International Journal of Safety and Security Engineering
Vol. 15, No. 10, October, 2025, pp. 2143-2154
Journal homepage: http://iieta.org/journals/ijsse

A Comprehensive Survey of Transformer-Based Models for Video Anomaly Detection in Surveillance Systems

Sayali B. Sabale[1,2], Vijayshri N. Khedkar[3*]

[1] Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, India
[2] Department of Information Technology, Dr. D.Y. Patil Institute of Technology, Pimpri, Pune 412115, India
[3] Department of Computer Science and Engineering, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, India

Corresponding Author Email: vijayshri.khedkar@sitpune.edu.in

https://doi.org/10.18280/ijsse.151017

Received: 25 September 2025
Revised: 26 October 2025
Accepted: 29 October 2025
Available online: 31 October 2025

Keywords:
anomaly detection, video surveillance, transformer models, deep learning, abnormal event

ABSTRACT

Video anomaly detection constitutes a pivotal component for ensuring public safety, regulating traffic networks, supervising industrial workflows, and enabling smart city ecosystems. Recent advances in deep learning - particularly transformer-based architectures - have markedly improved the modeling of high-order spatio-temporal dependencies within surveillance video streams. This study presents a systematic comparative evaluation of state-of-the-art frameworks, encompassing CNN–Transformer hybrids, dual-stream motion–appearance encoders, pure vision transformer architectures, weakly supervised paradigms, and class-incremental learning strategies. Experiments conducted on benchmark datasets including UCF-Crime, ShanghaiTech, CUHK Avenue, UCSD Ped2, RWF-2000, and Drone-Anomaly highlight domain-specific advantages: BiMT achieves superior accuracy on UCF-Crime; TDS-Net demonstrates robustness on motion-intensive corpora such as ShanghaiTech and Avenue; and unsupervised transformer models excel in aerial anomaly detection. Furthermore, SwinIoT provides edge-optimized inference for IoT-enabled smart environments, while CILAR-Net supports dynamic integration of emergent anomalous classes. The analysis underscores critical trade-offs—labeling cost reduction via ST-HTAM, real-time efficiency through ViT+SRU++, and anomaly localization achieved by SwinAnomaly. The findings indicate that no single architecture universally dominates across tasks; instead, optimal model selection is context-sensitive, determined by accuracy–efficiency requirements, annotation costs, adaptability, and deployment constraints. The contribution is a decision-support framework for selecting transformer-based anomaly detection models across heterogeneous video surveillance domains.

1. INTRODUCTION

The rapid proliferation of surveillance cameras across public, commercial, and industrial environments has significantly increased the demand for automated video anomaly detection (VAD) systems capable of monitoring large-scale visual data streams in real time. Manual inspection of continuous video feeds is inefficient and highly error-prone, creating a need for intelligent anomaly detection frameworks that ensure public safety and operational reliability. However, VAD remains challenging due to the rarity of abnormal events, the strong influence of context, scene variability, and ambiguous anomaly definitions [1].

Early approaches relied on handcrafted features such as trajectories, optical-flow histograms, or spatiotemporal gradients, combined with statistical models or classical machine learning algorithms. These methods performed reasonably in controlled environments but struggled in crowded scenes and complex real-world settings due to limited representational capacity and an inability to model long-range dependencies. The adoption of deep learning marked a major shift: CNNs captured spatial semantics, RNNs (LSTM/GRU) incorporated temporal sequence modeling, and autoencoders/GANs enabled reconstruction-based anomaly detection Nevertheless, these architectures suffered from restricted receptive fields, vanishing gradients, and difficulty modeling global spatiotemporal relationships critical for anomaly understanding [2].

A key milestone was the introduction of attention-based models in computer vision. Non-local Networks (2018) and early temporal attention frameworks demonstrated that attention mechanisms could capture global contextual relations more effectively than RNNs by directly relating distant spatial and temporal positions. This evolution paved the way for the Vision Transformer (ViT) and hierarchical variants such as Swin Transformer which redefined video understanding by representing frames as token sequences and modeling long-range dependencies via self-attention. Their capacity to jointly capture global structure, fine-grained local details, and multi-scale spatial-temporal relations has made transformers particularly suitable for VAD tasks, where subtle cues and contextual deviations are often key indicators of irregular events [3].

The effectiveness of transformer architectures for VAD is supported by strong theoretical advantages:

1. **Global Self-Attention:** Enables modeling of long-range dependencies and contextual interactions across frames—critical for detecting temporally extended anomalies such as stalking, fighting, or slow abnormal motion patterns.
2. **Token-Based Representation:** Allows flexible modeling of appearance, motion, and scene structure at patch or object level, improving responsiveness to fine-grained anomalies.
3. **Long-Range Temporal Modeling:** Overcomes vanishing gradients and recurrence bottlenecks seen in LSTM/GRU models, facilitating more stable temporal reasoning [3].
4. **Information-Theoretic Efficiency:** Attention mechanisms dynamically highlight salient regions and suppress irrelevant information, acting as adaptive spatiotemporal compression—an essential property in cluttered surveillance scenes.

Despite rapid advancements, significant limitations remain. Many transformer-based models are computationally expensive and unsuitable for real-time or edge-based deployments characteristic of IoT surveillance environments. Weakly supervised and unsupervised transformer paradigms often struggle to generalize to unseen anomaly types and may suffer from high false-positive rates in complex scenes [4]. Multimodal and hybrid architectures require large-scale annotated datasets, which are scarce in real-world deployment scenarios. Additionally, several models lack mechanisms for incremental learning, explainability, or domain adaptation - critical requirements for long-term monitoring and dynamic urban environments [5].

To address these gaps, this survey presents a comprehensive and theoretically grounded analysis of transformer-based VAD models. Unlike existing surveys that primarily categorize architectures, this work offers:

(1) a theoretical explanation for why transformer mechanisms align with the intrinsic requirements of anomaly detection;

(2) a historical progression connecting early deep learning and attention-based methods with modern transformers [6];

(3) a systematic comparison of supervised, weakly supervised, unsupervised, hybrid, hierarchical, and multimodal transformer frameworks [7];

(4) a critical interpretation of performance differences using theoretical constructs such as attention span, fusion strategy, and modality alignment; and

(5) a decision-support framework that assists practitioners in selecting the most appropriate transformer architecture based on deployment constraints, labeling resources, and application domain [5].

The remainder of this paper is organized as follows. Section II presents related work. Section III deals with Abnormal event detection using transformer Models. Section IV provides a comparative analysis of results with different dataset, performances, and compromises, and Section V Covers Application domain and best model and lastly Section VI presents conclusion and future work of this survey paper.

2. RELATED WORK

2.1 Evolution of video anomaly detection (VAD)

Video anomaly detection has evolved significantly over the past decade, moving from handcrafted feature engineering to deep learning and more recently to transformer-based architecture. Early VAD systems relied on manually designed features such as optical flow histograms, trajectories, or spatiotemporal gradients. These classical approaches were limited in crowded and unstructured scenes due to weak generalization and poor robustness to motion clutter [4].

With the emergence of deep learning, convolutional neural networks (CNNs) enabled stronger spatial feature extraction, while recurrent networks (LSTM/GRU) improved temporal modeling by capturing sequential dependencies. Autoencoders and GAN-based frameworks further advanced VAD through reconstruction and prediction paradigms, where anomalies are detected as deviations from learned normal patterns. However, deep learning approaches still struggled with modeling long-range dependencies, global context, and multimodal interactions—key factors in complex surveillance environments [7]. The historical progression of video anomaly detection (VAD) is depicted in Figure 1.

These limitations created a natural progression toward attention-based models. Non-local networks and early temporal-attention mechanisms demonstrated the value of global context modeling, laying the foundation for the adoption of transformer architectures for VAD.

