Can Event Cameras Replace RGB for Violence Recognition? An Empirical Study

Mira Adra
GTD International
Ramonville-Saint-Agne, France
mira.adra@gtd.eu

Sameer Hans
Data Science Department
EURECOM
Biot, France
sameer.hans@eurecom.fr

Jean-Luc Dugelay
Data Science Department
EURECOM
Biot, France
jean-luc.dugelay@eurecom.fr

***Abstract*—Violence recognition is dominated by abrupt temporal dynamics rather than static appearance. This creates a dilemma for surveillance systems: while RGB transformers achieve strong performance through explicit temporal modeling, event cameras are increasingly considered as an alternative sensing modality better aligned with the motion-centric nature of violent actions. To investigate whether event-based sensing can approximate powerful RGB pipelines, we first establish a strong RGB baseline by enhancing a frozen VideoMAE encoder with lightweight temporal modeling. We then compare this baseline against a state-of-the-art event-based model and the most recent vision model on the UCF-Crime benchmark dataset using simulated RGB–event pairs. Our results show that event-based recognition remains competitive while being substantially more efficient. We further analyze the relationship between modalities by fusing their prediction scores and show that RGB and event models capture complementary information. Finally, by unifying RGB and event data in a shared optical-flow domain, we find comparable performance when reduced to motion-only cues. Overall, our findings indicate that while event cameras do not yet surpass strong RGB transformer pipelines, they constitute an efficient and complementary modality for violence understanding.**

***Index Terms*—Violence Recognition, Event Cameras, Vision Transformers, Optical Flow .**

I. Introduction

Violence recognition is a critical task in public safety and surveillance. Although general action recognition has advanced substantially, detecting violence remains challenging due to its abrupt and highly dynamic temporal characteristics. As a result, effective violence recognition relies primarily on temporal dynamics rather than static spatial appearance. Event cameras [1], or dynamic sensors, are increasingly adopted in motion-driven surveillance, as their asynchronous pixel-level encoding emphasizes dynamic activity while discarding redundant appearance information. Their high temporal resolution, low latency, and privacy-preserving nature make them well suited for real-world surveillance. Despite these advantages, RGB-based transformers remain dominant in video understanding. Architectures such as VideoMAE have recently emerged as powerful self-supervised backbones capable of extracting rich spatio-temporal representations and achieving strong performance with effective temporal modeling. This motivated us to investigate whether event-based systems can realistically match state-of-the-art RGB pipelines for violence recognition.

To address this, we first establish a strong RGB baseline by attaching lightweight temporal heads to a frozen VideoMAE encoder and show that a TCN head offers the best balance between accuracy and efficiency. We then compare this optimized RGB model with an event-based architecture, SpikingFormer [2], and a recent vision model, VideoPrism [3], to assess the strengths of frame-based and event-driven sensing. Importantly, this work is a controlled empirical study focused on modality behavior rather than proposing a new architecture. Our objective is to analyze trade-offs between RGB and event-based representations under identical conditions. While the event-based model achieves slightly lower accuracy, it offers substantial gains in speed and efficiency, highlighting a clear efficiency–accuracy trade-off. Next, we evaluate modality complementarity by fusing predictions, observing consistent improvements over either model alone. Finally, we conduct an optical-flow experiment, mapping both RGB and event data into a shared motion domain and evaluating them using the same network. We show that both modalities achieve comparable performance when reduced to motion-only representations, suggesting that they capture similar motion dynamics for violence recognition.

II. Related Works

A. RGB

Automatic violence detection from video has been widely studied in computer vision, especially for surveillance, crowd monitoring, and public safety applications. Early approaches relied on handcrafted motion features; more recently, deep learning-based RGB video models (3D CNNs [4], CNN-LSTM [5], or two-stream architectures [6]) have become the dominant paradigm. Recent works [7] demonstrates a shift toward Transformer architectures, which model global dependencies via self-attention. Vision Transformers (ViT) [8] have shown significant gains in video understanding tasks by treating the sequence of patches as tokens and capturing long-range temporal relations. A Temporal-Aware Transformer approach [9] integrates MobileNetV2 for frame-level spatial feature extraction with two alternative temporal mechanisms: a BiLSTM module and a Transformer-based sequence encoder

