IET Image Processing

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Swin-3DART: An Efficient and Robust Lightweight Transformer for Video Anomaly Detection with TG-RGB$^+$

Intissar Ziani[1] | Gueltoum Bendiab[2] | Mourad Bouzenada[1] | Meriem Guerar[3]

[1]MISC Laboratory, Department of Computer Science and its Applications, Faculty of NTIC, University of Constantine 2 - Abdelhamid Mehri, Constantine, Algeria | [2]Department of Electronics, LIRE Laboratory, University of Frères Mentouri, Constantine, Algeria | [3]DIBRIS, University of Genoa, Genova, Italy

Correspondence: Meriem Guerar (meriem.guerar@unige.it)

Received: 1 July 2025 | **Revised:** 1 February 2026 | **Accepted:** 9 February 2026

ABSTRACT

Video anomaly detection is vital for public safety but remains challenging due to complex motion patterns, limited robustness to motion-related perturbations, and the heavy computation demands of modern transformers. To address these challenges, swin-3DART is introduced as a unified framework that improves both efficiency and resilience. First, the proposed TG-RGB$^+$ modality enhances RGB frames with temporal-gradient motion cues, improving motion sensitivity. Second, it designs T-GAP (temporal gradient adaptive perturbation), which generates worst-case perturbations to expose vulnerabilities and strengthen the model through adversarial training. Third, an adversarial defence mechanism is embedded to ensure robustness, achieving consistently low attack success rates (4%–7%) across datasets. Finally, the framework incorporates the 3DART (3D adaptive receive transformer), which reduces memory footprint by ~12% and FLOPs by ~11.9%, making it suitable for deployment in real-time surveillance or edge computing scenarios. Comprehensive evaluations show that swin-3DART achieves state-of-the-art AUCs of 95% on UBI fights, 86% on UCF-crime, and 99% on RLVS. These results highlight swin-3DART's potential as an efficient and robust solution for real-time, safety-critical video anomaly detection.

1 | Introduction

Automated violence detection in surveillance video has become increasingly critical for public safety, particularly as modern societies seek real-time responses to potential threats. This need is underscored by the rapid expansion of surveillance infrastructure worldwide. In Europe alone, the video surveillance market is projected to grow at a compound annual growth rate (CAGR) of 10.9% to 13.9% between 2024 and 2030 [1], reflecting a surge in camera installations across cities, transportation hubs and sensitive public areas. As surveillance feeds multiply, manual monitoring becomes infeasible, creating a pressing need for intelligent systems that can detect violent or abnormal behaviour in real time [2, 3]. Beyond urban infrastructure, surveillance systems are also widely used in workplaces, hospitals, airports and other public spaces, where they serve as both a deterrent and a forensic tool. However, the inefficiency, cost and error-prone nature of manual analysis make it unsuitable for real-time, large-scale use [4, 5]. As a result, video-based violence detection has emerged as a central focus in computer vision and artificial intelligence research [3, 5, 6].

However, accurately detecting violence in videos is challenging due to a combination of factors. Varying illumination, diverse camera angles, occlusions and cluttered or dynamic backgrounds can all obscure violent actions, while the subtle and often brief nature of some violent behaviours makes them difficult to distinguish from normal activities [3, 5, 6]. These challenges are further amplified in the context of real-time video anomaly detection (VAD) for public safety, where existing methods suffer from high computational costs, making them difficult to deploy on edge devices. They also lack sufficient robustness

against sophisticated, motion-aware adversarial perturbations [5]. Therefore, designing VAD systems that are both efficient and reliable in real-world conditions remains a significant research challenge.

Unlike standard action recognition, where activities are often well-defined and structured, violence detection requires identifying spatiotemporal anomaly detection and unpredictable behaviour. By analysing changes across both space (location, appearance) and time (motion, dynamics)[5, 6]. Furthermore, violence can manifest in multiple visual forms, such as physical aggression, shooting, explosions and armed attacks, making it necessary for detection models to generalize effectively across different scenarios [4, 7]. Despite recent progress, many state-of-the-art methods still perform poorly on challenging datasets for their real-world deployment [3–7]. To address this particular challenge, there is a need for a more reliable, efficient, and generalizable model that improves detection capabilities, ensuring its effectiveness in real-world scenarios [4]. Leveraging advancements in artificial intelligence, particularly deep learning (DL) and computer vision, can significantly enhance the identification of violent incidents, enabling faster response times and improving overall security [4, 5, 8].

In this work, a novel weakly supervised framework is proposed to balance detection performance, computational efficiency and resilience to environmental variability. The method detects violent events in videos, including deliberate actions like physical aggression and unintended ones like explosions, by addressing four core challenges: first, scene diversity, which undermines reliance on static spatial cues; second, motion complexity, where rapid, non-rigid motion and camera jitter obscure meaningful temporal patterns; third, adversarial vulnerability, wherein small perturbations can mislead fragile models; and fourth, domain shift, which causes model performance to degrade across datasets with different illumination, viewpoints and environments.

To solve these issues holistically without incurring high computational costs, an early fusion strategy is adopted that combines Temporal Gradient (TG) and RGB features into a unified input modality TG-RGB$^+$. This fusion enables the model to capture subtle, fast-moving actions and appearance context in a single representation, which is critical for detecting violent behaviours. The enriched stream is fed into a video swin transformer [9], enabling robust feature extraction with global context modeling. Instead of relying solely on complex feature extractors, swin-3DART incorporates temporal gradient adaptive perturbation (T-GAP), an adversarial attack strategy that simulates worst-case motion disturbances to improve the model's robustness. Complementing this, the parameter-free 3D adaptive receive transformer (3DART) dynamically adjusts spatial and temporal receptive fields based on internal feature dynamics. This adaptive mechanism strengthens spatiotemporal representations while simultaneously reducing computational redundancy and memory usage, these efficiency gains are particularly meaningful in the context of pure 3D video transformers, which are typically computationally intensive. Together, these components form swin-3DART, a robust and energy-efficient architecture that achieves strong accuracy, cross-domain generalization and real-time performance, making it well-suited for deployment in intelligent, resource-constrained surveillance environments. The main contributions of this work are summarized below:

1. A TG-RGB$^+$ modality is proposed that integrates temporal gradient (TG) features with RGB frames in a single stream. By combining TG, which emphasizes motion dynamics, with RGB's rich spatial information, this modality effectively captures both fine-grained spatial details and temporal motion patterns, enhancing the representation of dynamic activities in videos.
2. A temporal gradient adaptive perturbation (T-GAP) is designed as an adversarial attack strategy that exposes the model to carefully crafted perturbations over time. By simulating challenging adversarial conditions during training, T-GAP encourages the model to learn from these examples, thereby enhancing its robustness and improving performance in video-based violence detection, particularly against motion-aware attacks.
3. Swin-3DART is introduced as a unified architecture that combines video swin transformers with the parameter-free 3DART module, enabling efficient and robust violence detection under real-world constraints.

A comprehensive cross-dataset evaluation is conducted to assess the generalization capabilities of the proposed swin-3DART model. Even when trained and tested on different datasets, swin-3DART demonstrates significant improvements over existing methods, which struggle with domain shifts, achieving substantial performance gains in challenging settings. The rest of the paper is structured as follows: Section 2 explores video violence detection solutions, including modalities and methodologies, deep learning-based techniques, adversarial attacks, and TG-RGB$^+$ analysis, with a focus on spatiotemporal data. Section 3 introduces swin-3DART, detailing its architecture and key components. Section 4 describes the experimental setup, datasets, and performance evaluation of the proposed approach. Finally, Section 5 concludes the paper by summarizing key findings, discussing their implications and suggesting directions for future work.

2 | Landscape of Video Violence Detection

Video violence detection has emerged as a critical area of research in computer vision and AI, driven by the increasing need to enhance public safety and monitor sensitive environments. With the proliferation of video data from surveillance systems, social media platforms and public spaces, the ability to automatically identify violent behaviour in real time has become increasingly important. In particular, the increasing demand for reliable and scalable violence detection solutions has positioned the field as a dynamic and rapidly evolving domain, presenting significant opportunities and challenges for researchers and practitioners.

This section explores the landscape of video violence detection solutions, highlighting various modalities and methodologies. It introduces DL-based techniques, adversarial attack strategies, and TG-RGB$^+$ analysis, with a particular focus on their application to spatiotemporal data analysis.

2.1 | Modalities for Video Violence Detection

Video violence detection has increasingly relied on leveraging multiple modalities to enhance accuracy, robustness, and real-time performance in recognizing violent behaviour [5]. Among these, visual modalities are fundamental due to their ability to encode spatial layouts and motion dynamics inherent to violent events. The three most widely adopted visual modalities are RGB frames, optical flow, and temporal gradients (TG) [10, 11], each offering distinct advantages and limitations. RGB frames contain rich spatial information crucial for identifying scene context, objects and posture configurations [7]. However, they inherently lack motion sensitivity, making them insufficient for capturing dynamic cues essential for differentiating violent and non-violent actions. To overcome this, optical flow estimates pixel-level motion vectors between consecutive frames, enabling explicit modelling of movement trajectories and velocities [12]. Despite its effectiveness, optical flow suffers from high computational cost and sensitivity to camera motion, often leading to spurious motion artefacts [10, 12].

To address efficiency concerns, temporal gradient (TG) features have emerged as a lightweight alternative that highlights abrupt intensity changes across frames, capturing sudden motion bursts often indicative of violence [13]. However, despite their efficiency and simplicity, temporal gradient features are inherently limited in representing spatial structure and fail to capture subtle or gradual motions, reducing their discriminative power when used in isolation.

Recent advances have increasingly shifted toward multimodal frameworks that aim to integrate spatial and temporal information to improve performance across diverse scenarios [14, 15]. Notable approaches such as I3D [14], SlowFast [16], and MoViNets [17] (mobile video networks) incorporate multiple visual streams to capture both appearance and motion, achieving significant gains on benchmarks like UCF-Crime and UBI Fights. However, these architectures often incur increased complexity due to reliance on dual-stream processing and fusion strategies, which can hinder deployment in latency-sensitive environments [18].

Beyond these computational hurdles, recent advances in action recognition have shifted from simple temporal sequence modeling toward more expressive geometric representations of human motion. This shift toward finer-grained and structurally informed descriptions has led to growing interest in structural modalities. For instance, EMS-TAGCN [19] models human actions using skeletal graphs, whereas PRG-Net [20] captures body dynamics through 3D point-relation graphs. Although these representations provide high precision in controlled environments, their reliance on explicit structure makes them vulnerable in real-world violent scenarios. In particular, PRG-Net [20] shows that point-cloud disorder and rapid dynamics blur boundaries in fast-moving actions. Similarly, skeleton-guided approaches [19] often fail under heavy occlusions or when victims fall to the ground. In such cases, pose estimators are unable to reliably detect the joints required to construct the underlying graph.

In response to these structural limitations, the proposed TG-RGB$^+$ modality offers a more robust representation for unstructured environments. By performing early fusion, summing temporal gradients directly into the RGB manifold, the framework provides the model with intrinsic motion awareness. Unlike skeletal or point-cloud modalities that collapse when the human structure is obscured, TG-RGB$^+$ preserves the pixel-level motion energy of the entire scene. This allows for the detection of non-human anomalies (e.g. vehicular collisions or environmental hazards) that lack a defined skeletal topology and are thus invisible to graph-based human action recognizers.

2.2 | Deep Learning for Video Violence Detection

Understanding spatiotemporal representations is central to modern video analysis, but it continues to face two significant challenges. First, spatiotemporal redundancy persists: consecutive video frames often exhibit minimal differences, yet traditional models process each frame uniformly. Second, highly dynamic dependencies: object interactions evolve across temporal and spatial contexts. Inspired by the success of transformers in natural language processing [21], computer vision researchers have proposed architectures, such as ViViT [22] and VioNet [23], which improve long-range modelling compared to conventional CNNs. These models have achieved strong performance in video-based tasks such as action and violence recognition.

In the specific context of violence recognition, recent methods have leveraged Transformer variants to improve accuracy: mixture of experts (MoE) architectures have reached 92.4% [24], and cascaded attention pipelines using motion and pose features have achieved up to 90.9% [25]. While recent frameworks, such as [26, 27], rely on bidirectional GRUs to capture temporal context, their recursive nature processes frames as a linear sequence, creating a bottleneck that limits the correlation of distant temporal events and introduces significant latency. Despite these advances, transformer- and GRU-based models often incur prohibitive computational costs, especially in real-time or resource-constrained settings. To address this, several efficiency-oriented architectures have emerged. For instance, swin transformers [28] apply local attention within windows to reduce complexity; HST-MRF [29] introduces heterogeneous multi-receptive fields to capture multi-scale structure; and EfficientViT [30] processes subsets of features through grouped attention heads. Nevertheless, these approaches rely on fixed attention patterns and often fail to adapt to motion-salient dynamics, resulting in wasted computation on static or low-information regions.

These limitations motivate 3DART, a parameter-free module designed to improve spatiotemporal modeling efficiency in video transformers. Unlike fixed-window attention schemes (e.g. swin) or static multi-branch receptive field strategies (e.g. HST-MRF), 3DART adaptively modulates spatial and temporal attention spans based purely on internal feature dynamics, without introducing any additional learnable parameters or relying on handcrafted saliency or motion cues. This enables the model to implicitly emphasize regions with significant motion or appearance change while avoiding redundant computation in static or low-information areas. By dynamically adjusting its receptive field, 3DART enhances discriminative feature learning and reduces the floating-point operations (FLOPs), making it ideal for real-time, resource-constrained video understanding

tasks. When integrated into the swin transformer backbone, 3DART achieves a 12% reduction in FLOPs and memory footprint while delivering state-of-the-art accuracy on multiple benchmark datasets for violence recognition. Compared to ViViT, EfficientViT, and HST-MRF, the method achieves a superior efficiency-to-accuracy trade-off, demonstrating its advantage in resource-constrained environments. Although evaluated on violence recognition, 3DART is task-agnostic and applicable to other video understanding tasks, such as action recognition or video summarization.

