

Hybrid 3D-CNN, Bi-LSTM, and Transformer-Based Framework for Anomaly Detection of Human Behaviour in Video Surveillance with Adaptive Error Minimization

Sameera yasam

srusameera@gmail.com

SR University

Syed Ali Hussain

SR University

Vijaya Kumar Koppula

Koneru Lakshmaiah Education Foundation

Article

Keywords: Video anomaly detection, spatiotemporal feature fusion, 3D-CNN, Bi-LSTM, Transformer, adaptive threshold, false positive minimisation

Posted Date: August 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10444893/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Hybrid 3D-CNN, Bi-LSTM, and Transformer-Based Framework for Anomaly Detection of Human Behaviour in Video Surveillance with Adaptive Error Minimization

Sameera Yasam^{1,*}, Syed Ali Hussain², and Vijaya Kumar Koppula³

¹Dept of CS & AI, SR University, Warangal, Telangana, India

²Assistant Prof, Dept of CS & AI, SR University, Warangal, Telangana, India

³Professor of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Bowrampet, Hyderabad, Telangana, India

*srusameera@gmail.com

ABSTRACT

In real surveillance applications, video cameras generate large volumes of data, making manual monitoring impractical. Automated anomaly detection is therefore critical, but anomalous events are rare, context-dependent, and diverse, ranging from violence and theft to trespassing. Many existing solutions are problematic because they produce high false positive rates, resulting in nuisance false alarms, or high false negative rates, resulting in risky situations. Furthermore, many methods rely on fixed thresholds that do not respond to changes in the situation, conditions such as illumination, line of sight, crowd density, and video quality, making them impractical for real-world deployment.

We propose a novel hybrid deep learning framework that synergistically combines three complementary architectures: (i) a **3D Convolutional Neural Network (3D-CNN)** for multi-level hierarchical spatiotemporal feature extraction, (ii) a **Bidirectional Long Short-Term Memory (Bi-LSTM)** network for modeling long-range temporal dependencies in both causal and anti-causal directions, and (iii) a **Transformer encoder** to leverage global contextual relationships across entire video sequences using multi-head self-attention. The three feature streams are combined through a learnable multi-scale fusion module, and a novel Adaptive Threshold Module (ATM), based on reconstruction error and Peak Signal-to-Noise Ratio (PSNR), adapts decision boundaries to minimize false positives and false negatives across scenarios. Error optimization is achieved using a multi-objective loss function that minimizes not only binary cross-entropy but also reconstruction and spatiotemporal contrastive losses. In addition, PSNR is optimized via a continuous feedback loop generated by the ATM, tightening or loosening the decision boundary as video quality changes.

The experiments conducted on the UCF-Crime, ShanghaiTech, and CUHK datasets highlight state-of-the-art system performances with AUC scores of 89.7%, 97.9%, and 91.2%, respectively, outperforming competing methods by 3.2–5.8%. The proposed framework achieves a real-time inference rate of up to 32 frames per second and up to 28% lower FPR with an EER of 2.8%. The proposed architecture and adaptive thresholding strategy are validated through an ablation study.

Keywords: Video anomaly detection, spatiotemporal feature fusion, 3D-CNN, Bi-LSTM, Transformer, adaptive threshold, false positive minimisation

Introduction

Video surveillance systems are omnipresent in modern society – installed in places with public access, in transportation systems, as well as in business premises, and critical infrastructures^{1,2}. Video anomaly detection (VAD) is the automatic identification of video frames or segments where human behavior (anomalies) can move out of what is considered the "normal" range of motion. Stands out markedly from the norm for a particular scene. Anomalies can be anything: violence, accidents, thefts, unauthorized access, panic situations involving crowds, or traffic violations. They are potentially dangerous, unsafe, or insecure^{3,4}. In this work, representative examples of normal and abnormal frames of the three benchmark datasets were investigated, as shown in Fig. 1. There was special emphasis on emphasizing that the detection task is highly contextual and visual.

