent studies have shown that efficient 3D CNN architectures can achieve competitive performance while supporting real-time processing in surveillance environments [20], [21]. Moreover, such efficiency-oriented designs are increasingly favored in security applications that demand scalability and low latency [22]. Consequently, X3D represents an appropriate backbone for this study, supporting practical and deployable violence detection systems.

C. The Challenge of Domain Generalization

Domain adaptation is a continuing challenge in computer vision[19]. A model trained on a "Source Domain" usually performs poorly on a "Target Domain" because of changes in data distribution [10]. In violence detection, this problem is made worse by the use of datasets like MovieFight or HockeyFight. These datasets are both semantically and visually different from the Real-Life Violence Situations (RLVS) seen in public surveillance [11]. Current research focuses on learning domain-invariant spatiotemporal features to bridge this gap [17]. While recent studies look into Unsupervised Domain Adaptation (UDA), it is important to first understand the baseline performance of effective models across these domains. This understanding is essential for developing strong solutions. Recent survey studies further confirm that domain shift remains a major unresolved issue in video-based action and violence

recognition, particularly under real-world surveillance conditions [25].

III. Research Methodology

This section describes the method used to assess the performance and reliability of the proposed violence detection framework. We outline the experimental design, which includes gathering various datasets and setting up the X3D model. Additionally, the thorough training procedures and evaluation metrics presented here are designed to guarantee that the results can be replicated and that the cross-domain analysis is valid.

A. Research Framework

This study uses a quantitative experimental approach to assess how well efficient video architectures can generalize across different areas. The method is divided into three main phases:

a. Data Curation, which includes carefully categorizing the source (training) and target (testing) domains.

b. Model Training, which uses the X3D-XS architecture in single-source and multi-source aggregation scenarios.

c. Performance Evaluation, which performs zero-shot testing on new real-world data to measure the generalization gap.

B. Dataset Configuration

To clearly visualize the environmental variations and the domain shift challenge, Fig. 1 presents sample frames from the four datasets used in this study, contrasting the staged source domains with the real-world target domain.

Fig. 1. Sample frames from the datasets used in this study: (a) AVD (Real-world/CCTV), (b) HockeyFight (Sports), (c) MovieFight (Cinematic), and (d) RLVS (Target Real-Life). Note the visual diversity between the source domains and the target domain

a. Source Domains (Training Sets) These datasets represent the environments used to train the models:

i. Automatic Violence Detection Dataset (AVD) [12]: Curated by Bianculli et al., this dataset serves as a baseline for real-world surveillance. It consists of 350 video clips from CCTV and user-generated content. Unlike staged datasets, AVD contains real violent events. However, it is relatively small and has a specific selection bias compared to larger "in-the-wild" collections.

ii. HockeyFight Dataset [13]: Representing the "Sports" domain, this dataset includes 1,000 clips of fights during ice hockey games. The motion is very focused and aggressive, but the environment is limited with uniform backgrounds (white ice), specific attire, and consistent camera angles.

iii. MovieFight Dataset [14]: This dataset represents the "Cinematic" domain and includes clips from action movies. This domain has a lot of visual noise due to staged effects and exaggerated acting. It features dramatic choreography, professional lighting, and quick editing cuts, which are quite different from continuous surveillance footage.

b. Target Domain (Testing Set):

i. Real-Life Violence Situations (RLVS) [11]: Used only for Zero-Shot Evaluation. This dataset includes 2,000 clips collected from YouTube (CCTV, dashcams, mobile). It acts as the "Uncontrolled" target domain because of its wide range of lighting, resolution, camera stability, and background clutter.

C. Proposed Architecture: X3D-XS

The recognition model is X3D-XS (Extra Small), an efficient 3D ConvNet [5]. Unlike traditional 3D CNNs, such as C3D and I3D, which expand 2D kernels in the same way, X3D uses a coordinate descent optimization to expand the network along six specific axes, as illustrated in Fig 2, enabling efficient exploration of the model design space.