(TransformerSeq). While the CNN–BiLSTM pipeline captures local temporal evolution, the TransformerSeq leverages self-attention to model long-range temporal dependencies more effectively, enabling improved discrimination between violent and non-violent behaviors. Their work highlights two important trends in RGB violence detection: (1) the move toward computationally efficient architectures suitable for real-time deployment, and (2) the increasing use of attention-based temporal models to better capture complex motion dynamics in surveillance scenarios. Still, RGB-based methods often suffer in challenging surveillance conditions, such as low-light, blur, occlusion, and motion saturation. This limitation motivates exploring more motion-centric modalities.

B. Event

Motivated by the motion-driven nature of violent actions, event cameras have attracted growing attention in video understanding and surveillance research, as they capture scene dynamics through asynchronous brightness changes rather than static appearance information.

1) Spiking Neural Networks (SNNs): Early spiking neural networks (SNNs) were mainly effective for low-level event-based tasks and struggled to scale to higher-level analysis due to unstable training. As event cameras began to be adopted for more complex human-centered tasks, architectural advances enabled scalable learning. In particular, EventTransAct [10] introduced transformer-based modeling for event streams at high computational cost, while SpikFormer [11] proposed a spike-native transformer design that improves efficiency and enables end-to-end event processing. These developments made transformer-based SNNs suitable for high-level surveillance applications.

2) Violence detection: These architectural advances enabled progress toward high-level event-based video understanding, including surveillance tasks such as abnormality and violence detection. In [12], Chen et al. detected abnormal activity using optical-flow histograms from event streams, while Annamalai et al. [13] represented events as memory surfaces and employed a convolutional GAN for anomaly detection. For violence recognition, Bullying10K [14] applied SEW-ResNet to classify interpersonal aggression in event-based videos. However, progress remains constrained by dataset limitations. NeuroAED [12] and EvAn [13] datasets contain only around 150 and 130 sequences, respectively, and focus on simple anomalies, while Bullying10K is limited to close-range bullying scenarios. This lack of large-scale paired RGB–event violence datasets hinders fair cross-modal comparison, motivating the need to unify RGB and event representations to enable meaningful evaluation between event-based and RGB models under motion-dominated conditions.

III. Methodology

A. Vision Backbone Models

In this section, we describe the vision backbones and our proposed design. We use two models: (1) VideoMAE, as a benchmark to evaluate and select the most effective temporal head, and (2) VideoPrism, a recent state-of-the-art video foundation model used for comparison with event-based models.

We design a unified framework for temporal head evaluation with two components: (1) spatiotemporal feature extraction using a frozen vision backbone, and (2) lightweight temporal modeling via three different temporal heads. Figure 1 illustrates the overall pipeline. We use VideoMAE [15] as our main backbone, a transformer-based architecture for efficient video representation learning via masked autoencoding. Input clips are divided into spatio-temporal tubelets, typically consisting of two consecutive frames and a 16×16 spatial patch, each forming a token embedding. A stack of transformer layers with self-attention processes these tokens to capture spatial and short-range temporal dependencies.

Beyond VideoMAE, we further consider VideoPrism [3], a recent large-scale video foundation model trained on diverse web-scale data. Unlike VideoMAE, used to explore temporal modeling strategies, VideoPrism evaluates the generalization of the proposed framework. Specifically, we apply the best temporal configuration from VideoMAE to VideoPrism without re-evaluating other temporal heads.

B. Temporal Heads

While VideoMAE captures spatio-temporal features, its temporal modeling remains limited to short local cues. Fine-tuning the full backbone is computationally expensive (86M parameters), so we freeze it and evaluate three lightweight temporal heads operating on VideoMAE embeddings. For all heads, we reshape the backbone output into per-tubelet embeddings $[B, T', D]$, where B is the batch size, T' the number of tubelets, and D the hidden dimension. We first evaluate a 1D CNN, where depthwise temporal convolutions capture short-range motion patterns, followed by adaptive average pooling. We then consider a BiLSTM, in which a bidirectional LSTM with attention models long-range temporal dependencies and aggregates informative time steps. Finally, we test a Temporal Convolutional Network (TCN) that uses causal, dilated depthwise convolutions to capture multi-scale temporal dynamics, combined with residual connections and global pooling for stable learning. At the end of each temporal head, we use a two-layer MLP for binary classification.