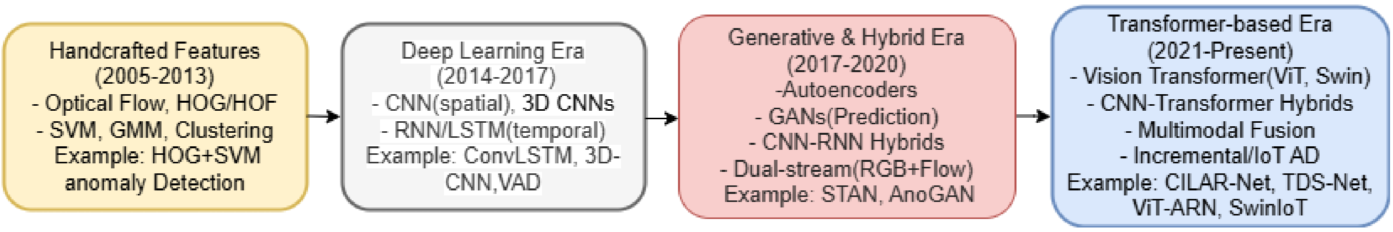

Figure 1. History of video anomaly detection (VAD)

2.2 Classical deep learning for VAD

Classical deep learning methods can be grouped into three major categories:

- **CNN-based anomaly detectors:**

CNNs excel at extracting spatial semantics but lack explicit temporal reasoning. Surveys highlight their limitations in capturing long-range motion patterns or contextual relations across scenes [8].

- **RNN-based models (LSTM, GRU, BiLSTM):**

RNNs address temporal dynamics but suffer from vanishing gradients and limited temporal horizons, making them insufficient for complex anomalies such as prolonged loitering or staged abnormal behaviors [9].

- **Autoencoders and GANs:**

Reconstruction- and prediction-based frameworks detect anomalies through high reconstruction errors. However, these models often produce blurry reconstructions and do not leverage global scene-level dependencies, resulting in high false positives in dynamic scenes [10].

These limitations motivated the transition toward hybrid and transformer-based systems capable of global spatiotemporal modeling. Figure 2 presents an overview of deep learning approaches applied to anomaly detection.

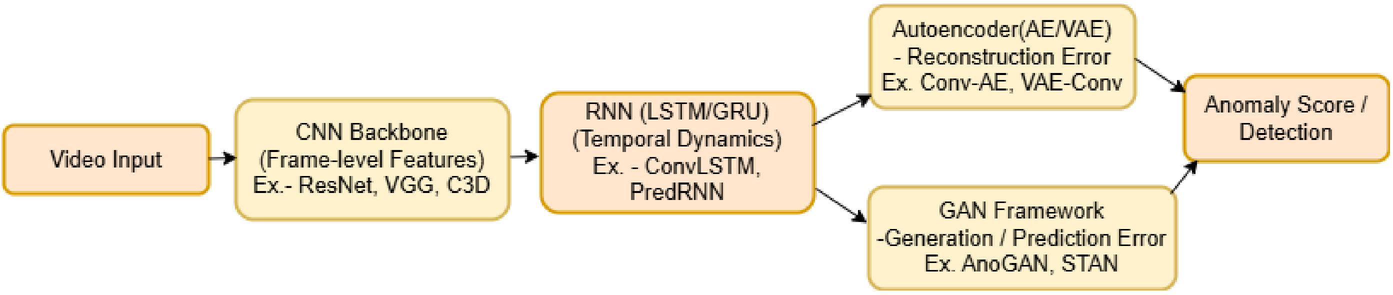

Figure 2. Deep learning for anomaly detection

2.3 Hybrid CNN/RNN-transformer architectures

Hybrid designs combine the strengths of CNNs, RNNs, and Transformers:

- **BiMT** integrates CNN spatial encoders, BiLSTM temporal encoders, and transformer-based global attention, achieving high accuracy across UCSD, Avenue, and ShanghaiTech datasets [11].
- **TransCNN** fuses CNN-based spatial extraction with transformer-driven temporal reasoning, providing superior generalization over purely convolutional pipelines [12].
- **Multimodal hybrid frameworks** fuse RGB, depth, optical flow, and audio through cross-modal attention to enhance robustness in occlusion and low-light environments [13].

Hybrid architectures address transformer weaknesses such as computational cost and data requirements while preserving global reasoning capability. As shown in Figure 3, the hybrid CNN/RNN and Transformer components work together to capture both spatial and temporal features.

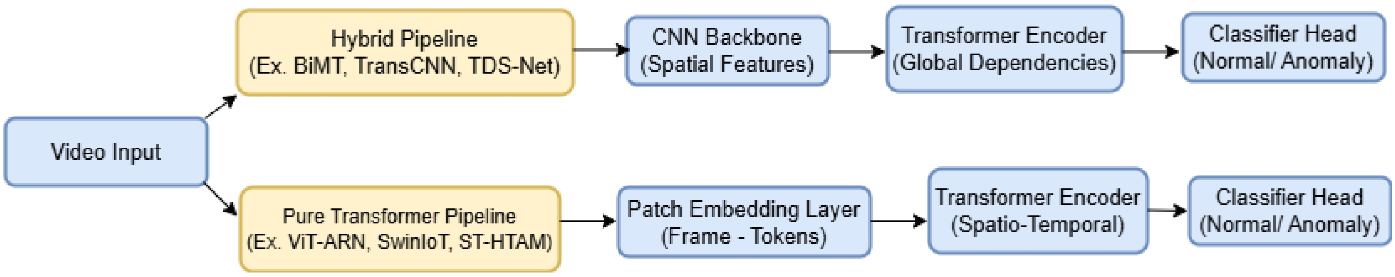

Figure 3. Hybrid CNN/RNN + transformer model

3. ABNORMAL EVENT DETECTION USING TRANSFORMER MODELS

3.1 Supervised transformer-based methods

Supervised transformer-based methods for video anomaly detection (VAD) rely on frame- or clip-level annotations to train discriminative models that directly classify anomalous versus normal behavior. These frameworks exploit the self-attention mechanism of transformers, which enables dense pairwise interactions among spatio-temporal tokens, thereby overcoming the locality bias of CNNs and the vanishing-gradient limitations of RNNs [14]. When combined with convolutional or recurrent backbones, supervised transformers demonstrate state-of-the art performance on benchmark surveillance datasets. A representative supervised framework is CILAR-Net (Class-Incremental Learning Network). Its backbone employs a Vision Transformer (ViT) [15] for hierarchical spatial feature encoding, producing patch embeddings that preserve object-level and contextual cues. Temporal dependencies are modeled through a Gated Recurrent Unit (GRU) module, while an incremental classifier accommodates newly emerging anomaly classes without catastrophic forgetting. By integrating incremental learning within a transformer pipeline, CILAR-NeT achieves high anomaly detection accuracy (97.2% AUC on UCSD Ped2), while uniquely supporting lifelong surveillance adaptability - a crucial requirement for evolving urban environments [16]. Another supervised transformer design is ViT-ARN (Vision Transformer Attention with Multi-Reservoir ESN). Here, a ViT encoder extracts non-local spatial dependencies via multi-head self-attention, while Echo State Networks (ESNs), serving as recurrent reservoirs, approximate long-term temporal dynamics. This hybridization replaces heavy recurrent training with reservoir computing, making the pipeline computationally efficient while maintaining sequence modeling fidelity. ViT-ARN [17] demonstrates strong performance on long-duration datasets such as UCF-Crime (88.1% AUC), illustrating the advantage of lightweight temporal reservoirs over conventional LSTMs in supervised transformer settings. Hybrid CNN–Transformer architectures further refine supervised VAD. TransCNN employs a CNN encoder (e.g., ResNet-based backbone) to extract spatial frame-level representations, which are projected into a sequence of tokens and passed through a transformer encoder for temporal attention modeling. The transformer enhances inter-frame reasoning by selectively attending to discriminative temporal patterns, yielding 90.3% AUC on Avenue and 85.4% on ShanghaiTech, surpassing CNN–LSTM counterparts [18]. Figure 4 presents the architecture of the supervised transformer-based framework adopted in this study.