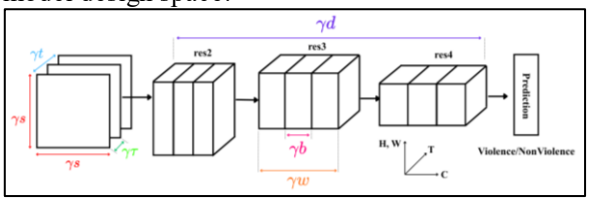

Fig. 2. Schematic of the X3D architecture expansion. The model expands a 2D ConvNet along six coarse-to-fine axes (γ). It generates efficient spatiotemporal features

a. Frame Rate ($\gamma\tau$): The temporal sampling frequency.

b. Duration (γt): The length of the input video clip.

c. Resolution (γs): The spatial dimensions of the frame.

d. Width ($\gamma\omega$): The channel capacity of the network.

e. Depth (γd): The number of network layers.

f. Bottleneck (γb): The inner channel width of the residual blocks.

The X3D-XS variant comes from shrinking these axes to reduce Floating Point Operations (FLOPs) while maintaining its representational power. It uses channel-wise separable convolutions, which makes it very suitable for edge deployment. Fig 3 shows the entire processing pipeline, from the input tensor to the binary classification output.

For this study, the final classification head is adapted for binary output: Violence and Non-Violence.

D. Experimental Setup

The models were implemented using the PyTorch framework with CUDA acceleration. To ensure a thorough evaluation and reproducibility, the training environment was set up as follows:

a. Optimization: The network was optimized with the Adam optimizer and a cosine annealing learning rate scheduler to maintain stable convergence during training.

b. Loss Function: Cross-Entropy Loss was used as the objective function for the binary classification task.

c. Training Protocol: To explore the effect of data diversity, models were trained in two different scenarios: individually on each source dataset (Single-Source) and then on a combined dataset (Combination). To ensure rigorous evaluation, the datasets were partitioned using a stratified split ratio of 75:15:10 for training, validation, and testing, respectively. This split ensures that the model is evaluated on unseen data while maintaining class distribution consistency.

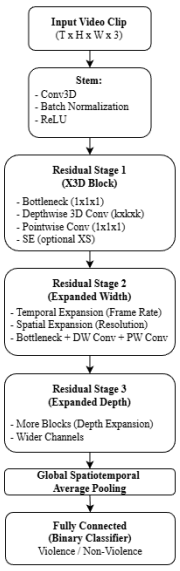

Fig. 3. Block diagram of the X3D-XS network structure used in this study. The architecture processes the input clip through a series of separable 3D convolutions and a global average pooling layer. This helps to predict the violence probability

Implementation Environment: All experiments took place on a high-performance workstation with an Intel Core i9-12900K processor (16 cores, 24 threads) and 32 GB of RAM, using an NVIDIA GeForce RTX 4090 (24 GB) for computation. To ensure reproducibility, the random seed was set to 42. This high-end hardware setup guarantees that the recorded inference latency reflects the model's best performance in a strong computational environment.

E. Evaluation Metrics

To ensure a comprehensive assessment, we evaluated the models using standard classification and efficiency metrics:

a. Classification Performance based on the confusion matrix (TP, TN, FP, FN), we calculate:

i. Accuracy: The overall correctness of predictions.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

ii. Precision: The reliability of positive predictions, which is critical for reducing false alarms.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

iii. Recall (Sensitivity): The ability to detect all violent instances and ensure no threat is missed.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

iv. F1-Score: The harmonic mean of Precision and Recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

b. Computational Efficiency to assess real-time viability, we measured:

i. Inference Time: Average processing time per video clip (ms).

ii. Throughput (FPS): Frames processed per second.

$$FPS = \frac{Total\ Frames}{Total\ Inference\ Time\ (s)} \quad (5)$$

IV. Results and Discussion

In this section, we present a detailed look at the experimental findings, divided into two main areas: computational efficiency and classification robustness. First, we demonstrate that the X3D-XS model can be used for edge deployment by measuring its inference speed. Next, we closely examine the model's ability to generalize. We look at the performance drop that occurs when transferring knowledge from controlled source domains to the unrestricted target domain, focusing on the reasons for the gap between the domains.