C. Event Data Simulation

Due to the lack of large-scale paired RGB–event datasets, prior works commonly rely on simulators to generate synthetic event streams from RGB videos.Such setup allows us to generate aligned RGB–event pairs for direct comparison, which are not available in existing real-world event datasets. In this paper, we adopt the DVS-Voltmeter simulator [16] to generate a large-scale event version of UCF-Crime. DVS-Voltmeter is a physics-based simulator that mimics the behavior of real DVS sensors and produces temporally continuous event streams, which avoids artificial temporal resets caused by frame-based differencing. We evaluate multiple parameter settings and find that configuration (c) is optimal (Fig. 2), as it highlights motion

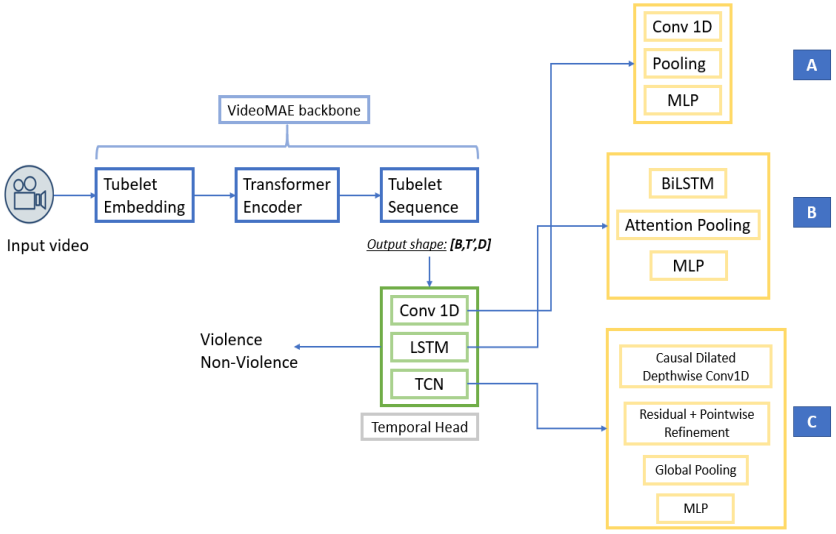

Fig. 1. Proposed architecture.

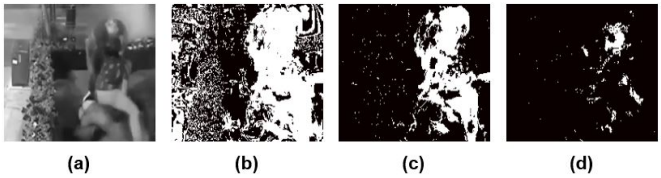

Fig. 2. Simulated event accumulations using different DVS-Voltmeter parameters.

regions while suppressing background activity. The parameter details are provided in Section 5.1. This choice enables a controlled and fair comparison between RGB temporal models and event-based architectures for violence detection, while preserving the underlying motion dynamics of the original data.

D. Event-based Architecture

We adopt SpikingFormer [2] as our event-based classifier, a benchmark transformer-based spiking neural network for event processing. Its spike-driven self-attention removes non-spike computation from the residual pathway, enabling efficient learning while modeling long-range temporal dependencies. Event data is represented as a tuple (t, x, y, p), where t denotes the timestamp, (x, y) the pixel location, and p the polarity of the intensity change. Since event streams are inherently spike-based, we feed them directly into SpikingFormer as spike tensors. The model converts inputs into spike-based patch embeddings using lightweight spike-driven convolutions, then applies stacked spiking transformer blocks to capture long-range motion via spike-based self-attention and residual connections. The resulting features are aggregated by global pooling and passed to a fully connected layer for classification.