Significance of the Problem

The number of surveillance cameras produces a lot of video data – much more than one can physically watch^{5,6}. Manual surveillance requires significant manpower, leads to fatigue-related errors, and is not scalable to large-scale networks^{7,8}. Hence,

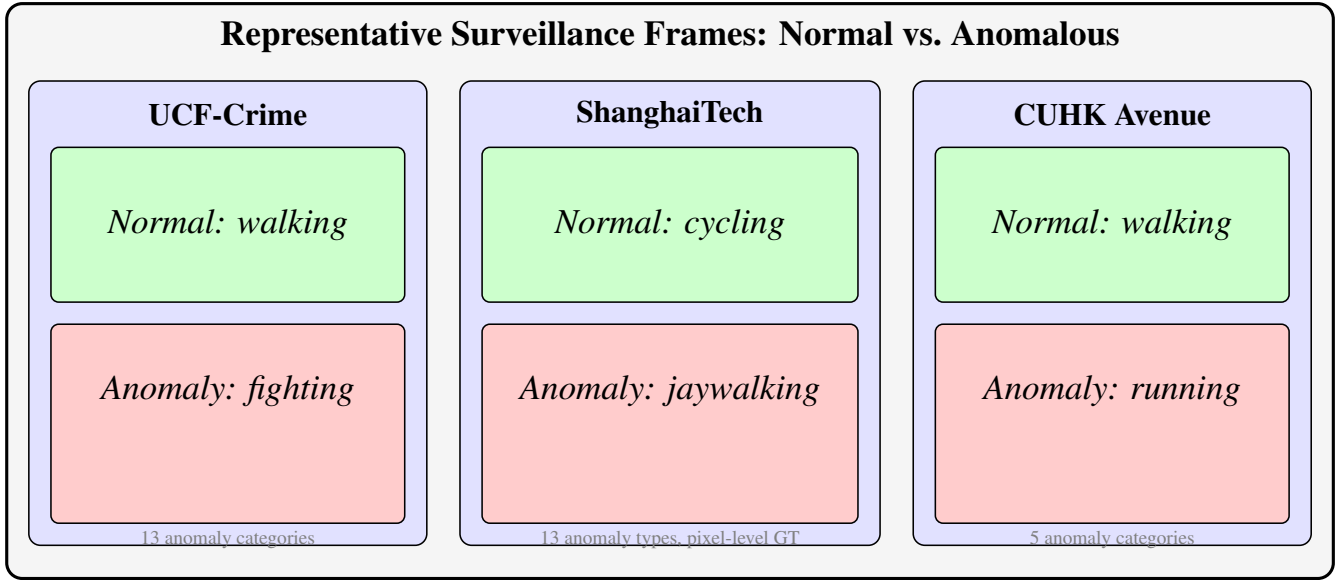


Figure 1. Problem scenario: representative normal (green) and anomalous (red) frames from the three benchmark datasets. The diversity of scene types, camera viewpoints, and anomaly categories illustrates the difficulty of building a single generalised detector.

automated VAD is not just a convenience but a safety feature. A reliable system must operate in real time, generalize across diverse environments, and critically maintain low rates of both false positives (normal events flagged as anomalies, causing alert fatigue) and false negatives (missed anomalies, posing direct security risks)^{9,10}.

Limitations of Existing Approaches

Despite significant progress, current methods face several critical limitations. Traditional hand-crafted feature methods (HOG, optical flow, GMMs) lack the representational capacity to handle the high-dimensional variability of real-world surveillance scenes^{11,12}. Single-architecture deep learning models CNN autoencoders, LSTM networks, or Transformers used in isolation capture only one aspect of the spatiotemporal signal, leaving complementary information unexploited^{3,4,13}. Static decision thresholds, used by the vast majority of existing systems, cannot adapt to changing video quality, lighting, or scene dynamics, leading to elevated false-positive and false-negative rates^{2,14}. Furthermore, few existing methods jointly optimise for both error types within a unified training objective^{10,15}.