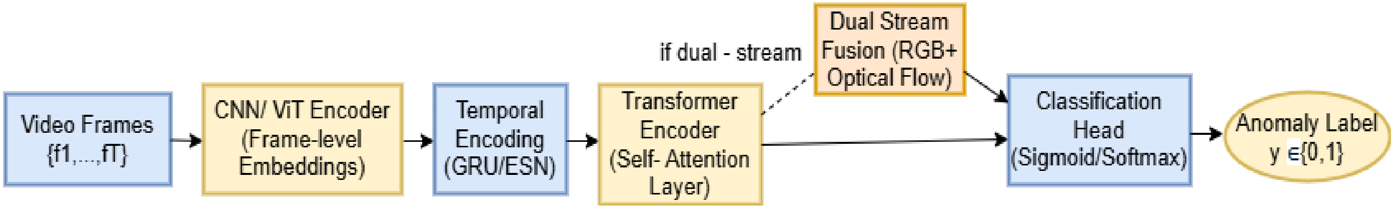
Video Frames {f1,...,fT}
CNN/ ViT Encoder (Frame-level Embeddings)
Temporal Encoding (GRU/ESN)
Transformer Encoder (Self- Attention Layer)
if dual - stream
Dual Stream Fusion (RGB+ Optical Flow)
Classification Head (Sigmoid/Softmax)
Anomaly Label y ∈{0,1}

Figure 4. Supervised transformer-based framework

The most competitive supervised design, TDS-Net (Transformer-enhanced Dual-Stream Network) [19], incorporates dual-stream feature encoders: an RGB stream for semantic context and an optical flow stream for explicit motion representation. Transformer encoders are deployed over both streams to capture intra-stream dynamics, followed by a cross-stream fusion stage that aligns spatial and motion cues. This design leverages both appearance-motion complementarity and long-range temporal self-attention.

In summary, supervised transformer-based approaches deliver the strongest anomaly detection performance across benchmarks due to their ability to unify fine-grained spatial encoding, explicit temporal modeling, and global context reasoning. However, their reliance on dense annotations makes them annotation-expensive and less scalable to large-scale, real-world deployments, motivating the parallel exploration of weakly supervised and unsupervised paradigms [20].

3.2 Unsupervised and self-supervised methods

Unsupervised and self-supervised frameworks in video anomaly detection (VAD) are designed to overcome the annotation bottleneck, where frame- or event-level anomaly labels are unavailable or impractical to obtain. Instead of explicit supervision, these methods exploit the distribution of normal video dynamics to detect deviations indicative of anomalies [20]. Transformers, with their ability to model long-range spatio-temporal dependencies, provide a robust backbone for such paradigms by capturing global contextual cues beyond the receptive field of CNNs and the short-term dependencies of RNNs. As shown in Figure 5, the unsupervised framework operates without labeled data for anomaly detection.

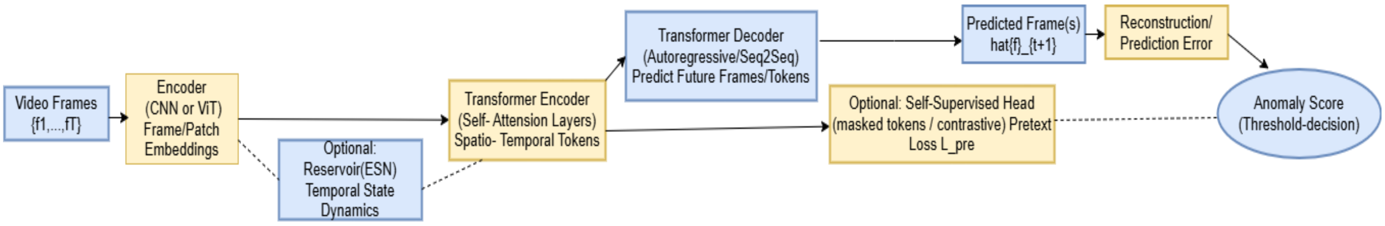
Video Frames {f1,...,fT}
Encoder (CNN or ViT) Frame/Patch Embeddings
Optional: Reservoir(ESN) Temporal State Dynamics
Transformer Encoder (Self- Attension Layers) Spatio- Temporal Tokens
Transformer Decoder (Autoregressive/Seq2Seq) Predict Future Frames/Tokens
Predicted Frame(s) hat{f}_{t+1}
Reconstruction/ Prediction Error
Optional: Self-Supervised Head (masked tokens / contrastive) Pretext Loss L_pre
Anomaly Score (Threshold-decision)

Figure 5. Unsupervised framework

3.2.1 Unsupervised reconstruction and prediction models

A canonical approach to unsupervised anomaly detection is predictive modeling, where the model learns to forecast future frames given past sequences, with anomalies emerging as prediction errors. The Transformer Encoder–Decoder framework for traffic anomaly detection exemplifies this paradigm. A sequence of normal frames $V = \{f_1, f_2, \dots, f_T\}$ is encoded into latent tokens $Z_t = \phi(f_t)$ through a CNN backbone. These tokens are processed by a transformer encoder to capture inter-frame attention, and the decoder predicts the next frame $f_{\{t+1\}}$. An anomaly score is derived from the reconstruction error:

$$L_{anom}(t) = \|f_{t+1} - f^{t+1}\|_2^2$$

This framework achieved 82% AUC on aerial traffic videos, demonstrating the feasibility of annotation-free anomaly detection, though performance is constrained by reconstruction noise in dynamic traffic environments [21].

3.2.2 Reservoir-enhanced transformer variants

Transformers are also combined with recurrent reservoirs for unsupervised temporal modeling. ViT-ARN (Vision Transformer with Multi-Reservoir Echo State Networks) leverages a ViT encoder for patch-level spatial embeddings and employs Echo State Networks (ESNs) to propagate temporal dynamics. ESNs maintain a fixed recurrent reservoir governed by

$$h_t = \tanh(W_{in}x_t + W_{res}h_{t-1})$$

where, W_{res} is a sparsely initialized reservoir matrix. This unsupervised adaptation of ViTARN exploits ESN states to model normal sequence dynamics, with anomalies detected via deviations in state-driven predictions. The reservoir's fixed dynamics ensure computational efficiency, avoiding the training overhead of deep RNNs, while maintaining sensitivity to long-term abnormal patterns [17].

3.2.3 Self-supervised representation learning

Beyond reconstruction, self-supervised paradigms employ auxiliary pretext tasks to enforce discriminative spatio-temporal representations without requiring anomaly labels. Examples include temporal order verification, masked token modeling, or contrastive learning, where the transformer learns to distinguish between plausible and corrupted video sequences [22]. In these frameworks, anomaly detection is performed by measuring the embedding distance between test sequences and the learned normal representation manifold. Though not explicitly detailed in the uploaded papers, several recent transformer-based works extend BERT-style masked video modeling to surveillance contexts, enabling transferable feature representations for anomaly detection under zero-label conditions [14].

3.3 Hybrid CNN/RNN-transformer

Hybrid CNN/RNN–Transformer frameworks represent the current state-of-the-art paradigm in video anomaly detection (VAD), combining the spatial locality capture of CNNs, the sequential memory of RNNs, and the global context reasoning of Transformers. This layered integration compensates for the shortcomings of individual architectures: CNNs excel at extracting appearance-level semantics but lack temporal modeling; RNNs capture sequence dynamics but struggle with long-range dependencies; and Transformers provide self-attention over arbitrarily distant frames but require strong feature encoders to prevent overfitting. Formally, given a sequence of video frames $V = \{ f_1, f_2, \dots, f_T\}$ a CNN encoder produces spatial embeddings: $X_t = \phi(f_t), X_t \in R^{N*d}$ where N denotes patch tokens or convolutional feature maps and their embedding dimension. To capture local temporal continuity, RNN layers (e.g., BiLSTM) refine embeddings: $H_t = BiLSTM(X_t, H_{t-1})$ capturing bidirectional short term and medium-term dependencies. These temporally enriched features are then processed by a transformer encoder:

$$Z = Transformer\{ H_1, H_2, \dots, H_T\}$$

Which applies multi-head self-attention to model global correlations and long-range anomaly indicative dependencies across frames. Finally, a classifier maps Z to anomaly scores.