IV. Datasets

For our main comparison, we use the violence-related subset of **UCF-Crime** [17] to generate the RGB-event pairs, focusing on fight, vandalism, and assault categories, yielding 607 violent and 503 non-violent clips. For temporal head selection, we evaluate on the same UCF-Crime subset together with four additional publicly available violence detection datasets commonly used in the literature. **RWF-2000** [18] is a large-scale surveillance dataset collected from YouTube, containing 1000 violent and 1000 non-violent scenes, and is widely used as a benchmark for violence detection. The **Real-Life Violence Situations (RLVS)** dataset [19] also contains 1000 violent and 1000 non-violent videos, capturing diverse real-world street fights across varying environments and lighting conditions. **Surveillance-Camera-Fight (SCF)** [20] is a smaller dataset composed of fight scenes extracted from surveillance-style videos, with samples primarily taken from movies or sports footage. The **Movies** dataset [21] consists of 123 violent and 123 non-violent clips taken from films, featuring more staged and visually clean scenes.

V. Results

A. Implementation Details

Data Preprocessing To reduce noise in UCF videos, we discard the first and last 20 seconds, crop each sequence to 5 minutes, and segment it into 20-second clips for training and evaluation.

Temporal Head For all experiments, we followed a standardized training setup. Videos were sampled into 16 frames per clip with a resolution of 224×224. Models were trained using the AdamW optimizer with an initial learning rate of 3e-4. Best model selection was based on the validation F1.

Event Simulation To generate the event streams, we use the DVS Voltmeter with the following sensor parameters: $k_1 = 1.5$ (event density), $k_2 = 25$ (contrast normalization), $k_3 = 5 \times$

TABLE I
COMPARISON OF FINETUNING VIDEOMAE VS OUR PROPOSED APPROACH.

Model	Metrics	Datasets				
		RWF	RLVS	SCF	Movies	UCF
VideoMAE finetuning	Accuracy [%]	75.75	94	76.1	94.74	89.73
	Precision [%]	76.96	94.59	75	90.48	97.17
	Recall [%]	73.5	93.33	78.3	**100**	83.74
	F1 [%]	75.19	93.96	76.6	95	89.96
	Epoch Time [min]	9.22	5.49	1.37	0.96	9.08
	Throughput [clips/sec]	2.85	4.25	3.75	3	1.42
VideoMAE frozen + 1D CNN	Accuracy [%]	80.7	96.7	84.78	94.7	97.3
	Precision [%]	81.2	97.3	80.77	90.5	96.1
	Recall [%]	78	96	91.3	**100**	99.2
	F1 [%]	79.6	96.6	85.71	95	97.6
	Epoch Time [min]	1.27	1.18	0.15	0.05	1.68
	Throughput [clips/sec]	17.47	19.51	24.49	55.13	7.68
VideoMAE frozen + LSTM	Accuracy [%]	85.3	**97.3**	80.4	97.4	96
	Precision [%]	84.1	95.5	79.2	95	94.5
	Recall [%]	87	**99.3**	82.6	**100**	98.4
	F1 [%]	85.5	97.4	80.9	97.4	96.4
	Epoch Time [min]	2.64	1.85	0.26	0.25	6.38
	Throughput [clips/sec]	8.54	12.64	13.5	11.59	2.03
VideoMAE frozen + TCN	Accuracy [%]	**92.5**	**97.3**	**95.7**	**100**	**97.8**
	Precision [%]	91.7	97.9	88	100	96.1
	Recall [%]	**93.2**	94.7	**95.7**	**100**	**100**
	F1 [%]	92.4	96.3	91.7	100	98
	Epoch Time [min]	1.46	0.84	0.15	0.05	1.73
	Throughput [clips/sec]	15.37	27.79	23.49	54.36	7.47

TABLE II
PERFORMANCE OF THE BEST RGB-BASED MODEL AND THE BEST EVENT-BASED MODEL ON UCF DATASET.