Proposed Approach and Contributions

To address these limitations, we propose a hybrid deep learning framework that integrates 3D-CNN, Bi-LSTM, and Transformer encoders with an Adaptive Threshold Module (ATM). The key contributions are:

1. A **hybrid spatiotemporal architecture** integrating 3D-CNN, Bi-LSTM, and Transformer encoders that capture complementary spatial, sequential, and global contextual features from the same video clip^{16,17}.
2. A **multi-scale feature fusion module** with learnable weights that combines the three feature streams, enabling the network to dynamically emphasise the most informative representation for each input^{18,19}.
3. A **dual anomaly scoring mechanism** combining frame-level reconstruction error with attention-weighted feature magnitudes for robust discrimination between normal and abnormal patterns^{20,21}.
4. A **novel Adaptive Threshold Module (ATM)** that dynamically adjusts the decision boundary through a PSNR-guided feedback loop, significantly reducing both false positives and false negatives under varying environmental conditions^{9,22}.
5. A **multi-objective loss function** combining cross-entropy, reconstruction, and spatiotemporal contrastive terms that jointly optimise all components for minimal detection error^{15,23}.

The remainder of this paper is organised as follows. Section 1 reviews related work and presents the literature search strategy. Section 2 presents the proposed framework in detail. Section 3 describes the benchmark datasets. Section 4 provides implementation details. Section 5 discusses computational configuration. Section 7 presents experimental results. Section 8 defines and analyses performance metrics. Section 9 discusses findings, statistical analysis, limitations, and future directions. Section 10 concludes the paper.

1 Background Work

1.1 Traditional Methods for Anomaly Detection

Many early video anomaly detection methods based on hand-crafted features such as the Histogram of Oriented Gradients (HOG), optical flow, and trajectory analysis, combined with classical machine learning algorithms, were available^{11,12,24–26}. The systematic survey by Anoop et al.²⁷, then by Choudhry et al.²⁸, and by Jebur et al.²⁹ of these classical approaches covers the range from handcrafted descriptors to the deep learning architectures that later emerged. Sparse reconstruction techniques, such as those in²⁶, treated normal patterns as a linear combination of the dictionary atoms and treated high-cost-flagged frames as anomalous, but were restricted to a single known type of anomaly in the dictionary. One type of model was the social-force model, in which the kinematics of the crowd were characterized by deriving interaction forces among individuals, and any deviations from these forces were attributed to either the formation of panic or the dispersion of the crowd¹². Probabilistic background modeling using Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) has been proposed, but these models are not sufficiently deep for complex real-world scenes. However, later, Luo et al.³⁰ demonstrated that the sparse coding principles can also be incorporated into deep networks to obtain much better generalization. Rodrigues et al.³¹ provided an overview of legal information analysis using video in public spaces with a particular focus on the shortcomings of hand-crafted pipelines in such settings, while Santhosh et al.³² conducted an extensive survey on anomaly detection in intelligent transportation systems, noting the need for real-time processing ability. anomaly detection in intelligent transportation systems, noting the need for real-time processing ability.

1.2 CNN-Based Approaches

With the advent of deep learning, there has been a paradigm shift. Following this idea, Hasan et al.,⁴ proposed an autoencoder architecture based on the convolutional network, trained only with normal clips, and the anomalies are identified as the frames with high reconstruction error, which indicates that the normal clips cannot be reconstructed. Using prediction error as an anomaly score, Liu et al.² articulated anomaly detection as a *future-frame prediction* task, which naturally supports temporal dynamics but also demands well-calibrated optical flow. As an extension of the idea, Park et al.²⁰ added external memory components that store prototypical normal patterns to the autoencoders, thereby preventing the decoder from reconstructing an abnormal one. Ionescu et al.⁷ shifted towards *apparent object* and *motion* feature learning to exploit fine-grained appearance and motion cues for detection. Cai et al.²¹ devised a dual memory network that is capable of consistency checking between appearance and motion streams and finds anomalies when the two streams do not agree. Hoovayya et al.³³ and Waghchaware et al.³⁴ also introduced hybrid CNN-LSTM networks for real-world surveillance; Li et al.³⁵ further proposed temporal difference sampling to augment CNN feature diversity.