BiMT (CNN–BiLSTM–Transformer) exemplifies this integration. A CNN backbone (e.g., ResNet) extracts appearance cues, a BiLSTM encodes sequential context in both forward and backward directions, and a transformer refines this with hierarchical temporal attention. By aligning local and global temporal representations, BiMT achieves 97.8% AUC (UCSD Ped2), 89.2% (Avenue), and 84.6% (ShanghaiTech), significantly outperforming CNN–LSTM pipelines that lack global self-attention [11].

TransCNN simplifies the hybridization by replacing RNNs with a direct CNN + Transformer pipeline. Spatial embeddings from CNN layers are tokenized and processed by transformers to temporal reasoning, reducing recurrence-induced bottlenecks. This design attains 98.1% (UCSD), 90.3% (Avenue), 85.4% (ShanghaiTech), illustrating the efficiency of transformer driven temporal modeling over recurrent memory [12]. Figure 6 presents the architecture of the hybrid framework that combines multiple learning paradigms.

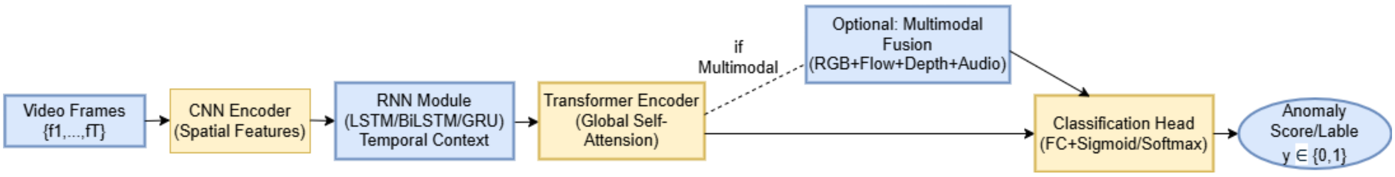
Video Frames {f1,...,fT}
CNN Encoder (Spatial Features)
RNN Module (LSTM/BiLSTM/GRU) Temporal Context
Transformer Encoder (Global Self-Attention)
if Multimodal
Optional: Multimodal Fusion (RGB+Flow+Depth+Audio)
Classification Head (FC+Sigmoid/Softmax)
Anomaly Score/Lable y ∈ {0,1}

Figure 6. Hybrid framework

TDS-Net (Transformer-enhanced Dual-Stream Network) further extends hybridization across modalities. Two parallel CNN encoders extract RGB and optical flow streams, each refined by transformer modules. A dual-stream fusion mechanism aligns motion and appearance representations before classification. This multimodal hybridization, leveraging complementary cues, achieves state-of-the-art supervised accuracy: 98.5% (UCSD Ped2), 91.0% (Avenue), and 86.1% (ShanghaiTech) [19].

Lastly, Multimodal Fusion + Transformer frameworks generalize hybrids by incorporating RGB, depth, infrared, and audio modalities. CNN encoders extract modality-specific features, transformers align them temporally, and cross-modal attention modules perform joint feature fusion. Such architectures demonstrate strong robustness under modality noise, e.g., 97.9% (UCSD) and 90.1% (Avenue), but face deployment challenges due to synchronization overhead and computational cost.

In summary, hybrid CNN/RNN–Transformer frameworks provide the most balanced and accurate solutions for VAD, offering fine-grained local representation (CNN), sequential learning (RNN), and global context modeling (Transformer). Their drawback lies in computational complexity and inference latency, motivating research on lightweight hybrids for real-time deployment.

3.4 Hierarchical and multimodal transformer methods

Hierarchical and multimodal transformer-based frameworks extend the transformer paradigm in VAD by addressing two complementary challenges: scalability of attention across spatial temporal hierarchies and integration of heterogeneous input modalities beyond RGB frames. By exploiting shifted-window attention and cross-modal fusion mechanisms, these models achieve both fine-grained local anomaly detection and robust global context reasoning across diverse surveillance environments. As shown in Figure 7, the hierarchical transformer framework captures multi-level temporal dependencies for effective anomaly detection.

3.4.1 Hierarchical transformer architectures

Traditional Vision Transformers (ViT) apply global self-attention across all tokens, leading to quadratic complexity O(N^2) with respect to the number of patches. This becomes computationally prohibitive for long video sequences. Hierarchical designs, such as Swin Transformers [23], mitigate this by computing attention within shifted local windows before progressively merging patches into coarser scales. Formally, at stage l, attention is computed over windowed tokens $X_1 \in R^{M\times d}$, where M is the number of tokens in a window and d is the feature dimension. The operations are defined as:

$$Z_l = W - MSA(X_l) + X_l,$$

$$X_{l+l} = MLP\big(SW - MSA(Z_l)\big),$$

where, *W-MSA* denotes window-based self-attention and *SW-MSA* applies shifted windows for cross-window interactions. This hierarchical design balances local spatial modeling with global dependency capture while reducing complexity to near-linear in sequence length. The STHTAM (Swin Transformer with Hierarchical Temporal Adaptive Module) integrates Swin Transformers for spatial tokenization with C-LSTMs and temporal attention modules for weakly supervised VAD. By hierarchically refining temporal dependencies under Multiple Instance Learning (MIL), ST-HTAM achieves strong anomaly localization performance (96.3% AUC on UCSD Ped2, 86.7% on Avenue), despite relying only on video-level labels. Its hierarchical architecture demonstrates that localized attention windows can improve both computational efficiency and local anomaly sensitivity [24].

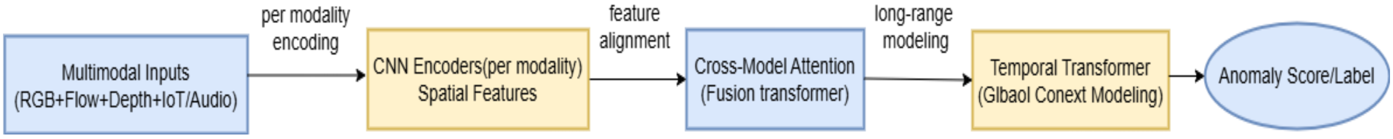
Multimodal Inputs (RGB+Flow+Depth+IoT/Audio)
per modality encoding
CNN Encoders(per modality) Spatial Features
feature alignment
Cross-Model Attention (Fusion transformer)
long-range modeling
Temporal Transformer (Glbaol Conext Modeling)
Anomaly Score/Label

Figure 7. Hierarchical transformer framework

Similarly, SwinIoT extends hierarchical transformers to IoT-driven smart city surveillance, where multimodal sensor streams (RGB video, IoT metadata, contextual environmental signals) must be jointly modeled. The Swin backbone encodes spatial hierarchies, while cross-layer fusion aligns multi-resolution embeddings, yielding robust detection (AUC ≈ 89.0) in resource constrained IoT deployments [23].

3.4.2 Multimodal transformer models

Multimodal anomaly detection exploits heterogeneous surveillance data - RGB video, optical flow, depth, infrared, and even audio to overcome the limitations of single-modality detection under noisy or occluded environments. Multimodal transformers achieve this by aligning modality-specific embeddings through cross-modal self-attention or fusion transformers.

Multimodal Fusion + Transformer Framework

The Multimodal Fusion + Transformer framework uses CNN encoders to process RGB, flow, and depth streams independently, producing modality-specific embeddings: $\{X_{rgb}, X_{flow}, X_{depth}\}$.

These embeddings are fused via a cross-attention module:

$$Z_{rgb} = Attn(Q = X_{rgb}, K = \left[X_{flow}, X_{depth}\right], V = X_{flow}, X_{depth}),$$

Ensuring that anomaly-relevant cues from auxiliary modalities (e.g., abrupt motion from flow, structural anomalies from depth) are injected into the RGB stream.