To the best of our knowledge, 3DART is the first transformer module to introduce dynamic 3D receptive field adaptation within a hierarchical vision transformer. This innovation enables the model to allocate attention flexibly across space and time, responding to motion and structural variation without additional parameters or handcrafted priors. Unlike models relying on fixed windowing or static multi-branch attention, 3DART offers a new direction for efficient, adaptive video modelling.

This combination of zero-parameter adaptability, flexible 3D receptive fields and seamless integration into hierarchical transformers establishes 3DART as a timely and impactful contribution to efficient video understanding.

2.3 | Motion-Aware Adversarial Threats in Video Models

Adversarial learning has rapidly evolved into a foundational pillar across modern machine learning, driving breakthroughs in video anomaly detection, speech recognition, autonomous driving, human action recognition and robust classification tasks [31–33]. By strategically injecting adversarial examples during training, these approaches enable models to fortify their resilience against malicious or unexpected inputs, significantly boosting real-world generalization and transferability [33]. With the advent of transformer-driven anomaly detectors and foundation-model pretraining, ensuring cross-modal robustness against adaptive perturbations has become more urgent than ever. In the context of video violence detection, adversarial learning serves as a high-impact tool that pressures models to confront challenging, deliberately crafted perturbations, forcing them to sharpen their decision boundaries and enhancing their adaptability in ambiguous, dynamic and noisy visual environments.

Several works have paved the way in video adversarial research. For example, Chen et al. [34] propose replacing a few consecutive frames with dummy frames, perturbing only these, to maximize the difference between the classifier's output on clean versus perturbed videos. Kim et al. [35] introduced temporal-aware attacks that extend image-based models to video by minimizing feature similarity across neighbouring frames. However, these approaches do not differentiate between high-motion and low-motion regions, increasing the risk of generating visually detectable artefacts, especially in videos with minimal movement. To improve robustness, Yin et al. [36] explored multimodal data transformations incorporating Gaussian noise, shape warping, and colour distortions to craft adversarial examples that remain effective under varied visual conditions. Yet, these strategies apply transformations uniformly across the video, without accounting for localized temporal motion patterns, limiting their stealth and precision in real-world scenarios (e.g. static surveillance feeds or subtle behavioural anomalies).

In safety-critical applications, such as public-space surveillance, autonomous vehicles and medical video analysis, stealthy motion-adaptive attacks pose catastrophic risks if left unaddressed. Moreover, recent studies on zero-shot vulnerability [40] reveal that even foundation models can be deceived by context-aware perturbations. The evolving landscape is summarized in Table 1, which compares key attack methods across critical features such as temporal-spatial coupling, flow alignment, motion adaptation, and anomaly targeting.

The final column in Table 1, "constraint type," specifies how each method restricts perturbations. Traditional approaches apply fixed ℓ_p norm constraints (e.g. ℓ_2, ℓ_∞), limiting the overall magnitude of adversarial noise. Some methods enforce temporal smoothness for realism. In contrast, the proposed method introduces a motion-adaptive ℓ_p constraint that varies the perturbation budget across frames, allowing stronger perturbations in dynamic regions while preserving imperceptibility in static ones. This principled design better aligns with real-world motion patterns in anomaly detection.

Despite these advances, the field still lacks a comprehensive, motion-adaptive adversarial attack framework explicitly designed for next-generation video anomaly detection systems. In light of these missing elements, a T-GAP framework is proposed, representing a truly motion-aware, context-aware adversarial framework that pushes beyond uniform transforms and static flow alignment. This work addresses the gap by introducing a novel temporal gradient adaptive perturbation (T-GAP) strategy that dynamically tailors adversarial signals to motion magnitudes within video sequences, making attacks stealthier, more effective, and broadly transferable across diverse, widely used anomaly detection benchmarks. To the best of our knowledge, this is the first systematic exploration of motion-adaptive adversarial attacks tailored specifically to challenge modern anomaly detection architectures, opening new avenues for understanding and fortifying video surveillance systems.

Table 2 provides a comprehensive summary of deep learning-based video anomaly detection methods across diverse modalities, architectural designs, and datasets. It illustrates the dominance of 3D CNN-based models (e.g. I3D, C3D) and transformer-based methods (e.g., ViViT, HST-MRF) in capturing spatiotemporal patterns. Notably, audio-text fusion remains under-explored despite its potential, while RGB+Flow modalities remain widely adopted due to their complementary temporal cues. Datasets like UCF-Crime are frequently used as benchmarks, reinforcing their value in comparative evaluations. However, several models face challenges in computational efficiency (e.g. ViViT), robustness to noisy inputs (e.g. raw skeletons), and generalization across scenes. These gaps highlight the need for adaptive, lightweight and multimodal fusion strategies, –a direction pursued in the proposed TG-RGB$^+$ and the 3DART framework.

TABLE 1 | Comparative feature coverage of video evasion attacks.

Attack	Spatiotemporal	Flow alignment	Motion adaptive	Targets anomaly	Constraint type
TSA (temporal shift)[37]	✓	✗	✗	✗	Temporal shift
ATN (Adv. transfer net)[38]	✓	✗	✗	✗	Learned perturbations
Temporal UAP [35]	✓	✗	✗	✗	Universal temporal ℓ_p
Flow-aligned attack [39]	✓	✓	✗	✗	Optical-flow aware
Chen et al. [34]	✓	✗	✗	✗	Frame replacement
Yin et al. [36]	✓	✗	✗	✗	Global transformations
T-GAP (ours)	✓	✓	✓	✓	Motion-adaptive ℓ_p

3 | Proposed Anomaly Detection Model

For video-based anomaly detection, a deep transformer network is proposed as the baseline model for video classification, which processes videos frame by frame while effectively capturing essential temporal dynamics. To improve motion recognition, temporal gradients (TG) are incorporated, embedding motion information directly within RGB frames and allowing a single-stream model to integrate temporal details efficiently. Furthermore, an adversarial temporal gradient adaptive perturbation (T-GAP) is introduced, representing a novel adversarial attack technique that utilizes temporal gradients and optical flow to create perturbations focused on disrupting motion patterns rather than introducing conventional pixel-level noise. This method enhances video anomaly detection by increasing robustness against adversarial attacks while maintaining the integrity of crucial motion dynamics. More details of the proposed methods will be given below:

3.1 | TG-RGB$^+$ Approach

There is extensive research on using multimodality (e.g. combining RGB and optical flow) or unimodality (e.g. using only RGB or only optical flow) approaches in video models. Various methods have explored these modalities for abnormal action recognition. Since the model processes videos on a frame-by-frame basis, capturing temporal dynamics within each frame is crucial. To achieve this, temporal gradients (TG) are incorporated, which encode motion cues directly within the RGB frames. Unlike traditional optical flow, temporal gradients enhance the representation of dynamic regions by capturing pixel-wise intensity variations across consecutive frames, as illustrated in equation 1.

$$\mathrm{TG}_t(x,y) = I_t(x,y) - I_{t-1}(x,y) \tag{1}$$

This operation provides a lightweight and direct measure of motion energy, highlighting regions undergoing significant intensity variation, a strong indicator of motion or interaction. As a result, TG effectively captures temporal changes without requiring additional pre-processing networks, making it suitable for real-time and large-scale video analysis.

Instead of explicitly computing motion vectors such as optical flow, temporal gradients directly reflect changes in pixel values over time, highlighting moving regions. This allows the model to incorporate temporal information within a single TG-RGB$^+$ (temporal gradient RGB plus) stream, improving its ability to recognize motion patterns relevant to violence detection.

In contrast, this paradigm employs a unified representation learning approach by combining appearance and motion cues within a single modality. Instead of relying on separate input streams, TG is embedded into RGB data, enriching it with motion information. This approach enhances the model's ability to extract interdependencies between spatial and temporal features, leading to a more comprehensive understanding of video content. As indicated in Equation (2) where R_t, G_t, B_t are the original RGB colour channels at frame t, and $\mathrm{TG}(x,y,t)$ is the temporal gradient function.

$$I_t^{\mathrm{aug}}(x,y) = [R_t(x,y), G_t(x,y), B_t(x,y)] \oplus \mathrm{TG}_t(x,y) \tag{2}$$

Through this additive fusion, temporal variations are embedded directly into each color channel, allowing the network to learn joint spatiotemporal dependencies in a compact and efficient manner. Unlike concatenation, which expands the feature space, the element-wise summation preserves the original input dimensionality, ensuring that the model remains lightweight while enhancing motion awareness. This element-wise summation strategy offers several theoretical and practical advantages:

- *Computational efficiency*: Temporal gradient computation and summation operations are linear in complexity ($O(N)$) and require no additional memory allocation for new channels. This allows temporal gradient features to be computed and integrated with existing data streams without increasing the model's memory footprint.
- *Compactness*: The resulting fused frame maintains the same dimensionality as the original RGB input, avoiding an increase in feature space size. This compact design enables seamless incorporation into existing architectures without additional computational or memory overhead, maintaining model efficiency while enriching spatiotemporal representation.
- *Improved feature synergy*: Since both motion and appearance cues are co-located in pixel space, the fusion enables the network to learn joint spatiotemporal patterns directly at the input level.

TABLE 2 | Overview of video anomaly detection deep learning. methods

Model	DL Approach	Modality	Datasets	AUC	Challenges
VioNet (2023) [23]	Framework that integrates Vision Transformer with BiLSTM and 3D-CNN to extract and fuse spatiotemporal features	RGB	Hockey fight, violent flow, movies	97% – 100%	High computational cost; limited adaptability to real-world motion variance.
Cascade transformer (2023) [25]	A framework utilizing human keypoint extraction and change detection, enhanced by a cascade transformer for spatiotemporal feature learning	RGB	RWF-2000 and campus violent video	83.2%, 90.9%	Depends on keypoint accuracy; motion sensitivity issues under occlusion or clutter.
TEVAD (2023) [41]	A weakly supervised framework employs a dual-branch for visual and text extraction followed by a fusion module	RGB + Text	ShanghaiTech, UCF-Crime, XD-Violence, and UCSD-Pedestrians	98%, 84%, 80%, 98%	Relies on paired text; inefficient without text; computationally heavy for fusion.
I3D-AD-AFF-MFL (2024) [42]	A dual-branch Weakly supervised training method, leveraging attention-based redundancy filtering, adaptive feature fusion, and temporal dependency modeling	RGB + optical flow	UCF-Crime and ShanghaiTech	85%, 97%	High fusion complexity; optical flow is slow to compute
Human skeletons and change detection (2021) [43]	A dual-pipeline framework that extracts human skeletons and captures dynamic temporal changes, integrating spatiotemporal features through ConvLSTM for sequence-level anomaly detection	RGB + Optical flow + Skeleton	Hockey fight, crowd, movies	94.50% 98.50% 94.30%	Pose detection fails in low quality; modality dependency; motion capture under occlusion is fragile.
WSVAD-CFA-HLGAtt (2024) [44]	A dual-branch cross-Modal feature fusion with adaptive weighting for Audio-Video Integration	RGB + Optical Flow + Audio	XD Violence	86%	Sensitive to audio noise; hard to balance modalities; fusion inefficiency.
ConvLSTM2D CBAM (2023) [45]	Involving convolutional block attention Modules Into Conv2DLSTM	RGB	UBI fights	93%	Limited temporal modeling depth; lacks robustness to rapid motion and adversarial perturbations
GMM (2021) [46]	Iterative learning training framework based on Bayesian filtering	RGB	UBI fight, UCF-Crime, and UCSD-Peds1	90%, 75%, 80%	Basic modeling fails with complex motion and has low robustness to adversarial scenarios.
3DSCT-3DCNNs (2024) [47]	A 3D convolutional attention framework with spatiotemporal coupling for enhanced abnormal behaviour detection in videos.	RGB	UCF-Crime and RLVS	80%, 98%	the computational complexity, data dependency, and generalizability across diverse real-world anomaly scenarios.

FIGURE 1 | Sequence of frames in RGB format alongside their TG-RGB$^+$ versions. Images were sourced from the UBI fights dataset [48].

From a signal processing perspective, this integration acts as a feature modulation mechanism, where motion energy adaptively enhances appearance features according to their temporal salience. Consequently, the fused representation emphasizes dynamic regions that are most relevant for violence or anomaly detection, without increasing network depth or parameter count.

Figure 1 illustrates a sequence of video frames in RGB format (top row) alongside their TG-RGB$^+$ versions (bottom row), generated using the proposed method. The original RGB frames were extracted from the UBI Fights dataset [48]. As shown in the figures, the augmentation process includes color jittering and motion blur simulation to enhance model robustness against variations in motion. The red parts highlight key regions of interest where significant motion or interactions occur, which are crucial for violence detection. These processed frames will serve as input for deep learning models to extract spatiotemporal features and improve classification performance.

3.2 | Adversarial Temporal Gradient Adaptive Perturbation

Adversarial training methods are mainly used to enhance the resilience of neural networks against adversarial attacks in video anomaly detection. This approach assumes that the model receives two inputs: the original RGB frames, transformed to TG-RGB$^+$ using temporal gradients, and adversarially perturbed samples. The adversarial perturbations act as structured noise, challenging the model to learn more robust feature representations. By integrating adversarial training, the model is encouraged to make accurate predictions while simultaneously improving its resistance to perturbations, ensuring stability even in the presence of adversarial manipulations.