Model	Accuracy [%]	Precision [%]	Recall [%]	F1 [%]	Epoch Time [min]	Throughput [clips/sec]
VideoMAE+TCN	97.8	96.1	**100**	98	1.73	7.47
VideoPrism+TCN	**98.2**	**98.36**	98.36	**98.36**	-	-
Spikingformer (Pretrained)	95.60	95	94.8	95	**0.14**	**42.6**

10^{-4} (variance control), and $k_4 = 0.0$, $k_5 = 0.0$, $k_6 = 10^{-5}$ (suppression of brightness and noise artifacts).

SpikingFormer For training, event streams are processed using the SpikingJelly dataset interface, which converts them into voxel grids. SpikingFormer is trained for 100 epochs with batch size 8 and learning rate 10^{-3} using AdamW optimizer, along with a StepLR scheduler (step size = 64, $\gamma = 0.1$).

B. Enhancing VideoMAE with Temporal Heads

Table I summarizes the performance of each temporal head across five datasets. We report accuracy, precision, recall, F1 score, epoch time, and throughput to evaluate both performance and computational efficiency.

The 1D CNN consistently outperforms the fully finetuned VideoMAE baseline while maintaining high efficiency. Across datasets, it improves accuracy by 5–10% and often doubles throughput. It performs particularly well on datasets dominated by short-term, high-intensity motion, such as RWF, SCF, and UCF. The 1D CNN reaches 97.3% accuracy on UCF and 96.7% on RLVS, while achieving the highest throughput (up to 55 clips/sec), making it attractive for real-time surveillance scenarios. However, its limited temporal receptive field reduces performance on longer or more structured violent sequences, such as Movies.

The BiLSTM improves over the 1D CNN on datasets with longer-duration interactions, such as Movies and RLVS. Its sequential modeling captures escalation and multi-person interactions unfolding over time, reaching 97.4% accuracy on Movies. It also achieves high recall across datasets (99.3% on RLVS), indicating its strong sensitivity to detect violent segments. This gain comes at the cost of lower efficiency, with reduced throughput compared to 1D CNN and TCN. Overall, the BiLSTM improves long-range temporal modeling but does not offer the best accuracy–efficiency trade-off.

The TCN head delivers the best and most consistent performance across all datasets. Its causal and dilated convolutions provide a multi-scale temporal receptive field that captures both short and long-range dependencies. As a result, it achieves the highest F1 scores and outperforms both the 1D CNN and BiLSTM, as well as the fully finetuned backbone. In terms of efficiency, TCN remains competitive, reaching up to 54 clips/sec, which is slightly slower than the 1D CNN but significantly faster than the BiLSTM. Across all datasets, TCN emerges as the most effective head, offering the best balance between accuracy, robustness, and computational efficiency.

C. RGB vs Event-Based Models

After selecting the best temporal head for VideoMAE (Section V-B),we further evaluate VideoPrism. We compare the

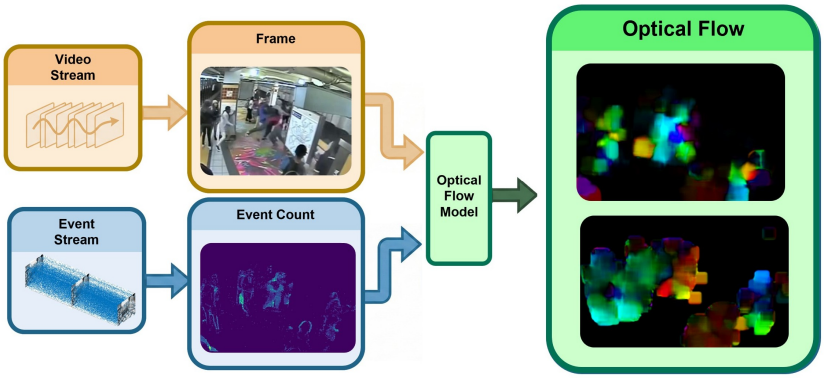

Fig. 3. Unified optical flow domain for RGB and event data.