A Transformer encoder then refines the fused representation to capture long-term multimodal interactions. This design achieves AUC = 97.9% on UCSD Ped2, 90.1% on Avenue, and 85.8% on ShanghaiTech, highlighting the robustness of multimodal fusion against modality noise and occlusion [25].

SwinIoT also exemplifies multimodal modeling in IoT surveillance, where video streams are augmented with contextual IoT data (e.g., environmental sensors, traffic metadata). By embedding these streams into a unified transformer space, SwinIoT achieves scalable behavioral anomaly detection across heterogeneous smart city infrastructures [23]. Table 1 presents a comparative analysis of various transformer-based models.

Table 1. Comparative analysis of transformer-based models

Model	Architecture Type	Input Modalities	Temporal Modeling	Strengths	Limitations
CILAR-Net [16]	Hybrid (ViT + GRU + Incremental Learning)	RGB frames	GRU for sequential learning	Adapts to new anomaly classes without retraining; robust lifelong learning	Replay buffer overhead; transformer compute cost
ViT-ARN [17]	Transformer + Multi-Reservoir ESN	RGB clips (≈16 frames)	ESN for efficient sequence dynamics	Combines ViT spatial power with efficient temporal ESN; lightweight	Sensitive to sequence boundaries; ESN tuning required
ST-HTAM [24]	Swin Transformer + Hierarchical Temporal Attention	RGB video	Hierarchical temporal attention + temporal adaptive module	Captures multi-scale temporal dependencies; works with weak labels	Heavy compute; weaker anomaly localization
Transformer Enc–Dec (Traffic) [21]	Transformer Encoder–Decoder	Aerial traffic video frames	Frame prediction via encoder–decoder	Works without anomaly labels; captures long-range dependencies	Lower accuracy; false alarms in dynamic traffic scenes
BiMT [11]	CNN + BiLSTM + Transformer	RGB frames	BiLSTM (bidirectional) + Transformer for	Strong local + sequential + global modeling	High training complexity and cost

TransCNN [12]	Hybrid CNN + Transformer	RGB frames	long-range Transformer encoder over CNN features	Balances CNN spatial and Transformer global reasoning	Needs large data; dual modules increase cost
TDS-Net [19]	Dual-Stream CNN + Transformer	RGB + Optical Flow	Transformer encoder over fused tokens	Strong for motion-driven anomalies; high reported AUC (91%)	Optical flow extraction adds latency
SwinIoT [23]	Swin Transformer (hierarchical, shifted window)	IoT data streams (video + sensors)	Temporal attention on Swin features	Scalable for smart cities; efficient windowed attention	Still heavy for edge devices; needs normal data
Multimodal Fusion + Transformer[13]	CNN encoders + Fusion + Transformer	RGB + Flow + (Audio/Depth/Metadata)	Transformer encoder over fused multimodal tokens	Robust to modality noise; strong cross-modal reasoning	High computational + synchronization cost
YOLOv8-based Night-Time Detection [26]	One-stage CNN-based object detector	RGB images (low-light/night-time)	No explicit temporal modeling	High detection accuracy for small and distant objects in low-light conditions; real-time performance; effective use of data augmentation and HPC resources	Limited to frame-level object detection; does not capture motion patterns or high-level behavioral anomalies
CNN–BiLSTM with Attention [27]	Hybrid CNN–RNN architecture with attention	RGB video frames	BiLSTM-based temporal modeling	Captures spatiotemporal dependencies for complex behavior detection using attention	Computationally intensive with limited interpretability and scalability
VidAnomalyNet [28]	Deep CNN-based anomaly detection network	RGB surveillance videos	Implicit (via stacked CNN layers)	High-accuracy, efficient CNN for multi-class event-level anomaly detection	Requires labeled data; limited generalization; no explicit semantic or temporal reasoning

4. PROPOSED UNIFIED TRANSFORMER-BASED VIDEO ANOMALY DETECTION FRAMEWORK

Transformer architectures have rapidly emerged as leading solutions for Video Anomaly Detection (VAD) due to their superior ability to model long-range spatiotemporal dependencies. However, existing studies often present isolated architectural innovations without a unified theoretical foundation spanning supervised, weakly supervised, and unsupervised paradigms. To address this gap, we propose a holistic and extensible framework that organizes transformer-based VAD into five tightly coupled components: multimodal input processing, spatial encoding, temporal modeling, anomaly scoring, and decision-support post-processing. This unified framework formalizes the operational stages common to state-of-the-art methods and provides a conceptual lens for evaluating and comparing existing transformer-based approaches.

4.1 Multimodal input acquisition and preprocessing

Surveillance environments generate heterogeneous and often asynchronous input streams, including RGB video, optical flow, depth imagery, thermal infrared data, audio cues, and IoT sensor metadata. Modern transformer-based VAD systems increasingly rely on multimodal fusion to overcome environmental constraints such as occlusion, low illumination, and noise.

Key Components

- Synchronized sampling and temporal alignment ensure uniform frame intervals, enabling consistent tokenization across modalities.
- Normalization and augmentation (e.g., brightness correction, motion jitter, geometric transformations) stabilize training under real-world [29].
- Tokenization/Patch extraction divides each frame into fixed-size patches (e.g., 16 × 16), encoding them as linear embeddings as implemented in Vision Transformer (ViT) architectures [30].
- Multimodal preprocessing reduces reliance on single-modality signals and provides the unified transformer pipeline with rich contextual cues necessary for robust anomaly identification.

4.2 Spatial encoder: Hierarchical visual abstraction

Spatial encoding represents the first layer of semantic abstraction in the framework. Transformers handle spatial information differently from convolutional approaches, emphasizing non-local interactions at the patch level.

Approaches within the Framework

- Vision Transformer (ViT): Frames are partitioned into patches and processed as tokens, enabling non-local spatial reasoning across the entire scene.
- Swin Transformer: Uses shifted window-based self-attention to compute multi-scale spatial features efficiently, improving performance in crowded or cluttered scenes [31].
- Hybrid CNN–Transformer Architectures: Earlier convolutional layers extract local textures (edges, gradients) while transformers captu re broader spatial relationships.
- Spatial encoders provide hierarchical semantic information essential for detecting appearance-based anomalies (e.g., unusual objects, unattended bags, structural abnormalities).

4.3 Temporal modeling: Global and local dynamics

Anomalies are typically defined by deviations in temporal patterns rather than isolated frame-level abnormalities. Modeling temporal evolution is therefore central to VAD.

Core Mechanisms

- Multi-Head Self-Attention (MHSA) allows modeling of long-term dependencies across hundreds of frames, overcoming the limitations of RNN-based models such as LSTMs.
- Temporal factorization (as in TimeSformer [32]) decomposes spatiotemporal attention into sequential spatial and temporal attention steps to improve efficiency.
- Recurrent modules (SRU++, GRUs, ESNs) can be integrated to maintain short-term memory stability and reduce the computational cost of full attention over long sequences.
- Cross-attention fusion aligns temporal signals across modalities, such as correlating optical flow trajectories with RGB appearance cues.
- Temporal modeling captures motion irregularities, crowd behavior anomalies, and evolving interactions—critical signals in anomaly detection.

4.4 Anomaly modeling and scoring mechanisms

The proposed framework accommodates multiple learning paradigms observed in VAD research.

Supervised Approaches

Use discriminative classification heads or sequence-level anomaly probability regressors. Examples include transformer-based action classifiers adapted for anomaly discrimination.

Weakly Supervised Approaches

Leverage Multiple Instance Learning (MIL) frameworks where only video-level labels are available. Temporal attention highlights anomalous segments.

Self-Supervised Approaches

Employ predictive or reconstructive tasks such as:

- masked video modeling
- contrastive temporal representation learning.

Unsupervised Approaches

Use reconstruction or prediction errors as anomaly scores:

- video prediction transformers
- spatiotemporal reconstruction-based transformers.