To further enhance the robustness of the proposed video anomaly detection system, a novel adversarial attack strategy, adversarial temporal gradient adaptive perturbation (T-GAP), is introduced. Unlike traditional adversarial perturbation methods that rely on pixel-wise noise addition, the proposed approach emphasizes temporal dynamics of video sequences, leveraging the temporal gradient and optical flow to generate perturbations that target motion patterns in video data. The main innovation in T-GAP is its adaptive nature, where the perturbations, noted as δ, are not fixed but dynamically adjusted according to the video content X, making them harder to detect and more effective in fooling anomaly detection models as shown in Equation (3).

$$X' = X + \delta \tag{3}$$

To generate adversarial perturbations that align with the motion dynamics of a video, an adaptive perturbation mechanism based on optical flow is introduced. Rather than standard random noise-based perturbations, this method modifies the input based on temporal motion patterns, making adversarial examples X' more effective and harder to detect. By incorporating motion awareness, the perturbation δ adapts to the intensity of movement, ensuring a more realistic and imperceptible modification of video frames. The adversarial perturbation δ is computed as follows:

$$\delta = \alpha \cdot \text{sign}(\nabla_t \text{Flow}(X)) \tag{4}$$

where, $\text{Flow}(X)$ represents the optical flow between consecutive frames X_{t-1} and X_t, capturing pixel displacements over time. ∇_t denotes the temporal gradient of the optical flow, describing how motion changes across frames. Whereas, the $\text{sign}(\cdot)$ function extracts the directional information from the gradient, ensuring that the perturbation aligns with motion patterns. Meanwhile, it α acts as a scaling factor that keeps the perturbation imperceptible while maintaining its effectiveness in misleading the model.

Adversarial perturbations are applied exclusively during training to enhance the model's robustness against attacks. During inference, there is no need to regenerate adversarial examples, as the model has already learned to withstand perturbations. The process of creating adversarial examples is generating perturbations from the input video frames. However, instead of applying perturbations uniformly across the entire video, they are selectively introduced only in the abnormal segments. This targeted approach ensures that the perturbations focus on the critical areas where anomaly detection models are most vulnerable. Through this method, a single adversarial example is generated for each video across the entire training set, effectively improving the model's resistance to adversarial manipulations while preserving the integrity of normal regions.

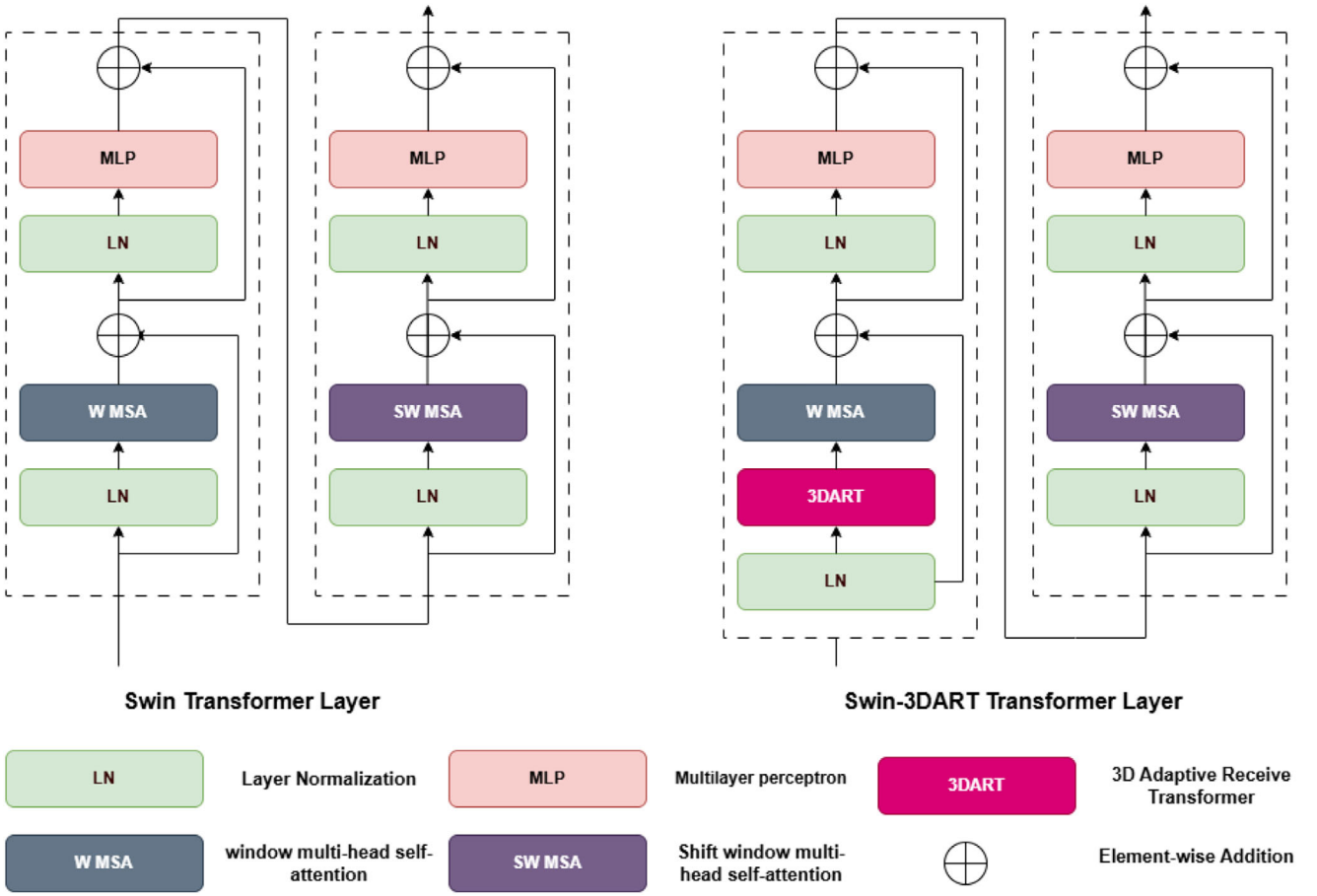


FIGURE 2 | Illustration of the integration of the 3D adaptive receive transformer (3DART) within the swin transformer layer. The 3DART module is positioned between the layer normalization (LN) and the window multi-head self-attention (W-MSA) block, enhancing spatial-temporal feature extraction. This modification adapts the transformer architecture for improved video-based processing.

3.3 | Transformer Network Architecture

In this work, a deep transformer network is adapted as the baseline classification model, anticipating that it will initially be vulnerable to adversarial examples. This choice is motivated by the strong performance of video swin transformers (VSTs), which have achieved state-of-the-art results on major video recognition benchmarks, surpassing traditional CNN-based approaches. The architecture of the model is derived from the original swin transformer, which was originally developed for image processing tasks. VSTs modify the attention mechanism to operate in three dimensions (3D), accounting for both spatial and temporal aspects of video data. This includes a 3D shifted window-based multi-head self-attention mechanism that allows for efficient processing of video frames [49].

To effectively capture motion dynamics, the model processes video clips as a sequence of non-overlapping 3D patches, where each patch consists of multiple consecutive frames. The input video is first divided into 3D patches, each representing a small spatial–temporal region. These patches are then projected into a high-dimensional feature space through a linear embedding layer. The core of the architecture consists of swin transformer blocks, as shown in Figure 2, where alternating layers of multi-head self-attention (MSA) and feedforward networks refine spatial-temporal features. A 3D shifted window mechanism ensures efficient self-attention computation while maintaining contextual continuity across frames. As the network progresses (see Figure 3), feature maps are hierarchically downsampled through patch merging, allowing deeper layers to capture more abstract motion patterns. Finally, the extracted features are processed by a classifier. This design enables the model to learn long-range dependencies while remaining computationally efficient. In the transformer block, a 3D adaptive receive transformer (3DART) module is designed to improve the spatial and temporal representation learning and reduce the dot-product computational complexity of the vanilla self-attention mechanism.

The self-attention mechanism in traditional transformers typically exhibits quadratic complexity concerning the spatial and temporal dimensions of the input. Swin transformers address this limitation by employing localized window-based attention, which reduces computational demands. Despite these improvements, processing high-resolution video data remains challenging due to the large number of tokens generated, leading to significant memory consumption and computational inefficiencies. Overcoming these issues is crucial for making swin transformers more effective and scalable for video applications. To address this challenge, the 3DART module is introduced, designed to enhance the transformer block's capacity to dynamically adjust its receptive field, enabling it to capture essential features across different scales and contexts. This is achieved through the use of

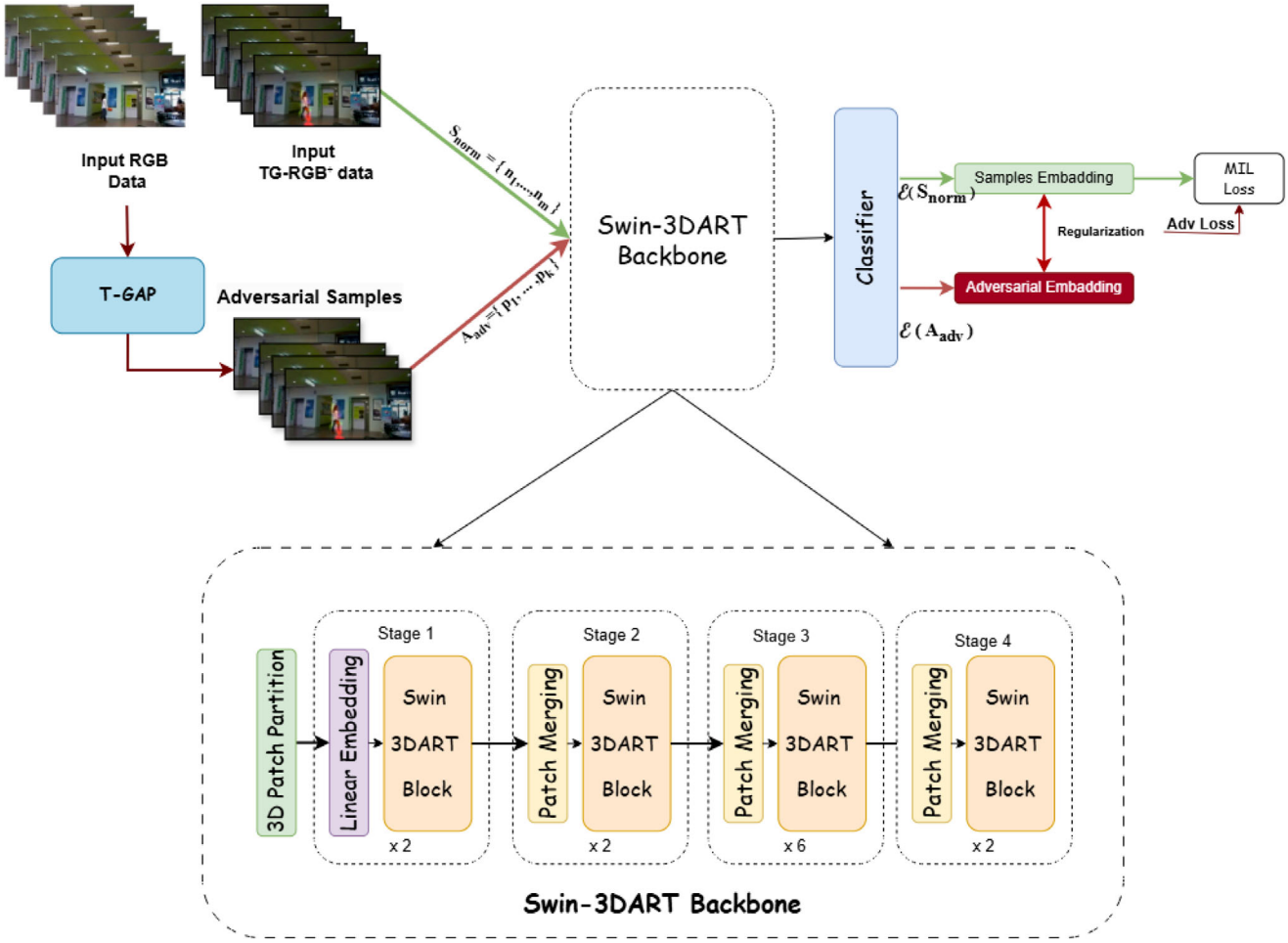


FIGURE 3 | High-level architecture of the proposed deep learning framework utilizing swin-3DART as the backbone feature extractor. The network is trained with adversarial training, taking two inputs: TG-RGB$^+$ frames and adversarial samples generated by T-GAP. It optimizes two loss functions (Equation 10), the multiple instance learning (MIL) loss [7] and the adversarial loss function, to enhance robustness and improve feature learning for violence detection.

an adaptive pooling technique applied to the depth and spatial dimensions, which not only reduces computational overhead but also incorporates multi-scale contextual information.

3.3.1 | 3D Adaptive Receive Transformer

As illustrated in Figure 4, the input token $X \in \mathbb{R}^{h \times w \times c \times d}$ is reshaped into $X_1 \in \mathbb{R}^{h' \times w' \times c \times d'}$, where h, w, and d represent the height, width, and depth of X, respectively. The resulting dimensions d', h', and w' are obtained by applying a reduction factor $r = \frac{3}{4}$, ensuring a controlled decrease in size that balances the trade-off between efficiency and information retention. Specifically,

$$d' = r \cdot d, \quad h' = r \cdot h, \quad w' = r \cdot w.$$

Meanwhile, adaptive pooling operations are applied to reduce both the depth (ADP) and spatial (ASP) dimensions of the key (K) and value (V) inputs before the self-attention mechanism. Specifically, the input X is passed through two parallel pooling branches, as described in Equations (6) and (5), to generate X_2 and X_3, where $X_2, X_3 \in \mathbb{R}^{h' \times w' \times c \times d'}$.

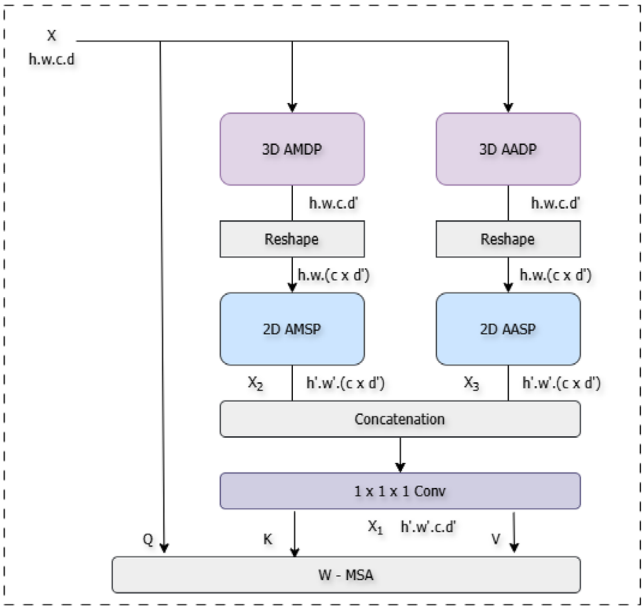


FIGURE 4 | 3DART module architecture, comprising adaptive max/avg depth pooling (A(M/A)DP) and spatial pooling (A(M/A)SP).