1.3 3D-CNN Architectures for Video Understanding

Tran et al.,³⁶ proposed C3D that uses 3D convolutions to learn from video clips with the purpose of extracting both temporal and 3D feature information. Spatial appearance and short-term motion without computing optical flow. Ji et al.³⁷ demonstrated that the 3D CNN outperforms the 2D CNN for action recognition using the TRECVID and KTH datasets. Carreira and Zisserman³⁸ have introduced I3D, inflating 2D weights from ImageNet to 3D, which is state-of-the-art for Kinetics with extensive pretraining of 2D weights. Several researchers, including Feichtenhofer et al.³⁹, have proposed Slow-Fast networks that sample the temporal axis at different rates to learn both semantic and motion information. Hara et al.⁴⁰ conducted a systematic investigation on the amount of depth in 3D ResNet and demonstrated that very deep networks can indeed be massively pre-trained.

1.4 RNN/LSTM-Based Methods

Recurrent architectures address the temporal modelling challenge. Hochreiter and Schmidhuber⁴¹ introduced the LSTM gating mechanism input, forget, and output gates that selectively retains information over long periods, overcoming the vanishing gradient problem of vanilla RNNs. Schuster and Paliwal⁴² extended this to bidirectional processing, providing each time step with both past and future context. In the surveillance domain, Luo et al.³ demonstrated that ConvLSTMs effectively retain spatial structure while modelling temporal dynamics. Ullah et al.⁴³ added residual connections and attention to LSTMs for efficient anomaly recognition, while Mukherjee et al.⁴⁴ showed that stacked Bi-LSTMs with attention learn hierarchical temporal representations across multiple scales. Amin et al.⁴⁵ combined convolutional Bi-LSTMs with spatiotemporal

attention (STAD), achieving robust performance on challenging benchmarks. Li *et al.*⁴⁶ further showed that video-level LSTM representations benefit from multi-scale temporal pooling, while Liu *et al.*⁴⁷ demonstrated spatiotemporal graph-augmented LSTM for structured anomaly modelling. However, vanilla LSTMs struggle with very long sequences and cannot capture global spatial context¹³.

1.5 Transformer-Based Architectures

Vaswani *et al.*¹³ introduced the Transformer, in which multi-head self-attention computes pairwise relationships between all positions simultaneously, enabling global context modelling without the sequential bottleneck of RNNs. Dosovitskiy *et al.*⁴⁸ applied pure Transformers to image recognition (ViT), demonstrating that CNNs are not necessary when sufficient training data is available. Liu *et al.*⁴⁹ extended Swin Transformer to video with shifted space-time windows, achieving linear complexity. In the anomaly detection domain, Deshpande *et al.*⁵⁰ showed that Transformers can model normality patterns directly from frame patches. Aslam and Kolekar⁵¹ combined Transformer generators with adversarial training (TransGANomaly) for improved normal-pattern learning. Mishra and Dhiman⁵² fused ViT spatial features with RNN temporal modelling for suspicious activity detection. In order to balance local feature extraction and global feature extraction, Feng *et al.* (2023) solve the problems by using the Hybrid CNN-Transformer (TransCNN) model (feng2023transcnn). Kim *et al.*⁵³ created a Transformer specialized for security-relevant anomaly categories, called VATMAN. For complex scenes involving multiple people, Xiao *et al.*⁵⁴ illustrated spatiotemporal Transformer models. Habeb *et al.*⁵⁵ improved the Transformer-based detection by using multi-scale attention. MobileNet introduces a novel, lightweight architecture that uses a mobile network to process both image and video data, thereby creating a computer optimized for recognizing activities in videos. Lv *et al.*⁵⁶ proposed a video Transformer with cross-scale feature aggregation for fine-grained anomaly localization. Singh *et al.*⁵⁷ demonstrated that attention-based Transformers can significantly decrease false alarms in complex scenes.