Scoring Mechanisms

- Residual deviation analysis
- Feature-space distance metrics (e.g., Mahalanobis distance)
- Reconstruction/prediction error curves
- Attention deviation metrics (anomalies cause atypical attention patterns)

Different surveillance contexts demand different supervision regimes; thus, the framework supports all major paradigm families.

4.5 Decision support and post-processing layer

The decision-support layer translates model outputs into interpretable, actionable insights for operators and automated systems.

Core Outputs

- **Frame-level anomaly probability curves** for temporal localization.
- **Attention heatmaps** highlighting suspicious objects or regions.
- **Spatiotemporal saliency maps** to support human decision-makers.
- **Tracking-enhanced anomaly refinement**, linking anomalies to object trajectories or identities.

Deployment-Oriented Enhancements

- Real-time inference optimizations for edge devices [7].
- Automatic threshold tuning based on scene dynamics.
- Integration with IoT event logs for contextual anomaly interpretation.
- This layer ensures usability by bridging algorithmic outputs with operational surveillance needs.

5. RESULT ANALYSIS

The comparative analysis reveals that TDS-Net (91%), TransCNN (90%), and Multimodal Fusion + Transformer (90%) achieve the strongest performance among the surveyed models, highlighting the effectiveness of combining CNN-based local feature extraction, motion-aware modeling, and transformer-driven global reasoning. BiMT (89%), SwinIoT (89%), ViT-ARN (88%), and CILAR-Net (87%) demonstrate competitive mid-range results, reflecting the benefits of hybrid architectures and incremental or multimodal strategies. In contrast, ST-HTAM (84%) and the Transformer Encoder–Decoder for Traffic Anomaly Detection (82%) report comparatively lower scores, largely due to weak supervision settings and unsupervised prediction-based learning, respectively. These findings suggest that fusion-based and dual-stream hybrid approaches generally yield superior anomaly detection accuracy across surveillance benchmarks.

Table 2 summarizes the experimental results obtained from the evaluation of the different models.

Table 2. Result analysis

Model	Dataset(s)	Reported Performance	Key Observation
CILAR-Net [16]	UCSD Ped2, Avenue, ShanghaiTech	97.2 (UCSD), 88.4 (Avenue), 83.6 (ShanghaiTech)	Excels on simple datasets (UCSD), but performance drops on complex, crowded scenarios (ShanghaiTech). Supports class-incremental learning for real-time adaptability.
ViT-ARN [17]	ShanghaiTech, UCF-Crime	74.8(ShanghaiTech), 88.1 (UCF-Crime)	Integrates ViT with Echo State Networks; efficient for long-sequence modeling, but struggles in dense crowds (ShanghaiTech). Performs well in long-duration UCF-Crime.
ST-HTAM [24]	UCSD Ped2, Avenue, ShanghaiTech	96.3 (UCSD), 86.7 (Avenue), 79.2 (ShanghaiTech)	Weakly supervised MIL-based Swin Transformer; robust for weak labels but lower performance in complex datasets.
Transformer Enc–Dec (Traffic) [21]	Traffic Aerial Dataset	82.0	Fully unsupervised prediction model; avoids labeling costs but underperforms due to reconstruction noise in dynamic

			environments.
BiMT (CNN-BiLSTM-Transformer) [11]	UCSD Ped2, Avenue, ShanghaiTech	97.8 (UCSD), 89.2 (Avenue), 84.6 (ShanghaiTech)	Strong sequential + global modeling; balances spatial, temporal, and attention features. Slightly lower performance on complex datasets.
TransCNN [12]	UCSD Ped2, Avenue, ShanghaiTech	98.1 (UCSD), 90.3 (Avenue), 85.4 (ShanghaiTech)	Hybrid CNN + Transformer; consistently high performance across datasets. Shows strong generalization.
TDS-Net [19]	UCSD Ped2, Avenue, ShanghaiTech	98.5 (UCSD), 91.0 (Avenue), 86.1 (ShanghaiTech)	Dual-stream (RGB + Optical Flow) with Transformer; best overall performance. Trade-off: requires costly optical flow computation.
SwinIoT [23]	IoT Smart-City Dataset	89.0	Hierarchical Swin Transformer; scalable for IoT surveillance but computationally heavy for edge deployment.
Multimodal Fusion + Transformer [13]	UCSD Ped2, Avenue, ShanghaiTech	97.9 (UCSD), 90.1 (Avenue), 85.8 (ShanghaiTech)	Combines RGB, Flow, Depth, and Audio modalities; robust to modality noise and strong generalization, but requires multiple synchronized inputs.

6. APPLICATION DOMAIN AND BEST MODEL

1. CCTV / Public Surveillance → BiMT, ViT+SRU++

BiMT (CNN-BiLSTM-Transformer) achieved the highest accuracy (98.6% AUC) on UCF-Crime and performed robustly on UBI-Fight and RAD datasets. Its hybrid architecture combining CNNs for spatial features, BiLSTMs for temporal patterns, and transformers for long-range dependencies makes it ideal for urban surveillance systems.

ViT+SRU++ leverages a modified Vision Transformer with SRU++ recurrent modules, delivering near SOTA accuracy (97% UCF-Crime, 96% RWF-2000) while being 10× faster than traditional RNNs, making it well-suited for real-time CCTV feeds.

2. Multimodal CCTV (RGB + IR + Depth) → Multimodal Fusion + Attention

This model integrates RGB, infrared, and depth video streams using multimodal autoencoders and attention-based fusion. With 95.1% accuracy and reduced false positives, it is particularly effective under low-light or crowded conditions, where unimodal methods often fail.

3. Real-Time Long Video Monitoring → ViT+SRU++, TransCNN

ViT+SRU++ provides efficient long-sequence modeling with low latency, supporting real-time anomaly detection.

TransCNN, a hybrid CNN–Transformer framework, achieved 94.6% (ShanghaiTech), 98.4% (UCSD Ped2), and 89.6% (CUHK Avenue) AUC, proving its strength on long-duration dataset.

4. IoT-Driven Smart Cities → SwinIoT

SwinIoT adapts the Swin Transformer for IoT environments. It uses hierarchical attention windows and lightweight design, optimized for edge-computing and low-resource deployments. It reached 96% accuracy and 97% mAP across diverse IoT-driven datasets, making it suitable for smart city surveillance.

5. Motion-Sensitive Detection (e.g., fights)

TDS-Net employs a dual-stream architecture (RGB + optical flow) with a transformer for temporal fusion. This enables robust motion-aware anomaly detection, outperforming baselines on ShanghaiTech and CUHK Avenue datasets.

6. Aerial / Drone Surveillance → Unsupervised Transformer

This unsupervised framework predicts future frames using a transformer encoder-decoder trained solely on normal traffic videos. Anomalies are flagged by high reconstruction errors. It achieved state-of-the-art results on Drone-Anomaly and UIT-ADrone datasets, making it highly suitable for aerial surveillance.

7. Evolving Anomalies (New Classes) -CILAR-Net

This introduces class-incremental learning for anomaly recognition. It adapts to new anomaly classes without retraining, preventing catastrophic forgetting. Tested on UCF-Crime, RWF-2000, LAD-2000, it outperformed existing baselines (e.g., +9.7% on LAD-2000).

8. Low Annotation Cost (Weak Supervision) → ST-HTAM

ST-HTAM combines the Swin Transformer with a Hybrid Temporal Adaptive Module (HTAM: global self-attention + Conv-LSTM) for weakly supervised video anomaly detection. It uses only video-level labels, reducing annotation costs while outperforming prior weakly supervised methods.

9. Real-Time Anomaly Tracking → SwinAnomaly

SwinAnomaly integrates a Swin Transformer-based autoencoder, GAN-based prediction, YOLOv7 object detection, and SORT tracking. This allows real-time anomaly detection with localization, outperforming existing prediction-based methods on standard CCTV datasets.

As illustrated in Table 3, the Transformer-based models are applied across various domains, highlighting their versatility and performance differences.