A. Computational Efficiency Analysis

To evaluate the practical use of the proposed framework, we measured the inference speed on a

high-performance workstation with an NVIDIA RTX 4090. Table I shows the performance metrics.

TABLE I. INFERENCE PERFORMANCE AND SCALABILITY OF X3D-XS

Model Source	*Inference Latency (ms)*	*Throughput (FPS)*	*Real-Time Factor (x)*
Single_AVD	5.23	191.34	6.3x
Single_Hockey	7.57	132.17	4.4x
Single_Movie	7.73	129.39	4.3x
TrainC_Comb	7.29	137.15	4.5x

The X3D-XS model achieved an average throughput of 191.34 FPS on the AVD dataset, exceeding the standard real-time requirement (30 FPS) by a factor of six. This performance confirms that the architecture is highly scalable; a single processing unit can effectively monitor multiple video streams simultaneously. Consequently, X3D-XS proves to be a viable solution for both centralized server processing and resource-constrained edge deployment, significantly lowering hardware barriers for large-scale surveillance.

B. Cross-Domain Generalization Analysis

The main finding of this study is the quantitative "Generalization Gap." Table II shows the specific performance metrics of the source-trained models. Fig. 4 offers a visual comparison that highlights the significant drop in performance when these models are used on the target RLVS dataset.

TABLE II. ZERO-SHOT CROSS-DOMAIN EVALUATION ON RLVS (TARGET DOMAIN)

Model Source	*Accuracy (%)*	*Precision*	*Recall*	*F1-Score (Macro)*	*AUC*
AVD	60.19	0.574	0.848	0.573	0.571
Hockey	61.17	0.671	0.467	0.604	0.669
Movie	49.51	0.503	0.924	0.369	0.448
TrainC Comb	62.62	0.582	0.943	0.579	0.666

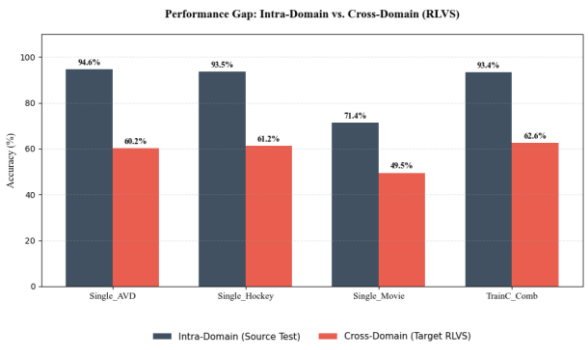

Fig. 4. Performance Gap Analysis. Comparison of expected Intra-domain performance vs. actual Cross-domain performance on RLVS

The quantitative generalization gap presented in Table II highlighting the severe performance degradation when models trained on single source domains are applied to the target RLVS dataset. The Single_Movie model performed poorly (49.51% accuracy), indicating that cinematic features such as dramatic angles and rapid editing, fail to align with the unrefined visual reality of CCTV footage. Similarly, the Single_AVD model, despite using real-world data, achieved only 60.19% accuracy. This suggests that limited-scale real data is insufficient to capture the diverse crowd densities and lighting variations present in the target domain.

However, the multi-source aggregation strategy (TrainC_Comb) successfully mitigated these limitations, achieving the highest accuracy and a significant Recall of 94.28%. By exposing the model to a diverse range of violent motions—from sports grappling to surveillance anomalies—the system developed a higher sensitivity to threats. In a safety-critical context, this improvement is vital, as a high-recall system ensures that genuine threats are rarely missed (minimizing False Negatives), even if it incurs a slight trade-off in Precision.