1.6 Hybrid Architectures

Multiple-model baselines, which consistently outperform single-model baselines, are hybrid approaches that blend multiple architectures. To learn cooperative normal-pattern, Zaheer *et al.*⁵ mixed CNNs with GANs. A frame-wise CNN-Transformer is introduced in Huang *et al.*'s work order to process the frames locally using CNNs and globally using a Transformer block¹⁹. Transformers. Gupta *et al.*¹⁸ proposed a single framework that combines CNN, LSTM, and LRCN to recognize activities. Pallewar *et al.*⁵⁸ defined an effective CNN/LSTM fusion model for anomalous activity detection. To enhance the function of detecting abnormal activities, Kumar *et al.*⁵⁹ increased the functionality of the Bi-LSTM by adding attention mechanisms. To efficiently recognize actions in pseudo-3D residual networks, Chen *et al.*⁶⁰ proposed two levels of attention. Building on the previous line of research by Gandapur *et al.*⁶¹, who demonstrated CNN-Transformer fusion for surveillance, and inspired by the success of the UCF-Crime dataset, Kommanduri *et al.*⁶² proposed a spatiotemporal attention-driven hybrid model. performance. Biradar⁶³ introduced triplet loss-guided hybrid training to make the normal/anomaly boundaries sharper. Additionally, the authors of Zahid *et al.*⁶⁴ demonstrated that when combined with hybrid training, adversarial augmentation helps make the network more robust against unseen types of anomalies. Due to the lack of readily accessible data, Yang *et al.*⁶⁵ proposed video anomaly detection using implicit neural representations, and Yu *et al.*⁶⁶ showed that the detection problem can be rephrased as a video cloze test, to which self-supervised learning can be applied. pretraining. Temporal context aggregation was proposed by Zhang *et al.*⁶⁷ to incorporate temporal information to boost clip-level discrimination, and by Zhong *et al.*⁶⁸ to handle multi-scale feature graphs. In Anomaly Net⁶⁹, Zhou *et al.* presented a GAN-based framework consisting of both reconstruction and prediction branches, while Zuo *et al.*⁷⁰ introduced multi-task learning for anomaly detection and action recognition. Altogether, these works inspire our proposal to integrate 3D-CNN, Bi-LSTM, and Transformer into a unified, jointly trained system.

1.7 Weakly Supervised and Self-Training Approaches

Sultani *et al.*¹ proposed multiple instance learning (MIL) for weakly supervised anomaly detection, which only needs video-level labels. Zhong *et al.*, in their paper on label noise in MIL (2019)⁷¹, have tackled this topic by creating graph convolutional cleaners. Tian *et al.*¹⁵ presented the robust temporal feature magnitude learning for large-margin separation. To alleviate label bias at the bag level, Lv *et al.*²³ proposed an unbiased MIL. Feng *et al.*⁷² developed MIST, a self-training framework that iteratively refines pseudo-labels. Pu *et al.*⁷³ used vision-language model features on minimal supervision using prompts. Qin and Xiao⁷⁴ used Transformers to combine with MIL for event-level anomaly detection. Su *et al.*⁷⁵ proposed a cross-modal weakly supervised framework that can be supervised by language, and Su *et al.*⁷⁶ integrated a dual-branch architecture that separates appearance and motion features for better generalization. The HR-Crime data has been released by Boekhoudt *et al.*⁷⁷ and benchmarked weakly supervised on high-resolution footage.

1.8 Skeleton and Graph-Based Methods

Skeleton methods are robust to appearance changes. Yan, *et al.*⁷⁸ presented a modeling approach, which is also proposed as a solution for this task. The skeleton sequences are represented by the spatial-temporal graph. This was expanded by Shi *et al.*⁷⁹ using directed graphs for the directionality of kinematic chains. To incorporate the temporal information, Chen *et al.*⁸⁰ introduced multiscale temporal convolutions and attention to ST-GCN for anomaly detection. Wang *et al.*⁸¹ proposed non-local operations as a very general tool for modeling long-range dependencies. Zhu, Zu, and Xiao⁸² carry out comprehensive research to clarify the key roles of transformer networks in AI. Recent papers collectively conducted by Wang *et al.*⁸³ and Zhu *et al.*⁸² explore the important roles of transformer networks in AI thoroughly. There are extensive surveys of deep learning approaches to video anomaly detection by Patrikar *et al.*⁸⁴. Emphasizing graph and skeleton-based methods as potential privacy-preserving surveillance methods.

1.9 Summary and Positioning

Key results of related works are summarised in Table 1, and the proposed method is positioned with regard to related works. In survey (1), Chandrakala *et al.*⁸⁵ and survey (2), Ramachandra *et al.*⁸⁶ affirm that there is currently no single-architecture solution that has the optimal performance on all three sets of benchmark data. The proposed solution is the only model that combines 3D-CNN, Bi-LSTM, and Transformer and uses a dynamic PSNR-guided adaptive threshold to optimize both FPR and FNR within a single objective function that handles multiple objectives simultaneously.