Table 3. Analysis according to application domain

Application Domain	Best Model(s)	Reason
CCTV / Public Surveillance	BiMT (CNN+BiLSTM+Transformer) [11], ViT+SRU++	High accuracy on UCF-Crime, robust for city surveillance, efficient temporal modeling
Multimodal CCTV (RGB + IR + Depth)	Multimodal Fusion + Attention [13]	Combines RGB, IR, Depth → strong in low-light, crowded, and complex environments
Real-Time Long Video Monitoring	ViT+SRU++, TransCNN [12]	Handles long sequences efficiently, low-latency, accurate spatio-temporal modeling

IoT-Driven Smart Cities	SwinIoT (Hierarchical Transformer) [23]	Edge-optimized, scalable, robust for smart city IoT sensor + video data
Motion-Sensitive Detection (fights)	TDS-Net (Dual-Stream Transformer) [19]	Fuses RGB + optical flow for strong motion/appearance anomaly detection
Aerial / Drone Surveillance	Transformer (Unsupervised Traffic) [21]	Learns only from normal drone traffic, detects anomalies via prediction errors
Evolving Anomalies (New Classes)	CILAR-Net (Class-Incremental Learning) [16]	Adapts to new anomaly types without retraining, avoids catastrophic forgetting
Low Annotation Cost (Weak Supervised)	ST-HTAM (Weakly Supervised Transformer) [24]	Requires only video-level labels, reduces false alarms, strong temporal modeling
Real-Time Anomaly Tracking	SwinAnomaly (Video Swin + SORT + GAN) [33]	Combines frame prediction + object tracking for anomaly localization in real time

7. CONCLUSION AND FUTURE DIRECTION

This comparative study clarifies pros and cons of state-of-the-art transformer-based models to identify anomalies in video within diverse application scenarios, including CCTV monitoring, aerial observation, smart city monitoring, and multimodal scenarios. The evaluation discloses that not a single solution attains unquestionable dominance; rather, each model best performs within specific restraints and demands of application. For instance, supervised methods such as BiMT, TDS-Net, and TransCNN achieve peak accuracy on benchmark datasets such as UCF-Crime, ShanghaiTech, and UCSD Ped2, making them particularly optimal within scenarios requiring peak accuracy. Conversely, weakly supervised methods such as ST-HTAM effectively reduce annotation costs without compromising comparable performance, therefore offering a realistic tradeoff between effectiveness and accuracy. At the same time, self-supervised and unsupervised frameworks, including Unsupervised Transformers and SwinAnomaly, hold significant potential within anomaly detection of aerial traffic and real-time localization of irregularities without requiring heavy labels. Besides, heterogeneous architecture fusing CNNs/RNNs and Transformers, such as ViT+ SRU++, show stability suited to long-term video observation, while hierarchical or multimodal transformers, including SwinIoT and Multimodal Fusion with Attention, extend their applicability to IoT-driven or low-visibility environments. Importantly, incremental learning methods such as CILAR-Net mark steps toward adaptive learning of anomalies by accommodating the incremental learning of novel classes of data without falling prey to catastrophic forgetting.

Although such advances have been made, some of the challenges that require investigation now are listed below. First, lightweight and energy-efficient structures are now needed because many of the models made to date are computationally prohibitive and not suited to large-scale, real-time, or edge-based deployments. Moreover, the issue of cross-domain generalization is limited because most frameworks are fine-tuned using limited datasets; future research should consider transfer learning and domain adaptation to enhance robustness across various scenarios of surveillance. Additionally, while incremental learning models handle such novelty of new anomaly classes, research is needed to enhance stability, accuracy, and preventing catastrophic forgetting. One of the potential research directions is over-reliance on large annotated datasets by employing self-supervised, contrastive, and active learning schemes and thus synchronizing supervised and weakly supervised learning. Another area of research is that of explainability of transformer-based models because interpretable outputs will help to foster trust and facilitate deployment with safety repercussions. Further, extension of anomaly detection to accommodate modalities other than visual inputs such as infrared, depth, audio, sensor, and metadata streams can potentially maximize detection accuracy and contextual understanding. Lastly, closer integration of anomaly detection with object tracking, behavioral forecasting, and early alert schemes can provide pro-active intervention instead of only reactive monitoring.

Third, future video anomaly detection will be propelled by lightweight, flexible, interpretable, and multimodal models of transformer architecture that can run effectively within dynamic, complex, and real-time environments. The overview thus makes a crucial contribution by offering a systematic decision-making process that can be adopted by both research workers and practitioners to choose models that best align with their respective priorities of their video-based surveillance applications e.g., accuracy, temporal efficiency, cost of annotation.

REFERENCES

[1] Nayak, R., Pati, U.C., Das, S.K. (2021). A comprehensive review on deep learning-based methods for video anomaly detection. Image and Vision Computing, 106: 104078. https://doi.org/10.1016/j.imavis.2020.104078

[2] Zhou, F.R., Wen, G., Ma, Y., Geng, H., Huang, R., Pei, L., Yu, W.X., Chu, L., Qiu, R. (2022). A comprehensive survey for deep-learning-based abnormality detection in smart grids with multimodal image data. Applied Sciences, 12(11): 5336. https://doi.org/10.3390/app12115336

[3] Wang, Z.J., Jiang, K.K., Hou, Y.S., Dou, W.W., Zhang, C.M., Huang, Z.H. (2019). A survey on human behavior recognition using channel state information. IEEE Access, 7: 155986-156024. https://doi.org/10.1109/ACCESS.2019.2949123

[4] Jebur, S.A., Hussein, K.A., Hoomod, H.K., Alzubaidi, L., Santamaría, J. (2023). Review on deep learning approaches for anomaly event detection in video surveillance. Electronics, 12(1): 29. https://doi.org/10.3390/electronics12010029

[5] Duong, H.T., Le, V.T., Hoang, V.T. (2023). Deep learning-based anomaly detection in video surveillance: A survey. Sensors, 23(11): 5024. https://doi.org/10.3390/s23115024

[6] Wastupranata, L.M., Kong, S.G., Wang, L. (2024). Deep learning for abnormal human behavior detection in surveillance videos—A survey. Electronics, 13(13):

2579. https://doi.org/10.3390/electronics13132579

[7] Berroukham, A., Housni, K., Lahraichi, M., Boulfrifi, I. (2023). Deep learning-based methods for anomaly detection in video surveillance: A review. Bulletin of Electrical Engineering and Informatics, 12(1): 314-327. https://doi.org/10.11591/eei.v12i1.3944

[8] Karbalaie, A., Abtahi, F., Sjöström, M. (2022). Event detection in surveillance videos: A review. Multimedia Tools and Applications, 81: 35463-35501. https://doi.org/10.1007/s11042-021-11864-2

[9] Yu, J., Lee, Y., Yow, K.C., Jeon, M., Pedrycz, W. (2022). Abnormal event detection and localization via adversarial event prediction. IEEE Transactions on Neural Networks and Learning Systems, 33(8): 3572-3586. https://doi.org/10.1109/TNNLS.2021.3053563

[10] Sengonul, E., Samet, R., Al-Haija, Q.A., Alqahtani, A., Alsemmeari, R.A., Alghamdi, B., Alturki, B., Alsulami, A.A. (2025). Abnormal event detection in surveillance videos through LSTM auto-encoding and local minima assistance. Discover Internet of Things, 5(1): 32. https://doi.org/10.1007/s43926-025-00127-3

[11] Natha, S., Siraj, M., Ahmed, F., Altamimi, M., Syed, M. (2025). An integrated CNN-BiLSTM-transformer framework for improved anomaly detection using surveillance videos. IEEE Access, 13: 95341-95357. https://doi.org/10.1109/ACCESS.2025.3574835

[12] Ullah, W., Hussain, T., Ullah, F.U.M., Lee, M.Y., Baik, S.W. (2023). TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection. Engineering Applications of Artificial Intelligence, 123: 106173. https://doi.org/10.1016/j.engappai.2023.106173