RGB-based models with the state-of-the-art event transformer, SpikingFormer. All models are trained independently on the same UCF splits and evaluated using identical metrics. As shown in Table II, VideoPrism+TCN achieves the highest accuracy in violence recognition (98.2%), reflecting the advantage of rich spatial appearance when combined with temporal modeling. In contrast, SpikingFormer achieves competitive accuracy (95.1%) and runs at 42.6 clips/s, compared to only 7.47 clips/s for the RGB model, and completes an epoch in 14 seconds vs 1.73 minutes. This makes the event-based approach 12.4× faster and far more suitable for resource-constrained or real-time surveillance systems, where latency and energy consumption are critical. Thus, RGB provides maximal accuracy, whereas event sensing offers lightweight, high-speed inference, showing that spatial and event-driven motion cues offer complementary advantages in real-world violence detection.

D. RGB-Event Fusion

To examine whether RGB and event models are complementary or learning the same underlying information, we perform a training-free late-fusion experiment. By combining the outputs of the two independently trained models, we can evaluate each modality's intrinsic contribution without introducing additional learned parameters. We first align the test set so that both networks process exactly the same scenes. Each model then produces its final logits, denoted as $\mathbf{z}_{\text{RGB}} \in \mathbb{R}^2$ and $\mathbf{z}_{\text{EVT}} \in \mathbb{R}^2$. The final prediction is obtained by linearly combining the pre-softmax outputs of the RGB and event models with a weighting factor $\alpha \in [0, 1]$ and selecting the class with the highest score. We select $\alpha = 0.5$ through hyperparameter tuning on a validation set, suggesting that RGB and event cues have a balanced contribution. The fusion yields higher performance for both modalities, improving RGB accuracy from 97.8% to 98.21% (+0.9%) and event accuracy from 95.60% to 98.21% (+2.61%). This improvement demonstrates that event representations capture complementary motion-driven information, as the simulated data preserves the characteristics of event-based sensing rather than collapsing into an RGB-equivalent signal. This provides empirical evidence of the validity of simulated data for evaluating event-based models and reinforces the role of event cameras as a robust and complementary modality for violence recognition.

E. Cross-Domain Unification via Optical Flow

So far, RGB and event data have been analyzed separately, although both encode the same underlying motion dynamics. To analyze both modalities under a shared motion representation, we unify both modalities within a shared optical-flow domain, allowing evaluation using motion cues alone. We convert event streams into event-count frames, a lightweight representation that accumulates events over fixed slices while preserving motion frequency. Then, we apply the same Farnebäck optical-flow estimation to both inputs and feed the resulting flows into the same VideoMAE model, so that both modalities are evaluated under the same conditions. Using this unified representation, RGB-derived optical flow achieves 79.0% accuracy, while event-derived optical flow reaches 77.6%, yielding only a 1.4% gap. Event-based flow appears noisier and more blob-like due to event sparsity and sensor noise, which may explain the slight difference. Overall, these results suggest that both modalities capture similar motion information when represented in the same way, highlighting the role of motion as a shared cue for violence recognition.

VI. Conclusions

This work presents a controlled empirical study of RGB and event-based modalities for violence recognition, where we present a detailed study of temporal modeling for RGB-based violence recognition and a comparative evaluation of event-based sensing for the same task. Using a frozen VideoMAE backbone, we evaluated several lightweight temporal heads and found that the TCN consistently provides the best balance between accuracy and computational efficiency across six datasets. These results confirm that effective violence recognition relies primarily on temporal modeling and multi-scale temporal reasoning to capture irregular and dynamic patterns. To address whether event-based systems can approximate strong RGB transformer pipelines, we evaluated an event-based SpikingFormer model. While its accuracy approaches that of the best RGB configuration, it offers significant gains

in speed and energy efficiency, highlighting event cameras as a promising modality for real-time surveillance, although still falling slightly behind RGB transformers on standard benchmarks. A natural extension of this work is the collection of a real RGB–event paired violence dataset under challenging conditions such as low light and motion blur for real-world evaluation, where RGB cameras degrade while event sensors remain reliable.