C. Implications and Prospects for Generalization

The experimental results provide important insight into the learning dynamics of the X3D architecture, showing its limitations and potential.

a. The Promise of Data Diversity: The combined model (TrainC_Comb) showed a clear improvement trend, moving from individual lows to the highest cross-domain metrics. This indicates that the limitation lies not in the lightweight X3D architecture's capacity, but in the homogeneity and small scale of the individual training domains. The notable difference between the Single_Movie model (49.51% Accuracy) and the combined model (62.62% Acc) highlights that model failure can be overcome by diversifying the input data. This finding is encouraging: the efficiency of X3D can work well in real-world applications if provided with a diverse training set.

b. Qualitative Analysis and Safety-First Deployment: A qualitative analysis of the failure cases shows that the high Recall (94.28%) reflects a "safety-first" approach. The model effectively identifies rare violent acts, such as those happening in chaotic crowds or poor lighting. However, this sensitivity affects Precision, as the model often flags non-violent, high-motion activities (False Positives). In practice, this is a favorable failure mode, since human intervention can quickly address a false alarm, while missing a real threat (False Negative) is costly. This confirms that the multi-source training strategy prepares the X3D model for immediate alerts.

c. Future Trajectory: The finding that X3D-XS can achieve a high recall rate with high throughput makes a strong case for cost-effectiveness. Since data diversity is the main bottleneck, future work should focus on implementing new learning approaches, such as Unsupervised Domain Adaptation (UDA). UDA, using methods like Adversarial Domain Alignment, can take advantage of the large amounts of unlabelled real-world surveillance footage. This will help X3D align its feature embeddings with the target RLVS domain, effectively closing the accuracy gap (from 62.62% to an acceptable level) without losing speed or incurring the high computational costs associated with heavy 3D architectures.

V. Conclusion

This study thoroughly examined the trade-off between efficiency and generalization using the X3D-XS architecture for video violence detection. We confirmed that the architecture is highly suitable for real-time edge computing, achieving a throughput of up to 191 FPS, which allows for cost-effective multi-stream processing on centralized servers. However, a significant generalization gap was observed during the zero-shot cross-domain evaluation. Training on a single domain resulted in lower accuracy, specifically 49.51% for the MovieFight dataset, indicating that reliance on choreographed or limited-scope data leads to failures in unconstrained environments like RLVS.

Importantly, the results demonstrate that a multi-source aggregation strategy significantly mitigates this failure by increasing Recall to 94.28%, suggesting that the primary limitation lies in data variability rather than the architecture itself. While X3D proves to be an excellent choice for deployment due to its efficiency, its practical success relies on robust data integration. Future research should focus on cost-effective data strategies, particularly Unsupervised Domain Adaptation (UDA) techniques, to align features from diverse sources with the target domain and address the remaining accuracy gap for deployable public safety systems.

References

[1] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. Syst. Man, Cybern. C, Appl. Rev.*, vol. 34, no. 3, pp. 334–352, 2004, doi: https://doi.org/10.1109/TSMCC.2004.829274.

[2] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 2625–2634, doi: https://doi.org/10.1109/CVPR.2015.7298878.

[3] D. Tran *et al.*, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 4489–4497, doi: https://doi.org/10.1109/ICCV.2015.510.

[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the Kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6299–6308, doi: https://doi.org/10.1109/CVPR.2017.502.

[5] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 203–213, doi: https://doi.org/10.1109/CVPR42600.2020.00028.

[6] H. Yao *et al.*, "A survey of video violence detection," *Cyber-Physical Systems*, vol. 9, no. 1, pp. 1–24, 2022, doi: https://doi.org/10.1080/23335777.2021.1940303.

[7] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *2012 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2012, pp. 1–6, doi: https://doi.org/10.1109/CVPRW.2012.6239348.

[8] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 20–36, doi: https://doi.org/10.1007/978-3-319-46484-8_2.

[9] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 6202–6211, doi: https://doi.org/10.1109/ICCV.2019.00630.

[10] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 5, pp. 1–46, 2020, doi: https://doi.org/10.1145/3400026.