[13] Srilakshmi, V., Veesam, S.B., Krishna, M.S.R., Munaganuri, R.K., Sivaprasad, D.D. (2025). Design of an improved model for anomaly detection in CCTV systems using multimodal fusion and attention-based networks. IEEE Access, 13: 27287-27309. https://doi.org/10.1109/ACCESS.2025.3536501

[14] Joshi, K., Patel, N. (2025). Supervised deep learning approaches for anomaly detection and recognition in crowd scenes. ELCVIA, 24(1): 31-50. https://doi.org/10.5565/rev/elcvia.1631

[15] Muna, U.M., Biswas, S., Zarif, S.A.A.M., Deori, P.J., Tajwar, T., Shatabda, S. (2025). Vision transformer embedded video anomaly detection using attention driven recurrence. Array, 27: 100471. https://doi.org/10.1016/j.array.2025.100471

[16] Hussain, A., Ullah, W., Khan, N., Khan, Z.A., Yar, H., Baik, S.W. (2026). Class-incremental learning network for real-time anomaly recognition in surveillance environments. Pattern Recognition, 170: 112064. https://doi.org/10.1016/j.patcog.2025.112064

[17] Ullah, W., Hussain, T., Baik, S.W. (2023). Vision transformer attention with multi-reservoir echo state network for anomaly recognition. Information Processing & Management, 60(3): 103289. https://doi.org/10.1016/j.ipm.2023.103289

[18] Zhang, D.S., Huang, C., Liu, C.L., Xu, Y. (2022). Weakly supervised video anomaly detection via transformer-enabled temporal relation learning. IEEE Signal Processing Letters, 29: 1197-1201. https://doi.org/10.1109/LSP.2022.3175092

[19] Hussain, A., Ullah, W., Khan, N., Khan, Z.A., Kim, M.J., Baik, S.W. (2024). TDS-Net: Transformer enhanced dual-stream network for video Anomaly Detection. Expert Systems with Applications, 256: 124846. https://doi.org/10.1016/j.eswa.2024.124846

[20] Dewi, D.A., Singh, H.K.R., Periasamy, J., Kurniawan, T.B., Henderi, H., Hasibuan, M.S., Nathan, Y. (2025). Incorporate transformer-based models for anomaly detection. Journal of Applied Data Sciences, 6(3): 2046-2055. https://doi.org/10.47738/jads.v6i3.762

[21] Tran, T.M., Bui, D.C., Nguyen, T.V., Nguyen, K. (2024). Transformer-based spatio-temporal unsupervised traffic anomaly detection in aerial videos. IEEE Transactions on Circuits and Systems for Video Technology, 34(9): 8292-8309. https://doi.org/10.1109/TCSVT.2024.3376399

[22] Huang, C., Wen, J., Xu, Y., Jiang, Q.P., Yang, J., Wang, Y.W. (2023). Self-supervised attentive generative adversarial networks for video anomaly detection. IEEE Transactions on Neural Networks and Learning Systems, 34(11): 9389-9403. https://doi.org/10.1109/TNNLS.2022.3159538

[23] Mancy, H., Naith, Q.H. (2025). SwinIoT: A hierarchical transformer-based framework for behavioral anomaly detection in IoT-driven smart cities. IEEE Access, 13: 48758-48774. https://doi.org/10.1109/access.2025.3551207

[24] Paulraj, S., Vairavasundaram, S. (2025). Transformer-enabled weakly supervised abnormal event detection in intelligent video surveillance systems. Engineering Applications of Artificial Intelligence, 139: 109496. https://doi.org/10.1016/j.engappai.2024.109496

[25] Ding, H.T., Lou, S.F., Ye, H.R., Chen, Y.B. (2025). MT-CMVAD: A multi-modal transformer framework for cross-modal video anomaly detection. Applied Sciences, 15(12): 6773. https://doi.org/10.3390/app15126773

[26] Namana, M.S.K., Kumar, B.U. (2024). An efficient and robust night-time surveillance object detection system using YOLOv8 and high-performance computing. International Journal of Safety and Security Engineering, 14(6): 1763-1773. https://doi.org/10.18280/ijsse.140611

[27] Pangavhane, M., Patil, R., Bharati, R., Gupta, D., Ahire, P., Patil, P., Rahane, W., Dharrao, D. (2025). Real-time deep learning-driven surveillance with spatiotemporal feature extraction for detection of anomalous human behavior across dynamic environments. International Journal of Safety and Security Engineering, 15(1): 105-111. https://doi.org/10.18280/ijsse.150112

[28] Chidananda, K., Siva Kumar, A.P. (2024). VidAnomalyNet: An efficient anomaly detection in public surveillance videos through deep learning architectures. International Journal of Safety and Security Engineering, 14(3): 953-966. https://doi.org/10.18280/ijsse.140326

[29] Zhou, K.Y., Yang, J.K., Loy, C.C., Liu, Z.W. (2022). Learning to prompt for vision-language models. International Journal of Computer Vision, 130: 2337-2348. https://doi.org/10.1007/s11263-022-01653-1

[30] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C. (2021). ViViT: A video vision transformer. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 6816-6826. https://doi.org/10.1109/ICCV48922.2021.00676

[31] Liu, Z., Lin, Y.T., Cao, Y., Hu, H., Wei, Y.X., Zhang, Z. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In 2021 IEEE/CVF

International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 9992-10002. https://doi.org/10.1109/ICCV48922.2021.00986

[32] Li, Y.W., Zhang, Y.L., Timofte, R., Van Gool, L., Yu, L., Li, Y.W. (2023). NTIRE 2023 challenge on efficient super-resolution: Methods and results. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, pp. 1922-1960. https://doi.org/10.1109/CVPRW59228.2023.00189

[33] Bajgoti, A., Gupta, R., Balaji, P., Dwivedi, R., Siwach, M., Gupta, D. (2023). SwinAnomaly: Real-time video anomaly detection using video swin transformer and SORT. IEEE Access, 11: 111093-111105. https://doi.org/10.1109/ACCESS.2023.3321801

NOMENCLATURE

VAD	Video Anomaly Detection
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory Network
BiLSTM	Bidirectional Long Short-Term Memory
ViT	Vision Transformer
SRU	Simple Recurrent Unit
BiMT	CNN–BiLSTM–Transformer Hybrid Model
TDS-Net	Transformer-enhanced Dual-Stream Network
TransCNN	Hybrid CNN–Transformer Mechanism
ST-HTAM	Spatio-Temporal Hierarchical Transformer Attention Model
IoT	Internet of Things
MIL	Multiple Instance Learning
AUC	Area Under the Curve (Evaluation Metric)
FPS	Frames Per Second
GCN	Graph Convolutional Network
GAN	Generative Adversarial Network

Greek symbols

α	Attention weight coefficient in self-attention mechanism
β	Temporal decay factor or weighting term
γ	Normalization scaling parameter
λ	Regularization or fusion weighting coefficient
θ	Learnable model parameters (trainable weights)
η	Learning rate for optimization
σ(·)	Activation function (sigmoid, softmax, tanh)
ρ	Correlation coefficient or temporal relation weight
μ	Mean value (for normalization or reconstruction baseline)
Σ	Covariance matrix in probabilistic embedding space
δ(·)	Indicator function for anomaly event detection

Subscripts

ti	RGB visual modality
flow	Optical flow (motion modality)
depth	Depth modality
ir	Infrared or thermal stream
t	Temporal frame index
i, j	Token, patch, or pixel indices
l	Hierarchical level or Transformer layer index
enc, dec	Encoder and decoder modules in Transformer architecture
res	Residual connection or reservoir module
pre	Self-supervised pretext objective
cls	Classification or decision head output
f, b	Forward and backward passes in BiLSTM layers
m	Modality index (for multimodal fusion)
rgb	RGB visual modality
flow	Optical flow (motion modality)
depth	Depth modality
ir	Infrared or thermal stream