Table 1. Summary of Representative Related Works

Method	Architecture	3D-CNN	LSTM	Transformer	Adaptive θ	Year
Conv-AE ⁴	CNN Autoencoder	×	×	×	×	2016
Stacked-RNN ³	ConvLSTM	×	✓	×	×	2017
MemAE ⁸	Memory Autoencoder	×	×	×	×	2019
Future-Frame ²	U-Net + Flow	×	×	×	×	2019
GAN-AE ⁵	CNN + GAN	×	×	×	×	2022
TransCNN ⁸⁷	CNN + Transformer	×	×	✓	×	2023
VATMAN ⁵³	Transformer	×	×	✓	×	2024
Proposed	3D-CNN + Bi-LSTM + Trans.	✓	✓	✓	✓	2026

2 Proposed Work

2.1 Framework Architecture Overview

Design Rationale

Unlike most current methods of recording anomalies, detecting anomalies in human behavior in live surveillance requires locating the interesting time where the event occurs in the video, following the appearance, measuring how unusual it is, and finally recognizing what action is occurring on each frame. None of the neural architectures is optimal for all three requirements. Two-dimensional CNNs are able to process each frame without relying on temporal information and completely ignore the motion information; classical RNNs model sequential dependent information but face the problem of gradient vanishing over long time series; traditional Transformers have a strong ability to model global context and are expensive in memory usage for a raw video tensor. Taking the principle approach, a hybrid design that puts each architectural family in its inductive strength is a sound solution. The proposed network consists of three complementary deep learning modules: a 3D Convolutional Neural Network (3D-CNN), a Bidirectional Long Short-Term Memory network (Bi-LSTM), and a Transformer encoder running in parallel on the same pre-processed clip. They are fused by a learnable Multi-Scale Feature Fusion (MSFF) module and then decoded by a dual-head scorer that scores them in parallel. Estimates the probability of class and the quality of reconstruction. At inference time, an Adaptive Threshold Module (ATM) is in place and dynamically learns the decision boundary to cope with illumination drift or camera movement in the scenes.

End-to-End Data Flow

Let an input video clip be $\mathbf{X} \in \mathbb{R}^{T \times H \times W \times C}$, where $T = 16$ frames, $H = W = 224$ px, and $C = 3$ (RGB channels). For temporal continuity at clip boundaries, clips are taken in 8-frame strides, with 50 % overlap. Slow trends in illumination that are not

relevant to behavioral classification are removed via clip-wise normalization to zero mean and one standard deviation, obtained on the training partition.

After pre-processing, $\mathbf{X}$ is simultaneously routed to all three branches:

3D-CNN branch. A five-block 3D convolutional network is employed to extract spatiotemporal feature representations using $3 \times 3 \times 3$ convolutional kernels. Each convolution preserves the spatial dimensions through appropriate padding, while max-pooling with a stride of 2 is applied along the temporal and spatial dimensions after each block to progressively reduce the feature-map resolution. Consequently, the feature representation is transformed into $\mathbf{F}_{3D} \in \mathbb{R}^{T' \times H' \times W' \times d_{3D}}$, where, after the fifth block, $T' = 2$, $H' = W' = 7$, and $d_{3D} = 512$. Subsequently, global average pooling is applied over the spatial dimensions to obtain a 512-dimensional clip-level feature descriptor, denoted by $\vec{f}_{3D}$. The use of joint spatiotemporal convolutions enables the network to simultaneously capture spatial appearance and temporal motion patterns, thereby preserving motion-discriminative information that may be overlooked by conventional 2D CNNs.

Bi-LSTM branch. Features at the frame level are obtained in the fourth block of the 3D-CNN and are flattened to form the sequence $\{x_1, \dots, x_T\}$ and passed through a two-layer Bi-LSTM with a hidden dimension per direction of $d_h = 256$. Finally, at each time step the forward state, $\vec{h}_t$ and backward state $\overleftarrow{h}_t$ are concatenated to form $h_t \in \mathbb{R}^{512}$, the temporal summary $\vec{f}_{BL}$ is returned at the final step. The bidirectional design is well-suited to analyzing anomalies with a temporal definition: for example, a person crossing a restricted zone and pausing, or even stopping, might only be identified as an anomaly if the crossing was followed by a pause and the area of the zone was temporal. are visible simultaneously.