[11] M. Soliman, M. H. Kamal, M. A. El-Massry, and M. El-Tarhony, "Violence recognition from videos using deep learning techniques," in *Proc. 9th Int. Conf. Intell. Comput. Inf. Syst. (ICICIS)*, 2019, pp. 80–85, doi: https://doi.org/10.1109/ICICIS46948.2019.9014714.

[12] M. Bianculli, N. Falcionelli, P. Sernani, S. Tomassini, P. Contardo, and A. F. Dragoni, "A dataset for automatic violence detection in videos," *Data in Brief*, vol. 33, p. 106587, 2020, doi: https://doi.org/10.1016/j.dib.2020.106587.

[13] K. Nievas, O. D. Suarez, G. B. Garcia, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Comput. Anal. Images and Patterns (CAIP)*, 2011, pp. 332–339, doi: https://doi.org/10.1007/978-3-642-23678-5_39.

[14] P. Wu, H. Liu, and X. Yao, "A deep learning approach for violence detection in surveillance videos," in *2020 IEEE Int. Conf. Image Process. (ICIP)*, 2020, pp. 2321–2325, doi: https://doi.org/10.1109/ICIP40778.2020.9191024.

[15] A. B. R. Javan, M. D. Namanloo, and M. H. J. Zahed, "Violent human behavior detection from videos using machine learning," *J. Real-Time Image Process.*, vol. 20, no. 1, p. 16, 2023, doi: https://doi.org/10.1007/s11554-023-01308-w.

[16] X. Lin, J. Zhu, and J. Wang, "Efficient and robust violence detection based on X3D-M architecture for edge devices," *Comput. Electr. Eng.*, vol. 108, p. 108660, 2023, doi: https://doi.org/10.1016/j.compeleceng.2023.108660.

[17] H. Chen *et al.*, "Learning domain-invariant spatiotemporal features for unsupervised video domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 789–803, 2024, doi: https://doi.org/10.1109/TPAMI.2023.3323067.

[18] S. K. Dilawari, M. Khan, and M. Al-Rakhami, "A review of domain adaptation for human action recognition," *IEEE Access*, vol. 9, pp. 159938–159954, 2021, doi: https://doi.org/10.1109/ACCESS.2021.3130948.

[19] Z. Xia *et al.*, "Overcoming domain shift in violence detection with contrastive consistency learning," *Big Data Cogn. Comput.*, vol. 9, no. 1, p. 286, 2025, doi: https://doi.org/10.3390/bdcc9110286.

[20] S. Mumtaz, L. Almagrabi, and F. Al-Otaibi, "An efficient deep learning framework for violence detection in edge computing environments," *J. Grid Comput.*, vol. 20, no. 3, p. 25, 2022, doi: https://doi.org/10.1007/s10723-022-09613-3.

[21] R. Zhang *et al.*, "Real-time violence detection with lightweight 3D convolutional neural networks," *IEEE Internet Things J.*, vol. 10, no. 15, pp. 13245–13256, 2023, doi: https://doi.org/10.1109/JIOT.2023.3283997.

[22] M. S. Islam *et al.*, "An efficient violence detection approach for smart cities surveillance system," in *Proc. IEEE Int. Smart Cities Conf. (ISC2)*, 2023, pp. 1–7, doi: https://doi.org/10.1109/ISC257844.2023.10293696.

[23] A. B. Mabrouk and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review," *Expert Syst. Appl.*, vol. 91, pp. 480–491, 2018, doi: https://doi.org/10.1016/j.eswa.2017.09.029.

[24] F. Ullah *et al.*, "Deep learning for violence detection in surveillance videos: A survey," *IEEE Access*, vol. 9, pp. 98622–98642, 2021, doi: https://doi.org/10.1109/ACCESS.2021.3096777.

[25] Y. Xu *et al.*, "Unsupervised domain adaptation for video action recognition: A survey," *ACM Comput. Surv.*, vol. 56, no. 4, pp. 1–36, 2024, doi: https://doi.org/10.1145/3638243.