- **Max-pooling branch**: Adaptive max-depth pooling (AMDP) is applied to emphasize the strongest temporal signals, yielding an intermediate tensor of shape

$$(h, w, c, d')$$

where c is the number of channels, d' is the reduced depth, and h, w are the spatial dimensions. For spatial reduction, the tensor is reshaped to

$$(h, w, (c \times d'))$$

enable adaptive max-spatial pooling (AMSP) to focus on the most salient spatial features, producing the branch output $X2$ with a tensor shape of

$$(h', w', (c \times d'))$$

the correspondence equation:

$$X_2(X) = \text{AMSP}(\text{AMDP}(X)) \tag{5}$$

- **Average-pooling branch**: Adaptive average-depth pooling (AADP) is applied to summarize the temporal dynamics by averaging, allowing the model to capture consistent, non-extreme patterns across the video sequence. yielding an intermediate tensor of shape

$$(h, w, c, d')$$

where c is the number of channels, d' is the reduced depth, and h, w are the spatial dimensions. For spatial reduction, the tensor is reshaped to

$$(h, w, (c \times d'))$$

enable adaptive average-spatial pooling (AASP) to compress the spatial dimensions effectively while retaining average regional patterns important for holistic scene understanding, producing the branch output X_3 with a tensor shape of

$$(h', w', (c \times d'))$$

the correspondence equation:

$$X_3(X) = \text{AASP}(\text{AADP}(X)) \tag{6}$$

After the pooling operations, the outputs X_2, and X_3 are reshaped back to

$$(h', w', c, d')$$

Finally, they are concatenated along the channel dimension, resulting in a combined feature map of size $\mathbb{R}^{h' \times w' \times 2c \times d'}$. To project the channels back to the original size, a $1 \times 1 \times 1$ convolutional layer is applied, as shown in Equation (7):

$$X_1(X) = \text{Conv}(1 \times 1 \times 1)(X_2 \oplus X_3) \tag{7}$$

By combining both branches, the model benefits from a dual perspective: one that preserves consistent global patterns and another that amplifies the strongest, most discriminative cues. This complementary fusion enhances the representational power of the pooled features, enabling the subsequent attention module to effectively balance general context with critical details for robust anomaly detection. 3DART can effectively improve the feature representation without additional parameters and can be directly inserted into the existing backbones. Note that the pooling (size or output size) layers in both branches can be adjusted according to the size of the input data.

3.3.2 | Loss Function

The proposed loss function integrates both supervised learning and adversarial regularization to enhance robustness in video anomaly detection. Given the anomaly score representation $\mathcal{R} = \mathcal{F}(\mathcal{S}, \mathcal{A})$, normal training samples are denoted as $\mathcal{S}_{\text{norm}} = \{n_1, \ldots, n_m\}$ and adversarially generated perturbation samples as $\mathcal{A}_{\text{adv}} = \{p_1, \ldots, p_k\}$, both of which serve as input to the model $\mathcal{F}$. The swin-3DART-based architecture processes these inputs, producing embeddings $\mathcal{E}(\mathcal{S}_{\text{norm}})$ and $\mathcal{E}(\mathcal{A}_{\text{adv}})$, where $\mathcal{E}(\mathcal{A}_{\text{adv}})$ is introduced only during training. To ensure effective anomaly separation, a weakly supervised loss is employed to combine multiple instance learning (MIL) ranking loss [7] with smoothness and sparsity constraints as illustrated in equation 8.

$$L_{\text{MIL}} = l(\mathcal{S}'_{\text{abnom}}, \mathcal{S}'_{\text{norm}}) + \lambda_1 s + \lambda_2 p + \|\mathcal{W}\|_F \tag{8}$$

where $l(\mathcal{S}'_{\text{abnom}}, \mathcal{S}'_{\text{norm}})$ enforces ranking constraints to distinguish anomaly scores of abnormal and normal snippets. s and p ensure temporal smoothness and sparsity, respectively, with λ_1 and λ_2 serving as weighting factors to balance their contributions in the loss function. The Frobenius norm $\|\mathcal{W}\|_F$ helps to prevent overfitting [7].

To further enhance robustness, an adversarial loss L_{adv} is incorporated based on cosine similarity, following the approach proposed in [50], to measure the relationship between normal and adversarial feature embeddings as illustrated in Equation (9). This adversarial constraint ensures that the model learns to discriminate anomalies even in the presence of perturbations.

$$L_{\text{adv}} = 1 - \cos(\mathcal{E}(\mathcal{S}_{\text{norm}}), \mathcal{E}(\mathcal{A}_{\text{adv}})) \tag{9}$$

To balance anomaly detection accuracy and adversarial robustness, the ultimate equation of the loss function is given as follows:

$$\text{Loss} = L_{\text{MIL}} + \alpha L_{\text{adv}} \tag{10}$$

where α is a scaling factor controlling the influence of adversarial regularization. This formulation enhances anomaly detection by leveraging structured data learning while mitigating adversarial vulnerabilities.

4 | Experimental Configuration and Performance Analysis

4.1 | Experimental Configuration

4.1.1 | Datasets

The proposed detection model is trained and evaluated using datasets preprocessed through the TG-RGB$^+$ representation

TABLE 3 | Description of the datasets used in the experiments.

Dataset	Types of anomalous actions included	Number of Classes	Number of Videos
UCF-Crime [7]	Robbery, assault, burglary, explosion, fighting, abuse, vandalism, road accident, shooting etc.	13 (12 anomalous + 1 normal)	1,900 videos (810 anomalous, 1,090 normal)
UBI Fights [48]	Physical fights between individuals or groups, pushing, and punching and kicking sequences.	2 (violence / non-violence)	1,000 videos (300 violent, 700 non-violent)
RLVS [51]	Physical aggression, group fights, and violent crowd behaviours in real-world scenes.	2 (violent/non-violent)	2,000 videos (1,000 violent, 1,000 non-violent)

rather than being directly applied in the initial set of experiments; according to the predetermined parameters for each dataset, each dataset has a distinct form of labelling. Two categories are used to classify these labelling methods. First, frame-level labelling, in which every frame is assigned a label based on whether or not it contains violent behaviour; this is the UBI Fights [48] case. Second, video-level labelling, where an entire video is assigned a single label that categorizes its overall content. This applies to the real life violence situations (RLVS) dataset [51]. Similarly, the UCF-Crime dataset [7] follows video-level labelling in its training set, while its testing set is labelled using frame-specific intervals, indicating the exact segments where violent activity occurs. This ensures that the TG-RGB$^+$ transformation is applied before feeding the data into the model for training and evaluation.

This diversity in annotation schemes allows evaluation of the TG-RGB$^+$ model under both fine-grained (temporal) and holistic (clip-based) supervision, highlighting its adaptability across heterogeneous datasets. UCF-Crime comprises 1900 real-world surveillance videos spanning 13 classes (12 anomalous and 1 normal), including robbery, assault, and vandalism. UBI Fights contains 1000 videos (300 violent, 700 non-violent) featuring interpersonal physical confrontations. RLVS includes 2000 videos (1000 violent, 1000 non-violent) depicting real-world violent scenarios. A detailed overview of these datasets is presented in Table 3.

4.1.2 | Implementation Details

All experiments were conducted on the Kaggle platform[1] using a TPU configuration with 330 GB of memory. The proposed swin-3DART model is trained using both clean TG-RGB$^+$ data and perturbed data with the T-GAP attack, enabling the model to learn robustness against adversarial interference. Training is performed consistently across all datasets, with hyperparameters and settings kept the same for fair evaluation. Testing is carried out under three conditions: clean TG-RGB$^+$ input, input data with T-GAP perturbations, and simulated low-light conditions. This setup allows for a thorough analysis of standard detection accuracy, adversarial robustness, and resilience to environmental variations. Additionally, ablation studies are performed to quantify the contribution of each model component and input modality, while cross-dataset experiments evaluate generalization. Efficiency versus accuracy trade-offs are also assessed to ensure that the model is suitable for real-time deployment on resource-constrained or edge devices. Furthermore, to ensure the statistical robustness and generalizability of the findings, a k-fold cross-validation framework ($k = 5$) was applied to the ablation configurations. Each model was trained for 7000 epochs per fold. While the training partitions were subjected to stochastic shuffling to enhance feature learning, the testing sets remained fixed to ensure a strictly unbiased evaluation of the model's performance.

4.2 | Evaluation Metrics

To evaluate the performance of swin-3DART, the receiver operating characteristic area under the curve (ROC AUC) was employed, which measures the trade-off between sensitivity and specificity across different classification thresholds. This metric assesses the model's ability to distinguish between anomaly and normal classes. Additionally, the equal error rate was utilized to quantify the proportion of misclassified instances. The EER is computed as the point where the false acceptance rate (FAR) equals the false rejection rate (FRR), providing an overall measure of the model's predictive errors.

The original accuracy (OAcc) metric is used to measure the overall correctness of the proposed model's predictions on clean (unaltered) test data (see Equation 11).

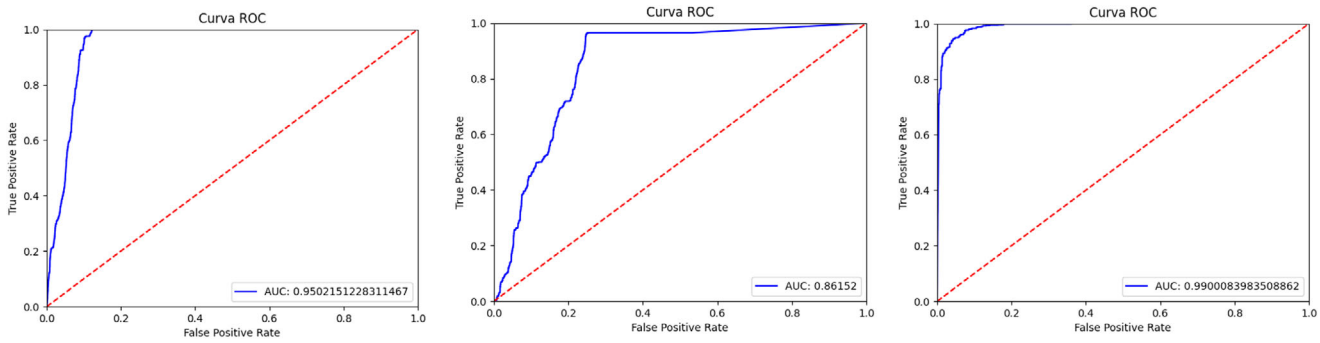
$$\text{OAcc} = \frac{\text{Correct predictions on clean data}}{\text{Total test samples}} \tag{11}$$

The attack success rate (ASR) is used to quantify how often an adversarial attack successfully fools the model by misclassifying previously correct samples (see Equation 12).

$$\text{ASR} = \frac{\text{Number of misclassified samples due to attack}}{\text{Total correctly classified clean samples}} \tag{12}$$

The accuracy under attack (AUA) metric is used to measure how well the proposed model performs on adversarially perturbed samples (see Equation 13).

$$\text{AccUA} = \frac{\text{Correct predictions on adversarial data}}{\text{Total test samples}} \tag{13}$$



(a) AUC UBI Fight (b) AUC UCF-Crime (c) AUC RLVS

FIGURE 5 | AUC performance of swin-3DART: (a) UBI Fight, (b) UCF-Crime and (c) RLVS datasets. Each plot shows the ROC curve highlighting the anomaly classification effectiveness.

In addition, a statistical validation was conducted using a two-stage pipeline by first applying the Shapiro–Wilk test [52] to evaluate the null hypothesis (H_0) that results are normally distributed relative to $\alpha = 0.05$. Depending on this outcome, either a paired t-test [53] or a Wilcoxon test [54] is selected, providing a rigorous, distribution-agnostic comparison between architectures.

4.3 | In-Domain Performance Evaluation

To evaluate the effectiveness and robustness of the proposed detection model, swin-3DART, in anomaly detection, three in-domain experimental settings were designed. Across all settings, the model is trained with the same configuration using TG-RGB$^+$ data and TG-RGB$^+$ perturbed data (generated using T-GAP). The evaluation is conducted on the three benchmark datasets. The three experimental settings are designed to provide a comprehensive analysis by assessing standard performance, adversarial robustness, and performance under low-light conditions. This multi-faceted evaluation ensures a rigorous analysis of swin-3DART's generalization and resilience.

4.3.1 | Evaluation With Clean TG-RGB$^+$ Data

In the first set of experiments, the evaluation is performed on clean TG-RGB$^+$ data to assess the model's performance under normal conditions, ensuring its ability to accurately detect anomalies in real-world scenarios without adversarial interference and emphasizing the impact of the TG-RGB$^+$ modality in achieving superior detection accuracy. As shown in Figure 5, swin-3DART demonstrated remarkable performance across the three datasets, achieving ROC AUC scores of 95.02%, 86.15% and 99.00% on the UBI Fight, UCF-Crime and RLVS datasets, respectively. These results highlight the model's strong ability to distinguish between normal and anomalous instances, with particularly high accuracy on the RLVS dataset, indicating its robustness in detecting anomalies. The UBI Fight dataset also exhibits a high AUC, suggesting that swin-3DART effectively captures complex patterns in violent activity detection. Although the UCF-Crime dataset has a comparatively lower AUC, the model still maintains strong discriminative capability, proving its effectiveness across diverse scenarios. It is worth noting that in Figure 5b, the ROC curve for UCF-Crime exhibits a long flat region. This behaviour arises due to the dataset's highly imbalanced nature, where long stretches of normal video without anomalies lead to a plateau in the true positive rate despite increases in the false positive rate. This characteristic highlights the difficulty of detecting sparse anomalies over time and reinforces the importance of the temporal modeling provided by the TG-RGB$^+$ design, which helps the model remain sensitive even in such challenging conditions.