REFERENCES

[1] M. Adra, S. Melcarne, N. Mirabet-Herranz, and J.-L. Dugelay, "Event-based solutions for human-centered applications: A comprehensive review," 2025. [Online]. Available: https://arxiv.org/abs/2502.18490

[2] C. Zhou, L. Yu, Z. Zhou, Z. Ma, H. Zhang, H. Zhou, and Y. Tian, "Spikingformer: Spike-driven residual learning for transformer-based spiking neural network," *arXiv preprint arXiv:2304.11954*, 2023.

[3] L. Zhao, N. B. Gundavarapu, L. Yuan, H. Zhou, S. Yan, J. J. Sun, L. Friedman, R. Qian, T. Weyand, Y. Zhao, R. Hornung, F. Schroff, M.-H. Yang, D. A. Ross, H. Wang, H. Adam, M. Sirotenko, T. Liu, and B. Gong, "VideoPrism: A foundational visual encoder for video understanding," in *International Conference on Machine Learning (ICML)*, 2024.

[4] W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, and A. Wang, "A novel violent video detection scheme based on modified 3d convolutional neural networks," *IEEE Access*, vol. 7, pp. 39 172–39 179, 2019.

[5] R. Halder and R. Chatterjee, "Cnn-bilstm model for violence detection in smart surveillance," *SN Computer Science*, vol. 1, 06 2020.

[6] M. Mahmoud, B. Yagoub, M. F. Senussi, M. Abdalla, M. S. Kasem, and H.-S. Kang, "Two-stage video violence detection framework using gmflow and cbam-enhanced resnet3d," *Mathematics*, vol. 13, no. 8, 2025. [Online]. Available: https://www.mdpi.com/2227-7390/13/8/1226

[7] A. Alshalawi, W. Abdul, and G. Muhammad, "Advanced detection of violence from video: Performance evaluation of transformer and state of the art of convolution of neural network transformer," *IEEE Access*, vol. 13, pp. 74 200–74 216, 2025.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[9] R. Chatterjee, R. Roy Choudhury, M. Kumar Gourisaria, S. Banerjee, S. Dey, M. Sahni, and E. León-Castro, "Temporal-aware transformer approach for violence activity recognition," *IEEE Access*, vol. 13, pp. 70 779–70 790, 2025.

[10] T. de Blegiers, I. R. Dave, A. Yousaf, and M. Shah, "Eventtransact: A video transformer-based framework for event-camera based action recognition," 2023.

[11] Z. Zhou, Y. Zhu, C. He, Y. Wang, S. Yan, Y. Tian, and L. Yuan, "Spikformer: When spiking neural network meets transformer," 2022. [Online]. Available: https://arxiv.org/abs/2209.15425

[12] G. Chen, P. Peng, G. Li, F. Jiang, J. Han, and S. Zhou, "Neuroaed: Towards efficient abnormal event detection in visual surveillance with neuromorphic vision sensor," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 923–936, 2020.

[13] L. Annamalai, A. Chakraborty, and C. S. Thakur, "Evan: Neuromorphic event-based sparse anomaly detection," *Frontiers in Neuroscience*, vol. 15, 2021.

[14] Y. Dong, Y. Song, Z. Zhou, Y. Zhu, C. He, Y. Sun, Y. Wang, and Y. Tian, "Bullying10k: A large-scale neuromorphic dataset towards privacy-preserving bullying recognition," *Advances in Neural Information Processing Systems*, vol. 36, pp. 1923–1937, 2023.

[15] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Advances in Neural Information Processing Systems*, 2022.

[16] S. Lin, Y. Ma, Z. Guo, and B. Wen, "Dvs-voltmeter: Stochastic process-based event simulator for dynamic vision sensors," in *European Conference on Computer Vision*. Springer, 2022, pp. 578–593.

[17] 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*, 2018, pp. 6479–6488.

[18] 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 (ICPR)*, 2021, pp. 4183–4190.

[19] 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 (ICICIS)*, 2019, pp. 80–85.

[20] S. Akti, G. A. Tataroglu, and H. K. Ekenel, "Vision-based fight detection from surveillance cameras," in *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, Nov. 2019, p. 1–6. [Online]. Available: http://dx.doi.org/10.1109/IPTA.2019.8936070

[21] P. Sonawane, "Movies-violence/non-violence videos," Last accessed: 10 September 2025.