Table 4 provides a comparative analysis of swin-3DART against state-of-the-art methods, showing it consistently outperforms them across multiple evaluation metrics. On UBI Fights, earlier models like BayesianNet [46] (2021) and GMM [48] (2020) rely solely on RGB, reaching ROC AUC scores of 0.84 and 0.90, respectively, but missing temporal dynamics. ViT (Vision transformer, 2023), using only optical flow, achieves 0.89 ROC AUC but lacks the spatial richness of RGB. DeVTrV2 [55] (2021) and ConvLSTM2D+CBAM [45] (2023) improve performance by mixing convolution and attention (up to 0.93 ROC AUC). However, these approaches introduce additional computational overhead due to stacked CNN and attention layers and largely focus on local spatial patterns captured by CNNs, which limits their capacity to model long-range temporal dependencies. The swin-3DART surpasses them all, achieving 0.95 ROC AUC and 0.07 EER, due to the integration of TG-RGB$^+$, which combines temporal gradient cues with RGB, enriching spatiotemporal learning beyond RGB or optical flow alone.

On UCF-Crime, methods like deep MIL ranking [7] (2018), GMM [48] (2020), and CNN-TCN [57] (2023) stay below 0.85 ROC AUC, largely due to unimodal designs. Recent works such as RTFM (swin transformers) [56] (2021), contrastive attention [59] (2021), and TEVAD [41] (2023) slightly improve results by leveraging attention or better temporal modeling. Even the most recent state-of-the-art methods (2024–2025) encounter performance plateaus: I3D-AD-AFF-MFL [42] achieves 0.850 ROC AUC by stacking multiple heavy modules (I3D backbone, AD, AFF, and MFL), while AFM-WSVAD [58] reaches 0.854 ROC AUC by incorporating auxiliary text-based features. In contrast, swin-3DART attains 0.861 ROC AUC with a simpler, unified architecture, demonstrating that higher complexity is not always necessary for improved performance. This demonstrates that the TG-RGB$^+$ modality is more effective than standard RGB, optical flow, or

TABLE 4 | Comparison with state-of-the-art methods.

Dataset	Category	Models	Year	Modality	ROC AUC	EER
UBI Fight	General models	BayesianNet [46]	2021	RGB	0.840	0.210
		DeVTrV2 CNN + transformer [55]	2021	RGB	0.918	—
		GMM [48]	2020	RGB	0.900	0.160
		ViT (vision transformer)	2023	Optical flow	0.890	—
		ConvLSTM2D+CBAM [45]	2023	RGB	0.930	0.060
		Swin-3DART (ours)	2025	TG-RGB$^+$	**0.950**	**0.070**
UCF-Crime	General models	Deep MIL ranking [7]	2018	RGB	0.750	0.300
		GMM [48]	2020	RGB	0.750	0.302
		RTFM (swin transformers) [56]	2021	RGB	0.843	—
		CNN-TCN [57]	2023	RGB	0.840	—
		TEVAD [41]	2023	RGB	0.849	—
		I3D-AD-AFF-MFL [42]	2024	RGB+optical Flow	0.850	—
		AFM-WSVAD [58]	2025	RGB + text	0.854	—
	Lightweight models	Contrastive attention [59]	2021	RGB	0.846	—
		Watanabe [60]	2022	RGB	0.847	—
		Be-WVAD [61]	2024	RGB	0.841	—
		Light-WVAD [62]	2025	RGB	0.847	—
		Swin-3DART (ours)	2025	TG-RGB$^+$	0.861	0.210
RLVS	General models	Deep MIL ranking [7] *	2018	RGB	0.940	0.130
		3DSCT-I3D [47]	2024	RGB	0.875	0.201
		3DSCT-ResNet152 [47]	2024	RGB	0.986	0.055
		Swin-3DART (ours)	2025	TG-RGB$^+$	**0.990**	**0.050**

TABLE 5 | Feature refinement modules on UCF-Crime dataset: evaluation against other methods.

Modules	AD [42]	SGA [63]	STGA [64]	3DART
ROC AUC (%)	81.69	80.28	81.02	**86.15**

text-augmented pipelines, offering a clean and scalable solution that surpasses current benchmarks.

When comparing to lightweight models designed for efficiency, including contrastive attention [59] (2021), Watanabe [60] (2022), Be-WVAD [61] (2024) and Light-WVAD [62] (2025), performance consistently peaks around 0.847 ROC AUC. While these methods prioritize efficiency, they often sacrifice the complex spatiotemporal reasoning necessary for hard-to-detect anomalies. On RLVS, prior methods such as 3DSCT-I3D and 3DSCT-ResNet152 [47] (2024) achieve high performance (up to 0.986 ROC AUC) by using heavy 3D backbones with additional modules. Swin-3DART goes even further, hitting 0.99 ROC AUC and 0.05 EER, using its parameter-free adaptive pooling within the swin transformer, demonstrating its strong generalization across diverse video anomaly detection tasks.

Further, Table 5 provides a comparative analysis of swin-3DART against other feature refinement modules on the UCF-Crime dataset. The 3DART module establishes a new benchmark by achieving the highest ROC AUC (86.15%), outperforming modules such as self-guided attention (SGA) (80.28%), spatial–temporal graph attention network (STGA) (81.02%), and the attention de-redundancy module (AD) (81.69%). This represents a relative gain of 6.06% over the next best method (AD) and an absolute improvement of 5.13%, 5.26% and 4.95% compared to SGA, STGA and BSP-style approaches, respectively. These margins underscore swin-3DART's feature refinement in distinguishing subtle anomalies from patterns.

While previous methods target redundancy through external supervision or fixed-structure modules, swin-3DART fills a critical gap by integrating adaptive pooling directly into a transformer backbone: unlike SGA, which depends on pseudo-label guidance that can be sensitive to labelling noise; STGA, which prunes redundant links via spatial and temporal graphs but remains constrained by local graph connectivity; the AD module, which applies specialized attention blocks yet still processes features at a uniform granularity; or BSP-style distillation, which introduces dependency on an external teacher network.

By contrast, swin-3DART's adaptive pooling mechanism dynamically resizes and refines its attention regions, focusing computation on the most informative spatial and temporal patches while reducing dot-product complexity. Coupled with the transformer's global receptive field, this design captures both fine-grained local

TABLE 6 | Comparison of performance and safety metrics for the swin-3DART model, across attack T-GAP method, and target dataset sequences. OAcc: original accuracy; ASR: attack success rate; AUA: accuracy under attack.

Datasets	OAcc (%)	ASR (%)	AccUA (%)
UBI Fights	92.47	4.5	91.91
UCF-Crime	77.85	7.5	75.48
RLVS	95.11	4.0	94.49

details and long-range dependencies, yielding a more discriminative feature space without additional supervision or heavy graph operations. The substantial uplift in ROC AUC (over 5% across all baselines) confirms that the 3DART parameter-free yet powerful refinement module closes the gap between localized redundancy reduction and global context modeling, establishing a new state of the art in video anomaly detection.

4.3.2 | Evaluation With T-GAP Perturbation

These experiments aim to evaluate the impact of T-GAP perturbation on swin-3DART's robustness. The model is trained using the same configuration with TG-RGB$^+$ data and TG-RGB$^+$ perturbed data (generated using T-GAP). Testing is conducted under two conditions: (1) on clean TG-RGB$^+$ data to measure standard detection performance and (2) on TG-RGB$^+$ perturbed data to evaluate resilience against adversarial interference. This evaluation allows us to analyse the trade-off between accuracy and robustness, as incorporating adversarial perturbations may improve resilience but could also affect standard detection performance. This ablation study examines the ability of swin-3DART to adapt to perturbations while preserving strong anomaly detection performance

The evaluation is conducted under the $\ell_\infty(\epsilon \approx 13/255)$ threat model, which represents a high perturbation level, as typical low and average perturbation values are around 2/255 and 8/255, respectively. Such a high perturbation significantly alters input features, making it challenging for deep models to retain their predictive performance. The adversarial perturbation δ is constrained by the ℓ_∞ threat model, ensuring that each pixel change remains within it ϵ, making the attack imperceptible yet effective in misleading the anomaly detection model. This level of adversarial manipulation closely mimics real-world attack scenarios, where malicious entities attempt to deceive models in critical applications such as video surveillance and security monitoring.

The assessment is conducted on the UBI Fights, UCF-Crime, and RLVS datasets, demonstrating the model's resilience against adversarial perturbations. As shown in Table 6, swin-3DART achieves an original accuracy (OAcc) of 92.47% on UBI Fights, with an attack success rate (ASR) of 4.5%, resulting in an accuracy under attack (AccUA) of 91.91%. On the UCF-Crime dataset, the model attains an OAcc of 77.85%, an ASR of 7.5%, and an AccUA of 75%. The lower accuracy on UCF-Crime can be attributed to its high intra-class variability and severe class imbalance, where normal frames dominate, making anomaly detection more challenging. Despite this, the model maintains strong discriminative capability, with only a 2.85% drop in accuracy under attack. Similarly, on the RLVS dataset, the model achieves an OAcc of 95%, an ASR of 4% and an AccUA of 94.49%. These results confirm swin-3DART's ability to withstand strong adversarial perturbations, particularly against unseen attacks.

To further validate generalization, the resilience of the swin-3DART model was tested against unseen, classic adversarial attacks that do not utilize the temporal gradient structure: PGD (projected gradient descent)[65] and FGSM (fast gradient sign method)[66]. The results in Table 7 demonstrate that the swin-3DART framework successfully maintains a low Attack success rate (ASR) and high accuracy under attack (AccUA) even against PGD and FGSM. The performance against T-GAP is marginally superior in most cases, which is expected since the model was trained against it. Crucially, the consistently low ASR values against PGD and FGSM validate that T-GAP training yields a generalizable robustness against various forms of adversarial interference.

Furthermore, resisting high-perturbation adversarial attacks is crucial in safety-critical applications. For instance, in crime detection, adversarial perturbations could manipulate video frames to mislead anomaly detection systems, potentially allowing unauthorized activities to go undetected [67]. Likewise, in autonomous systems, adversarial perturbations could disrupt decision-making processes, leading to severe consequences [68]. By successfully reducing the impact of such attacks, swin-3DART strengthens security and improves decision-making. This makes it a dependable solution for domains vulnerable to adversarial threats. Overall, these findings establish the proposed detection model as a new benchmark for adversarial robustness, achieving an optimal balance between accuracy and resilience under strong perturbations in video analysis tasks.

4.3.3 | Robustness to Environmental Illumination

To further assess deployment feasibility in real-world surveillance scenarios, model performance under low-light conditions was evaluated. To verify the robustness of the core TG-RGB$^+$ modality, specialized tests were conducted by simulating typical suboptimal illumination. Specifically, a mid-level low-light transformation was applied to the raw RGB frames of the test sets prior to the computation of the temporal gradient (TG) features. This ensures that the TG component is derived from noisy, degraded input, reflecting a realistic scenario where motion cues must be extracted from poor-quality video. The results, detailed in Table 8, quantify the impact across three diverse datasets:

The observed performance drops are minimal, confirming the model's strong resilience to environmental noise. This resilience validates the hypothesis that the temporal gradient (TG) features—even when derived from low-quality RGB input—are inherently less susceptible to pixel-level noise and illumination fluctuations than appearance features alone. By providing stable motion cues, the TG-RGB$^+$ fusion mechanism ensures the swin-3DART framework is a reliable solution for deployment in surveillance settings where lighting quality is frequently suboptimal.

TABLE 7 | Comparison of robustness under adversarial attacks across datasets.

Attack method	UBI Fight		UCF-Crime		RLVS	
	AccUA (%)	ASR (%)	AccUA (%)	ASR (%)	AccUA (%)	ASR (%)
PGD	91.6	5.0	71.99	8.2	94.0	4.8
FGSM	91.5	5.0	71.79	8.3	93.92	4.9
T-GAP (ours)	**91.91**	**4.5**	**75.48**	**7.5**	**94.49**	**4.0**

TABLE 8 | Performance degradation of Swin-3DART (TG-RGB$^+$) under simulated low-light conditions (AUC %).

Dataset	AUC (normal light)	AUC (low-light)	Degradation (%)
UBI Fight	95.00	91.00	−4.00
UCF-Crime	86.15	85.55	−0.60
RLVS	99.00	97.00	−2.00

TABLE 9 | Sensitivity analysis of the scaling factor α (from Equation 10) across UCF-Crime, UBI-Fights and RLVS datasets.

Dataset	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 1.0$
UBI-Fights	94.36	94.55	**95.04**	94.87	94.76
UCF-Crime	84.58	84.83	**86.16**	84.86	84.62
RLVS	98.90	98.99	**99.00**	98.99	98.70

4.4 | Component and Fusion Analysis

To thoroughly examine the behaviour and effectiveness of the proposed swin-3DART framework, this section investigates how individual components and fusion strategies contribute to the overall performance. The evaluation encompasses multiple configurations, moving incrementally from a simple baseline model (C1) to the final, complete model (C5). These settings are designed to assess three main objectives: (1) the sensitivity of the model to primary loss function hyperparameters, (2) the contribution of each architectural module to anomaly detection accuracy, and (3) the efficiency of the fusion mechanism in capturing complementary temporal–spatial cues across modalities.

4.4.1 | Hyperparameter Sensitivity Analysis

This subsection evaluates the stability and robustness of the proposed framework by analysing the impact of the weighting factors defined in the loss formulation. While λ_1 and λ_2 in Equation (8) are maintained at the standard values established in the literature (8×10^{-5}) to ensure a fair comparison with baseline MIL methods, the scaling factor α in Equation (10) is a specific parameter introduced to balance the MIL-based ranking loss (L_{MIL}) with the adversarial regularization (L_{adv}).

To verify that model performance is not the result of sensitive hyperparameter tuning, α was varied within the range $[0.1, 1.0]$. This parameter controls the influence of the cosine-similarity-based adversarial loss, designed to detect anomalies in the presence of perturbations. The resulting performance trends across three major benchmarks are illustrated in Table 9. Experimental results indicate that the framework is remarkably robust across all three datasets. Optimal performance for every benchmark is achieved at $\alpha = 0.5$, where the AUC reaches 86.16% for UCF-Crime, 95.05% for UBI-Fights, and 99.00% for RLVS.

The convergence of the optimal value at 0.5 across disparate datasets suggests that the integration of L_{adv} provides a universal benefit to the feature embedding space. In the case of UCF-Crime, which contains complex real-world anomalies, a significant improvement is observed as α approaches the midpoint, confirming that adversarial regularization effectively mitigates vulnerabilities in high-variance environments. For UBI-Fights and RLVS, the performance remains steady with only negligible fluctuations, indicating that the model is not overly dependent on precise parameter tuning. The marginal performance drop observed when α exceeds 0.5 suggests that while adversarial regularization is crucial, excessive weight can slightly diminish the focus on the primary ranking constraints. Consequently, $\alpha = 0.5$ was selected as the default setting for all experiments to provide an ideal trade-off between standard ranking accuracy and adversarial robustness.

4.4.2 | Ablation Study

To systematically quantify the individual contribution and marginal utility of each novel component within the proposed swin-3DART framework, an ablation study was conducted focusing on three primary elements: the TG-RGB$^+$ modality, the 3DART module, and the T-GAP attack. This study is essential to provide empirical evidence that the substantial performance gains demonstrated in Section 4.3 are not attributable to a single dominant factor but rather to the synergistic effect of the integrated modules. Five configurations were defined, where the first three establish the foundational impact of the core architectural choices:

- **C1 baseline (RGB + 3D swin-B)**: This configuration serves as the true baseline, utilizing standard RGB input with the off-the-shelf 3D swin-transformer backbone. It establishes the minimum performance achievable using contemporary vision transformer architectures for comparison.
- **C2: input validation (TG − RGB$^+$ + 3D swin-B)**: By replacing the RGB input of C1 with the TG-RGB$^+$ modality while keeping the baseline backbone, the marginal gain provided

TABLE 10 | Ablation study of the Swin-3DART components across three benchmarks. Results are reported as mean AUC (%) ± standard deviation across k-fold cross-validation.

Configuration	UBI Fights	UCF-Crime	RLVS
C1: RGB + Swin-B	82.70 ± 0.30	73.07 ± 0.23	95.62 ± 0.15
C2: TG-RGB$^+$ + Swin-B	91.97 ± 0.03	79.16 ± 0.21	98.28 ± 0.10
C3: RGB + Swin-3DART	92.48 ± 0.53	80.89 ± 0.68	98.64 ± 0.06
C4: TG-RGB$^+$ + Swin-3DART	94.96 ± 0.35	84.58 ± 0.86	98.90 ± 0.02
C5: **TG − RGB$^+$ + Swin-3DART + T-GAP**	**94.99 ± 0.24**	**85.82 ± 0.20**	**98.94 ± 0.07**

solely by the robust, motion-aware input representation is isolated and quantified.

- **C3: architectural validation (RGB + swin − 3DART)**: By replacing the standard swin-B backbone of C1 with the swin-3DART module while retaining the standard RGB input, the intrinsic feature refinement capabilities and efficiency gains of the novel parameter-free architectural enhancement are isolated.
- **C4: synergistic validation (TG − RGB$^+$ + swin − 3DART)**: This configuration combines the TG-RGB$^+$ input (validated in C2) with the swin-3DART backbone (validated in C3). This stage is crucial for assessing the synergy between the input and architectural innovations, showcasing maximal non-adversarial performance.
- **C5: Full model validation (TG − RGB$^+$ + swin-3DART + T-GAP)**: This final configuration incorporates the T-GAP adversarial strategy, exposing C4 to challenging perturbations during training. It quantifies the final, critical performance gains achieved through enhanced adversarial robustness, resulting in the complete, state-of-the-art swin-3DART framework.

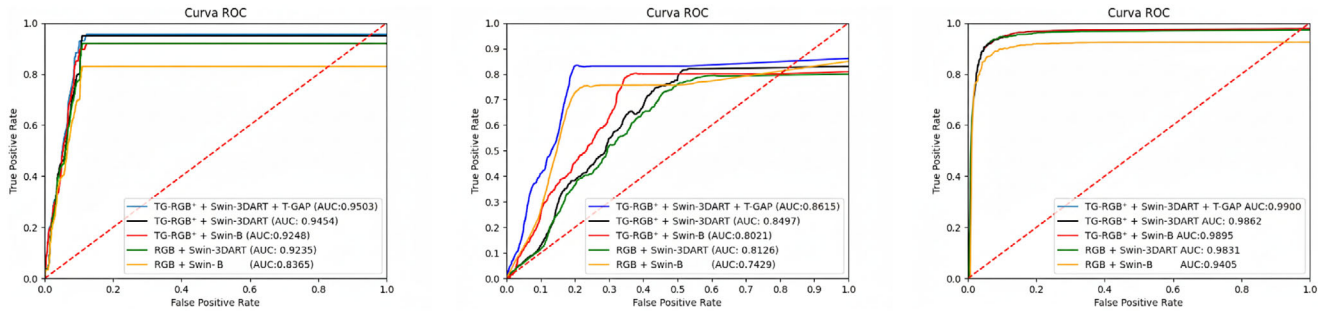
To isolate the contribution of each module, a cumulative configuration approach was employed, starting with C1 as the benchmark. The proposed components were sequentially added, and performance was evaluated across all three benchmark datasets: UBI Fights, UCF-Crime and RLVS. All results are measured using the area under the ROC curve (AUC%). To ensure the robustness of these findings, the statistical significance of the ablation results is verified according to the protocol in Section 4.2. This procedure aims to confirm that the observed performance gains are statistically sound ($\alpha = 0.05$) via distribution-appropriate hypothesis testing. The comparative results across the five configurations are detailed in Table 10.

The numerical results in Table 10 further confirm the necessity and effectiveness of each module, highlighting both individual and combined contributions. The main findings are summarized as follows:

1. **Impact of the TG − RGB$^+$ modality (C1 → C2)**: The transition from a standard RGB stream to the proposed temporal-gradient fusion (TG-RGB$^+$) yielded the most substantial single performance gain. On the UBI Fights dataset, a leap of +9.27% in AUC was observed. This improvement is statistically validated by a P-value < 0.001. Furthermore, the marked reduction in standard deviation (from 0.30 to 0.03) indicates that priming the network with motion-aware features effectively regularizes the learning process and enhances robustness to environmental noise and lighting changes.
2. **Efficacy of the 3DART module (C1 vs. C3)**: The introduction of the 3D adaptive receptive field transformer (3DART) backbone, even in the absence of motion-gradient fusion, resulted in a remarkable AUC increase (e.g. +9% on UBI Fights and +7% on UCF-Crime). This validates the intrinsic value of the proposed architectural enhancement. By enabling the network to adaptively calibrate its receptive field view based on spatiotemporal complexity, 3DART successfully captures the long-range dependencies that standard swin-B architectures fail to model. The near-perfect performance on the RLVS dataset (98.28%) further underscores the precision of this mechanism.
3. **Synergistic effect of configuration C4**: Configuration C4 represents the integration of both the TG-RGB$^+$ input representation and architectural (swin-3DART) innovations. This configuration achieves near-peak performance across all benchmarks (e.g. 84.58% on UCF-Crime). The statistical analysis suggests a strong synergistic effect; the 3DART module appears most effective when refining the high-entropy, motion-aware feature maps supplied by the TG-RGB$^+$ modality.
4. **Adversarial robustness and T-GAP (C5)**: In the final configuration (C5), the model is subjected to adversarial training using the adversarial temporal gradient adaptive perturbation (T-GAP). This stage provides a decisive performance boost, yielding state-of-the-art results: 94.99% on UBI Fights and 85.82% on UCF-Crime. The improvement from C4 to C5 is statistically significant ($P < 0.01$), confirming that exposing the model to synthetic motion perturbations forces the attention mechanisms to prioritize only the most discriminative anomaly signatures. This adversarial hardening is critical for ensuring performance stability in real-world, deployment-ready scenarios.

To visually confirm the distinct performance advantage delivered by the core architectural innovation, receiver operating characteristic (ROC) curves are presented for each dataset, comparing the performance of the full swin-3DART architecture (C5) against key baselines. While the preceding quantitative analyses report aggregate metrics derived from k-fold cross-validation, the visualizations in Figure 6a–c are generated using the conventionally established data partitions to provide a high-fidelity snapshot of the discriminative trajectory under standardized train-test splits.

The plots in Figure 6–c reveal a consistent and significant pattern of separation, demonstrating that each module contributes a unique, non-overlapping improvement to the model's feature discrimination capabilities. The curves do not converge or move



(a) AUC UBI Fight (b) AUC UCF-Crime (c) AUC RLVS

FIGURE 6 | Ablation study ROC curves of swin-3DART components: (a) UBI Fights, (b) UCF-Crime and (c) RLVS. Each curve shows the impact of progressively adding components on anomaly detection performance.

together because each configuration is learning fundamentally different and increasingly specialized features:

1. Input separation (TG-RGB$^+$): The initial separation of the TG-RGB$^+$ curve C2 (red) from the RGB baseline C1 (orange) is the first major lift. This proves that the TG-RGB$^+$ modality does not merely upscale the RGB output; it introduces new, motion-centric features that are highly discriminative for anomaly detection.
2. Architectural separation (3DART): The lift from the C2 (red) baseline to the C4 (black) architecture validates the adaptive mechanism. The 3DART module allows the model to dynamically select the most relevant spatiotemporal features, discarding noise and focusing classification confidence on the anomaly's true signal.
3. Robustness separation (T-GAP): The final decisive separation (C4 → C5) (black → blue) confirms that the training using T-GAP hardens the entire feature set. The model's learned features become invariant to subtle motion-based adversarial noise, compelling it to rely only on the most robust and generalized signals.

This graphical evidence conclusively demonstrates the synergistic effect of the integrated modules, which, when combined, result in a model (C5) that establishes the highest possible discriminative threshold for video anomaly detection.

The ablation studies confirm that the superior performance of the swin-3DART framework is not due to a single component, but rather the synergistic integration of the TG-RGB$^+$ modality (for robust motion cues), the 3DART module (for adaptive feature refinement) and the T-GAP strategy (for enhanced adversarial robustness). Each component contributes a necessary and quantifiable marginal gain, validating the overall architectural design.

4.4.3 | Fusion Mechanism Analysis of TG-RGB$^+$ Modality

The core novelty of the proposed fusion mechanism lies in demonstrating that direct summation is the optimal strategy, uniquely balancing superior computational efficiency–critical for real-time deployment–with robust performance gains derived from combining motion and appearance cues while preserving compatibility with pretrained 3D backbones. To further establish both the theoretical foundation and practical efficiency of the proposed TG-RGB$^+$ modality, an in-depth analysis of the fusion mechanism was performed, focusing on how temporal gradient (TG) and RGB features are integrated. The goal is to elucidate why direct summation outperforms channel concatenation and to evaluate the consequences on computational efficiency, representational capability and real-time feasibility.

The TG-RGB$^+$ modality combines motion-sensitive TG features with spatial and appearance-based RGB cues at the preprocessing stage—prior to model ingestion. Two fusion paradigms were examined: (i) channel concatenation, in which TG (2 channels) and RGB (3 channels) are stacked to form a 5-channel tensor, and (ii) direct summation, where normalized TG and RGB components are combined pixel-wise, preserving the standard 3-channel configuration. Although concatenation could theoretically increase feature diversity, it simultaneously enlarges the kernel dimension of the first convolutional layer, increases parameter count, and disrupts compatibility with RGB-pretrained models. In contrast, summation preserves semantic alignment, maintains pretrained weight coherence and promotes smoother gradient propagation between motion and appearance cues, resulting in a more stable and computationally lightweight representation.

To quantify these differences, the preprocessing phase configuration of input clips was analysed. Table 11 presents a detailed comparison for a 32-frame clip, measuring GFLOPs, energy consumption, inference time, and memory footprint. Summation requires 0.1108 GFLOPs and 472.84 MB of memory, while concatenation increases the computational cost to 0.1171 GFLOPs and 485.46 MB, corresponding to an approximate 6% increase in FLOPs and a 2.7% increase in memory footprint. Energy consumption follows a similar trend, with summation consuming 0.1125 mJ per clip, compared to 0.1190 mJ for concatenation. The additional two channels introduced by concatenation expand the first-layer convolutional kernel and intermediate feature maps, causing redundant computation and slower frame processing. In contrast, summation preserves compact input dimensions, efficiently leverages pretrained RGB backbones, and enables faster, more stable fusion with minimal overhead.

The cross-validation analysis demonstrates that the TG-RGB$^+$ (summation) strategy consistently and significantly outperforms

TABLE 11 | Preprocessing phase configuration of inputs comparing TG-RGB$^+$ fusion strategies (sum vs. concatenation) for a 32-frame clip.

Fusion type	Input channels	GFLOPs (per clip)	Energy (mJ)	Avg memory (MB)	Avg time/frame (ms)
TG-RGB$^+$ (concat)	5	0.1171	0.1190	485.46	1.941
TG-RGB$^+$ (sum)	3	0.1108	0.1125	472.84	1.829

TABLE 12 | Statistical significance analysis and performance comparison of fusion strategies using k-fold cross-validation (mean AUC ± SD).

Fusion strategy	UBI Fights	*P*-value	UCF-Crime	*P*-value	RLVS	*P*-value
RGB baseline	82.70 ± 0.30	—	73.07 ± 0.23	—	94.97 ± 0.49	—
TG-RGB$^+$ (concat)	88.77 ± 0.03	<0.001	70.19 ± 0.09	<0.001	67.03 ± 0.32	<0.001
TG − RGB$^+$ (sum)	**91.97 ± 0.03**	**<0.001**	**79.16 ± 0.21**	**<0.001**	**98.28 ± 0.10**	**<0.001**

both baseline and concatenation approaches. As shown in Table 12, this is proved on the UBI Fights dataset by a +9.27% AUC gain ($P < 0.001$) and a noticeable reduction in standard deviation from 0.30 to 0.03. This increase in performance and stability indicates that pixel-wise summation effectively guides and enhances the model's attention. Furthermore, independent two-sample *t*-tests yield *P*-values well below the $\alpha = 0.05$ significance threshold, enabling confident rejection of the null hypothesis (H_0) and confirming that these gains are genuine rather than stochastic artefacts of the training process.

Consistent with these findings, the high-complexity UCF-Crime dataset further highlights the advantages of the summation strategy. Specifically, summation achieves a robust +6.09% AUC improvement ($P < 0.001$), while the concatenation approach fails to surpass the baseline. This distinct divergence emphasizes the validity of the manifold preservation theory: by preserving a 3-channel input, the proposed method remains fully compatible with the pretrained spatial–temporal weights of the swin-transformer, whereas concatenation disrupts the learned feature manifold. This is confirmed by RLVS results, where summation achieves a near-perfect AUC of 98.28% ($P < 0.001$), compared to 67.03% for concatenation. This performance drop can be explained by a mismatch between input dimensionality and temporal density. In RLVS, video clips are short (5–7 s), resulting in sparse motion cues. Concatenation spreads this limited motion information across an expanded 5-channel input, forcing the transformer to relearn input representations without sufficient temporal context. In contrast, the TG-RGB$^+$ summation strategy preserves the original 3-channel structure, reinforcing the pretrained spatial–temporal manifold. It effectively acts as an attention prior, guiding the model toward motion-relevant regions without requiring weight re-initialization.

The effectiveness of the summation strategy is further supported by the hardware efficiency analysis in Table 11. By preserving the 3-channel input manifold, the summation strategy requires only 0.1108 GFLOPs and 472.84 MB of memory per 32-frame clip. In contrast, concatenation expands the input to five channels, resulting in a 6% increase in computational cost and a 2.7% increase in memory usage. More critically, this channel expansion disrupts the pretrained weight coherence of the swin-transformer backbone, which directly explains the manifold collapse observed on the RLVS dataset, where AUC drops to 67.03%. On the other hand, the synergy between low-level efficiency and high-level accuracy is particularly evident in long-duration surveillance streams. The 3.2% reduction in energy consumption (mJ) and faster frame processing time (1.829 ms) achieved by summation prevent thermal throttling and memory bottlenecks during continuous inference. Furthermore, the reduction in standard deviation (from 0.30 to 0.03 on UBI Fights) confirms that summation acts as an intrinsic prior motion, stabilizing the attention mechanism by reinforcing motion-salient regions within the existing spatial manifold.

Consequently, the analysis provides both the theoretical grounding (via manifold preservation) and empirical justification (via GFLOPs/AUC trade-off) for adopting direct summation. Unlike comparative studies that often accept the null hypothesis for similar model variants, these results demonstrate a clear architectural superiority. This approach ensures that the inclusion of temporal-gradient cues enhances anomaly discrimination without compromising the real-time feasibility of the swin-3DART framework in safety-critical, resource-constrained environments.

4.5 | Cross-Domain Generalization

These experiments aim to evaluate the generality of the violent action concept through cross-dataset experiments. To achieve this, the UCF-Crime and UBI Fights datasets are alternately used for training, with the model being evaluated on the other dataset in a cross-dataset approach. This methodology enables the assessment of the model's ability to recognize violent actions across different data distributions and recording conditions. UCF-Crime consists primarily of real-world surveillance footage, often captured in uncontrolled environments [7], whereas UBI Fights features more structured fight scenes. The generalization ability of the learned representations is assessed by training swin-3DART on one dataset and evaluating its performance on another. This approach helps analyse how well the model adapts to datasets with differing characteristics, such as variations in scene complexity, motion dynamics and environmental conditions [48].

Building on this, the remarkably high performance of swin-3DART with TG-RGB$^+$, approaching 90% across multiple evaluation metrics, raises the question of whether the model is overfitting to its training data. To investigate its generalization

TABLE 13 | Results of cross-dataset experiments on UCF-Crime and UBI Fights for binary classification.

Training dataset	Testing dataset	Method	ROC AUC
UBI Fights [48]	UCF-Crime [7]	CrimeNet [69] (2024)	71.72%
		Swin-3DART	73.65%
UCF-Crime [7]	UBI Fights [48]	DMIL ranking [7] * (2025)	70.26%
		Swin-3DART	74.83%

*Recalculated result.

ability, cross-domain experiments were conducted using the UCF-Crime and UBI Fights datasets. This experimental setup not only assesses the transferability of learned representations but also provides insight into the model's adaptability to diverse surveillance scenarios, –from the varied real-world anomalies in UCF-Crime to the focused violent interactions in UBI Fights.

The results demonstrate that swin-3DART effectively captures the concept of violent actions, maintaining high accuracy even on previously unseen datasets. Table 13 shows accuracy results from cross-dataset experiments, highlighting the model's performance when tested on unseen datasets. Specifically, the UBI Fights dataset, which focuses on a single category of violent interactions, achieved an AUC of approximately 74%, while UCF-Crime, a more diverse multi-scenario dataset encompassing various real-world anomalies, attained an AUC of 75%. The slightly higher performance on UCF-Crime suggests that training on a dataset with diverse scenarios provides a broader learning signal, improving the model's ability to generalize. Additionally, it is rare to find state-of-the-art studies that conducted this experiment to assess the robustness of their models.

By using TG-RGB$^+$ datasets, feature diversity is improved, which helps mitigate overfitting and ensures robust anomaly detection across different domains. The high accuracy indicates that the model is not simply memorizing dataset-specific patterns but effectively learning transferable features applicable across multiple surveillance environments. However, further experiments on additional datasets would be beneficial to confirm this trend and explore potential failure cases where the model may struggle. These findings underscore swin-3DART's strong generalization capability across different domains, demonstrating its robustness in handling diverse datasets. This adaptability makes it a reliable solution for real-world security applications, where effective anomaly detection is crucial.

4.6 | Efficiency and Scalability Analysis

In real-world video understanding applications, –such as smart surveillance, edge-enabled security networks and mobile cognitive platforms, –efficiency is as critical as accuracy. Particularly in the context of video anomaly detection, where continuous streams must be monitored under strict energy and hardware constraints, models must balance computational cost with high spatio–temporal sensitivity. To this end, the swin-3DART architecture is assessed not only on predictive performance but also in terms of FLOPs, parameter count, and memory usage.

The video swin transformer base (swin-B) is adapted as the primary baseline due to its strong performance and wide adoption in video recognition tasks [49]. Pretrained on Kinetics-400, swin-B provides a meaningful reference point to evaluate the impact of the 3DART integration (Tables 14, 15).

Despite having an identical parameter count (88.8M), swin-3DART achieves significant gains in efficiency over swin-B. The model reduces the floating point operations (FLOPs) from 321 GFLOPs to 282.79 GFLOPs per video clip–an improvement of over 12%. This translates into lower energy consumption and faster inference, crucial for edge-based anomaly detection where latency and resource budgets are tight. Even more compelling is the reduction in memory usage. Swin-3DART consumes 4389.72 MB of memory during inference, compared to 4990.23 MB for swin-B. This 600+ MB saving directly enhances model scalability and deployability, particularly on edge devices or real-time systems with strict memory budgets.

Compared to other state-of-the-art architectures, –such as MViT-B [71](455 GFLOPs), ViViT-L [22] (903 GFLOPs) or TimeSformer-HR [70] (1703 GFLOPs)swin-3DART achieves a highly favourable trade-off. It offers competitive or superior accuracy while substantially lowering both FLOPs and memory consumption. These results clearly demonstrate the effectiveness of the 3DART module in optimizing transformer-based video models for real-world applications.

Comprehensive evaluations on public benchmarks further reinforce this balance, as swin-3DART achieves state-of-the-art AUCs of 95% on UBI Fights, 86% on UCF-Crime and 99% on RLVS. These results clearly demonstrate the effectiveness of the 3DART module in optimizing transformer-based video models for real-world applications.

Regarding the T-GAP perturbation, it begins with optical flow extraction, which can be precomputed offline. This design ensures that the online inference stage is dominated by the swin-3DART forward pass rather than the optical flow calculation itself, maintaining high efficiency even when adversarial perturbations (via T-GAP) are applied. Notably, the additional computational cost introduced by T-GAP and TG-RGB$^+$ is extremely small compared to the swin-3DART backbone. Specifically, the TG-RGB$^+$ fusion adds merely **0.11 GFLOPs** (**58.53 ms**), representing less than **0.04%** of the total computation. Swin-3DART's forward pass remains the dominant operation, requiring **282.79 GFLOPs** (**13.93 s per clip**), leading to a total pipeline cost of **282.90 GFLOPs** and **13.99 s per clip**, which demonstrates the high computational efficiency of the proposed design.

Despite the use of temporal gradients as part of the TG-RGB$^+$ transformation, swin-3DART achieves a highly favourable efficiency–accuracy trade-off. The model maintains state-of-the-art detection performance across benchmarks, achieving AUCs of 95% on UBI Fights, 86% on UCF-Crime, and 99% on RLVS, while consuming fewer FLOPs and memory than larger models such as TimeSformer-HR (1703 GFLOPs), ViViT-L (903 GFLOPs), and MViT-B (455 GFLOPs). By separating offline and online

TABLE 14 | Comparison of recent video recognition models. Swin-3DART achieves better efficiency while maintaining competitive performance. FLOPs are reported per clip.

Method	Pretrain	Dimension	FLOPs (G)	Params (M)	Memory (MB)
TimeSformer-HR [70]	ImageNet-21K	3D	1703	121.4	23200
ViViT-L/16x2 [22]	ImageNet-21K	3D	903	352.1	—
MViT-B, 64×3 [71]	K400	3D	455	36.6	—
Swin-B [72]	K400	3D	321	88.8	4990.23
Swin-3DART (ours)	K400	3D	**282.79**	**88.8**	**4389.72**

TABLE 15 | Computational overhead and inference time of the proposed method

Component	GFLOPs per clip	Inference time (ms/clip)	Energy (mj)
TG-RGB$^+$	0.11	58.53	0.112
Swin-3DART	282.79	13 932.68	101.80
Total pipeline	**282.90**	**13,991.21**	**101.912**

computations, swin-3DART ensures practical deployability in real-time, resource-constrained environments, such as edge-based surveillance systems, without sacrificing accuracy. The combination of TG-RGB$^+$ input and the 3DART feature refinement module allows rich spatiotemporal representations to be processed efficiently, confirming the model's suitability for energy-aware, high-performance video anomaly detection.

Overall, swin-3DART demonstrates a highly favourable trade-off between efficiency and accuracy, validating the role of the 3DART module in optimizing transformer-based video models for resource-constrained anomaly detection scenarios. This positions swin-3DART as a practical solution for energy-aware, intelligent surveillance systems deployed at the edge.

5 | Conclusions and Future Work

This paper presented swin-3DART, a unified and efficient framework for video-based violence detection that enhances motion sensitivity, adversarial robustness and computational efficiency. By introducing the TG-RGB$^+$ modality, the T-GAP adversarial perturbation strategy, and the 3D adaptive receive transformer (3DART)–which reduces memory footprint by ~12% and FLOPs by ~11.9% without adding extra parameters, swin-3DART addresses core limitations of existing transformer-based detectors. Extensive evaluations demonstrate swin-3DART's state-of-the-art performance, achieving AUCs of 95% on UBI Fights, 86% on UCF-Crime, and 99% on RLVS, with strong cross-dataset generalization (~74% and 75% ROC AUC). These results show that swin-3DART effectively detects anomalies in both dense and sparse environments, even under unseen conditions. The efficiency gains from 3DART enable real-time, on-device inference, making the model suitable for deployment in resource-constrained or edge-computing scenarios. Furthermore, the low attack success rates (4%–7%) under T-GAP are vital for ensuring robustness in safety-critical applications.

Robustness to environmental illumination was additionally validated through specialized low-light tests. By applying mid-level low-light transformations to RGB frames before computing temporal gradient features, swin-3DART demonstrates strong resilience: performance degradation is minimal across datasets (UBI fights: −4%, UCF-crime: −0.6%, RLVS: −2%), These results confirm that TG-RGB$^+$ features effectively preserve motion cues under poor lighting conditions, ensuring reliable deployment in suboptimal surveillance settings.

Despite its strengths, swin-3DART has limitations. First, the use of a fixed temporal-gradient interval can miss ultra-fast motions or underperform on videos with variable frame rates, suggesting a need for adaptive Δt mechanisms. Second, the current adversarial strategy targets only the temporal gradient channel; joint perturbations across both RGB and gradient streams could expose residual vulnerabilities. Third, although 3DART improves efficiency, the overall transformer backbone remains more computationally demanding than lightweight CNNs, which could hinder deployment on highly constrained devices. Finally, the evaluations were conducted on three datasets, UBI Fights, UCF-Crime, and RLVS, which do not fully cover extreme environmental conditions or complex occlusions. In addition, some datasets (e.g. UCF-Crime) exhibit severe class imbalance, challenging anomaly detection for sparse events.

To address these limitations, future work will focus on developing adaptive temporal encoding techniques capable of dynamically adjusting Δt for each video segment and on extending T-GAP into a multi-channel adversarial training framework that also perturbs RGB inputs. Additional modalities (e.g. audio or depth) will be incorporated to enhance contextual understanding, and self-supervised pretraining on large, unlabelled video corpora will be explored to reduce reliance on annotated data. Furthermore, expand evaluations to challenging conditions like adverse weather, and address class imbalance and domain-specific artefacts. Finally, efforts will be directed toward optimizing the model for edge deployment, considering latency, memory, and computational efficiency. These efforts will ensure swin-3DART remains robust, generalizable, and deployable under real-world constraints.

Author Contributions

Intissar Ziani: conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing – original draft preparation, writing – review and editing, visualization, project

administration. **Gueltoum Bendiab:** conceptualization, methodology, validation, writing – original draft preparation, writing – review and editing. **Mourad Bouzenada:** validation, writing – review and editing, supervision. **Meriem Guerar:** writing – review and editing, supervision, funding acquisition. All authors have read and agreed to the published version of the manuscript

Acknowledgements

This work was partially supported by project SERICS - "Security and Rights in CyberSpace" (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

Open access publishing facilitated by Universita degli Studi di Genova, as part of the Wiley - CRUI-CARE agreement.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The datasets used in this study are publicly available and were utilized solely for academic research purposes.

Endnotes

[1]https://www.kaggle.com/

References

1. Mordor Intelligence, "Europe Video Surveillance Market Size & Share Analysis—Growth Trends & Forecasts (2025–2030)," accessed June 3, 2025, https://www.mordorintelligence.com/industry-reports/europe-video-surveillance-systems-market (2025).

2. J. L. S. González, C. Zaccaro, J. A. Álvarez-García, L. M. S. Morillo, and F. S. Caparrini, "Real-Time Gun Detection in CCTV: An Open Problem," *Neural Networks* 132 (2020): 297–308.

3. A. Karbalaie, F. Abtahi, and M. Sjöström, "Event Detection in Surveillance Videos: A Review," *Multimedia Tools and Applications* 81, no. 24 (2022): 35463–35501.

4. G. Sreenu and S. Durai, "Intelligent Video Surveillance: A Review Through Deep Learning Techniques for Crowd Analysis," *Journal of Big Data* 6, no. 1 (2019): 1–27.

5. E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence Detection in Video Using Computer Vision Techniques," in *Computer Analysis of Images and Patterns* (Springer, 2011), 332–339.

6. M. Ramzan, A. Abid, H. U. Khan, et al., "A Review on State-of-the-Art Violence Detection Techniques," *IEEE Access* 7 (2019): 107560–107575.

7. 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* (IEEE, 2018), 6479–6488.

8. P. Negre, R. S. Alonso, A. González-Briones, J. Prieto, and S. Rodríguez-González, "Literature Review of Deep-Learning-Based Detection of Violence in Video," *Sensors* 24, no. 12 (2024): 4016.

9. Z. Liu, J. Ning, Y. Cao, et al., "Video Swin Transformer," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2022), 3202–3211.

10. B. K. Horn and B. G. Schunck, "Determining Optical Flow," *Artificial Intelligence* 17, no. 1–3 (1981): 185–203.

11. L. Wang, Y. Xiong, Z. Wang, et al., "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," in *Proceedings of the European Conference on Computer Vision* (Springer, 2016), 20–36.

12. S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang, "Optical Flow Guided Feature: A Fast and Robust Motion Representation for Video Action Recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), 1390–1399.

13. J. Xiao, L. Jing, L. Zhang, et al., "Learning from temporal gradient for semi-supervised action recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2022), 3252–3262.

14. J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2017), 6299–6308.

15. J. H. Park, M. Mahmoud, and H. S. Kang, "Conv3D-Based Video Violence Detection Network Using Optical Flow and RGB Data," *Sensors* 24, no. 2 (2024): 317.

16. C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast Networks for Video Recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, 2019), 6202–6211.

17. D. Kondratyuk, L. Yuan, Y. Li, et al., "MoViNets: Mobile Video Networks for Efficient Video Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2021), 16020–16030.

18. N. Jaafar and Z. Lachiri, "Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance," *Expert Systems With Applications* 211 (2023): 118523.

19. F. Mehmood, X. Guo, E. Chen, M. A. Akbar, A. A. Khan, and S. Ullah, "Extended Multi-Stream Temporal-Attention Module for Skeleton-Based Human Action Recognition (HAR)," *Computers in Human Behavior* 163 (2025): 108482.

20. Y. Du, Z. Hou, E. Lin, X. Li, J. Liang, and X. Zhou, "PRG-Net: Point Relationship-Guided Network for 3D Human Action Recognition," *Neurocomputing* 635 (2025): 130015.

21. T. Wolf, L. Debut, V. Sanh, et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (Association for Computational Linguistics, 2020), 38–45.

22. A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A Video Vision Transformer," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2021), 6836–6846.

23. M. A. R. Iftee, M. M. Rahman, and S. Das, "VioNet: An Enhanced Violence Detection Approach for Videos Using a Fusion Model of Vision Transformer With Bi-LSTM and 3D Convolutional Neural Networks," in *Proceedings of the 2nd International Conference on Big Data, IoT and Machine Learning* (Springer, 2023), 139–151.

24. H. Mohammadi, E. Nazerfard, and T. Firoozi, "Reinforcement Learning-Based Mixture of Vision Transformers for Video Violence Recognition,"*arXiv:2310.03108* (2023).

25. L. Zhou, W. Li, Y. Chen, H. Liu, M. Yang, and Z. Liu, "Human Keypoint Change Detection for Video Violence Detection Based on Cascade Transformer," in *BIM: International Conference on Big Data, IoT and Machine Learning* (Springer, 2023), 88–94.

26. M. Jayamohan and S. Yuvaraj, "A Novel Human Action Recognition Using Grad-CAM Visualization With Gated Recurrent Units," *Neural Computing and Applications* 37, no. 17 (2025): 10835–10850.

27. J. Manoharan and Y. Sivagnanam, "A Novel Human Action Recognition Model by Grad-CAM Visualization With Multi-Level Feature Extraction Using Global Average Pooling With Sequence Modeling by Bidirectional Gated Recurrent Units," *International Journal of Computational Intelligence Systems* 18, no. 1 (2025): 118.

28. Z. Liu, Y. Lin, Y. Cao, et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2021), 10012–10022.

29. X. Huang, H. Gong, and J. Zhang, "HST-MRF: Heterogeneous Swin Transformer With Multi-Receptive Field for Medical Image Segmentation," *IEEE Journal of Biomedical and Health Informatics* 28, no. 7 (2024): 4048–4061.

30. M. Heidari, R. Azad, S. G. Kolahi, et al., "Enhancing Efficiency in Vision Transformer Networks: Design Techniques and Insights,"*arXiv:2403.19882* (2024).

31. S. Li, A. Neupane, S. Paul, C. Song, S. V. Krishnamurthy, A. K. R. Chowdhury, and A. Swami, "Adversarial Perturbations Against Real-Time Video Classification Systems,"*arXiv:1807.00458* (2018).

32. F. Dong, Y. Zhang, and X. Nie, "Dual Discriminator Generative Adversarial Network for Video Anomaly Detection," *IEEE Access* 8 (2020): 88170–88176.

33. Y. Izza and J. Marques-Silva, "The Pros and Cons of Adversarial Robustness,"*arXiv:2312.10911* (2023).

34. Z. Chen, L. Xie, S. Pang, Y. He, and Q. Tian, "Appending Adversarial Frames for Universal Video Attack," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE, 2021), 3199–3208.

35. H. S. Kim, M. Son, M. Kim, M. J. Kwon, and C. Kim, "Breaking temporal consistency: Generating video universal adversarial perturbations using image models," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2023), 4325–4334.

36. M. Yin, Y. Xu, T. Hu, and X. Liu, "A Robust Adversarial Example Attack Based on Video Augmentation," *Applied Sciences* 13, no. 3 (2023): 1914.

37. Y. Pan, J. J. Huang, Z. Chen, W. Zhao, and Z. Wang, "Svastin: Sparse Video Adversarial Attack via Spatio-Temporal Invertible Neural Networks," in *2024 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2024), 1–6.

38. Z. Wei, J. Chen, Z. Wu, and Y. G. Jiang, "Cross-Modal Transferable Adversarial Attacks From Images to Videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2022), 15064–15073.

39. J. Li, K. Gao, Y. Bai, J. Zhang, S. T. Xia, and Y. Wang, "FMM-Attack: A Flow-Based Multi-Modal Adversarial Attack on Video-Based LLMs," *arXiv:2403.13507* (2024).

40. N. Inkawhich, G. McDonald, and R. Luley, "Adversarial Attacks on Foundational Vision Models,"*arXiv:2308.14597* (2023).

41. W. Chen, K. T. Ma, Z. J. Yew, M. Hur, and D. A. A. Khoo, "TEVAD: Improved Video Anomaly Detection With Captions," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (IEEE, 2023), 5549–5559.

42. W. Sun, L. Cao, Y. Guo, and K. Du, "Multimodal and Multiscale Feature Fusion for Weakly Supervised Video Aanomaly Detection," *Scientific Reports* 14, no. 1 (2024): 22835.

43. G. Garcia-Cobo and J. C. SanMiguel, "Human Skeletons and Change Detection for Efficient Violence Detection in Surveillance Videos," *Computer Vision and Image Understanding* 233 (2023): 103739.

44. A. Ghadiya, P. Kar, V. Chudasama, and P. Wasnik, "Cross-Modal Fusion and Attention Mechanism for Weakly Supervised Video Anomaly Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2024), 1965–1974.

45. M. A. B. Abbass and H. S. Kang, "Violence Detection Enhancement by Involving Convolutional Block Attention Modules Into Various Deep Learning Architectures: Comprehensive Case Study for UBI-Fights Dataset," *IEEE Access* 11 (2023): 37096–37107.

46. B. Degardin and H. Proença, "Iterative Weak/Self-Supervised Classification Framework for Abnormal Events Detection," *Pattern Recognition Letters* 145 (2021): 50–57.

47. I. Ziani, G. Bendiab, M. Bouzenada, and S. Shiaeles, "Video-Based Abnormal Human Behaviour Detection for Video Forensics," in *2024 IEEE International Conference on Cyber Security and Resilience (CSR)* (IEEE, 2024), 401–406.

48. B. M. Degardin, "Weakly and Partially Supervised Learning Frameworks for Anomaly Detection," (master's thesis, Universidade da Beira Interior, 2020).

49. Z. Liu, H. Hu, Y. Lin, et al., "Swin Transformer V2: Scaling Up Capacity and Resolution," in *2024 IEEE International Conference on Cyber Security and Resilience (CSR)* (IEEE, 2022), 12009–12019.

50. H. S. Heo, J. W. Jung, H. J. Shim, I. H. Yang, and H. J. Yu, "Cosine Similarity-Based Adversarial Process," *arXiv:1907.00542* (2019).

51. M. M. Soliman, M. H. Kamal, M. A. E. M. 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)* (IEEE, 2019), 80–85.

52. S. S. Shapiro and M. B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika* 52, no. 3–4 (1965): 591–611.

53. H. Hsu, P. Lachenbruch, P. Armitage, and T. Colton, "Paired *t* Test. in Encyclopedia of Biostatistics," (Wiley, 2005).

54. E. A. Gehan, "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples," *Biometrika* 52, no. 1–2 (1965): 203–224.

55. A. R. Abdali, "Data Efficient Video Transformer for Violence Detection," in *2021 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)* (IEEE, 2021), 195–199.

56. Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-Supervised Video Anomaly Detection With Contrastive Learning of Long and Short-Range Temporal Features,"*arXiv:2101.10030* (2021).

57. W. Ullah, F. U. M. Ullah, Z. A. Khan, and S. W. Baik, "Sequential Attention Mechanism for Weakly Supervised Video Anomaly Detection," *Expert Systems with Applications* 230 (2023): 120599.

58. W. Dang, X. Mao, and L. Pan, "Weakly Supervised Video Anomaly Detection Based on Adaptive Fusion of Multimodal Features," *International Journal of Pattern Recognition and Artificial Intelligence* 40, no. 4 (2025): 2550041.

59. S. Chang, Y. Li, S. Shen, J. Feng, and Z. Zhou, "Contrastive Attention for Video Anomaly Detection," *IEEE Transactions on Multimedia* 24 (2021): 4067–4076.

60. Y. Watanabe, M. Okabe, Y. Harada, and N. Kashima, "Real-World Video Anomaly Detection by Extracting Salient Features in Videos," *IEEE Access* 10 (2022): 125052–125060.

61. Z. Yang, Y. Guo, J. Wang, D. Huang, X. Bao, and Y. Wang, "Towards Video Anomaly Detection in the Real World: A Binarization Embedded Weakly-Supervised Network," *IEEE Transactions on Circuits and Systems for Video Technology* 34, no. 5 (2023): 4135–4140.

62. Y. Wang, J. Zhou, and J. Guan, "A Lightweight Video Anomaly Detection Model With Weak Supervision and Adaptive Instance Selection," *Neurocomputing* 613 (2025): 128698.

63. J. C. Feng, F. T. Hong, and W. S. Zheng, "MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2021), 14009–14018.

64. H. Chen, X. Mei, Z. Ma, X. Wu, and Y. Wei, "Spatial–Temporal Graph Attention Network for Video Aanomaly Detection," *Image and Vision Computing* 131 (2023): 104629.

65. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks,"*arXiv:1706.06083* (2017).

66. I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples,"*arXiv:1412.6572* (2014).

67. H. Xu, Y. Ma, H. C. Liu, et al., "Adversarial Attacks and Defenses in Images, Graphs and Text: A Review," *International Journal of Automation and Computing* 17 (2020): 151–178.

68. F. Mumcu, K. Doshi, and Y. Yilmaz, "Adversarial Machine Learning Attacks Against Video Anomaly Detection Systems," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (IEEE, 2022), 206–213.

69. F. J. Rendón-Segador, J. A. Álvarez-García, and L. M. Soria-Morillo, "Transformer and Adaptive Threshold Sliding Window for Improving Violence Detection in Videos," *Sensors* 24, no. 16 (2024): 5429.

70. G. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding?" in *Proceedings of the 38th International Conference on Machine Learning* (PMLR, 2021), 1–12.

71. H. Fan, B. Xiong, K. Mangalam, et al., "Multiscale Vision Transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, 2021), 6824–6835.

72. L. Chen, Y. Bai, Q. Cheng, and M. Wu, "Swin Transformer With Local Aggregation," in *2022 3rd International Conference on Information Science, Parallel and Distributed Systems (ISPDS)* (IEEE, 2022), 77–81.