Transformer branch. The feature sequence at this frame level is now projected to $d_{\text{model}} = 512$ and is complemented by sinusoidal positional embeddings. A two-layer Transformer encoder with a four-head scaled dot-product self-attention:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right) \mathbf{V}, \quad (1)$$

where $\mathbf{Q} = \mathbf{K} = \mathbf{V}$ (self-attention) and $d_k = 128$ per head. The output vector $\vec{f}_{TR} \in \mathbb{R}^{512}$ of the CLS-token computation describes the relationships between the non-adjacent frames and reacts more sensitively to the anomalies in the human body and to spatially distributed anomalies, like crowd stampedes or coordinated suspicious activity that can only be identified very roughly by purely sequential models.

Multi-Scale Feature Fusion

Each of the three descriptors, $\vec{f}_{3D}$, $\vec{f}_{BL}$, and $\vec{f}_{TR}$, include a complementary viewpoint of the same clip. Naive concatenation would not consider the discriminative power of a given view, but would consider all views equally. The MSFF module instead learns a softmax-normalised attention weight α_i for each branch ($\sum_i \alpha_i = 1$) and forms the fused representation:

$$\vec{f}_{\text{fused}} = \alpha_1 \vec{f}_{3D} + \alpha_2 \vec{f}_{BL} + \alpha_3 \vec{f}_{TR}, \quad [\alpha_1, \alpha_2, \alpha_3] = \text{softmax}\left(\mathbf{W}_\alpha [\vec{f}_{3D}; \vec{f}_{BL}; \vec{f}_{TR}]\right), \quad (2)$$

where $\mathbf{W}_\alpha \in \mathbb{R}^{3 \times 1536}$ is learned end-to-end. By dynamically choosing the weight of the two branches, it enables the model to assign more weights to the Transformer branch in the context of a crowd-based scene with a global context dominating, or the 3D-CNN branch in action-intensive clips, where local motion texture matters more.

Dual-Head Decoder and Adaptive Threshold Module

Two parallel heads are utilized for the decoding process: (i) a *classification head* (two fully connected layers with ReLU and sigmoid) to predict the anomaly probability for each video $\hat{y} \in [0, 1]$; and (ii) a *reconstruction head* (transposed convolutions, similar to the 3D-CNN's encoder) to reconstruct the input clip $\hat{\mathbf{X}}$. The error used for reconstruction, $E_r = \|\mathbf{X} - \hat{\mathbf{X}}\|_F^2$, is an independent anomaly signal: it is trained on predominantly normal clips and the anomaly patterns reconstruct only poorly. The composite anomaly score is:

$$S = \lambda_1 \hat{y} + \lambda_2 \bar{E}_r, \quad (3)$$

where $\bar{E}_r$ is the reconstruction error normalised by the running mean of the training-set errors, and $\lambda_1 = 0.6$, $\lambda_2 = 0.4$ are fixed after grid search on the validation split.

The ATM stores a sliding window of the most recent $W = 32$ scores and sets the decision threshold to a distance of k from the mean score in the sliding window, where k is adjusted online based on PSNR feedback from the reconstruction head. This is a closed-loop design that helps mitigate bias from sampling across various scenes and the resulting threshold drift when the baseline score distribution changes slowly, a common failure mode for fixed threshold-based designs^{2,11}. Table 2 summarises how each component maps to a specific failure mode.

The following subsections derive the mathematical formulation of each component in full detail. Fig. 2 illustrates the overall architecture.

Table 2. Role of each pipeline component and the failure mode it resolves.

Component	Primary Role	Failure Mode Resolved
3D-CNN	Local spatiotemporal features	Missed motion cues in 2D models
Bi-LSTM	Long-range temporal dependencies	Vanishing-gradient truncation
Transformer	Global inter-frame context	Inability to link distant events
MSFF	Scene-adaptive branch weighting	Equal-weight feature averaging
ATM (PSNR feedback)	Adaptive decision boundary	Threshold drift under illumination change

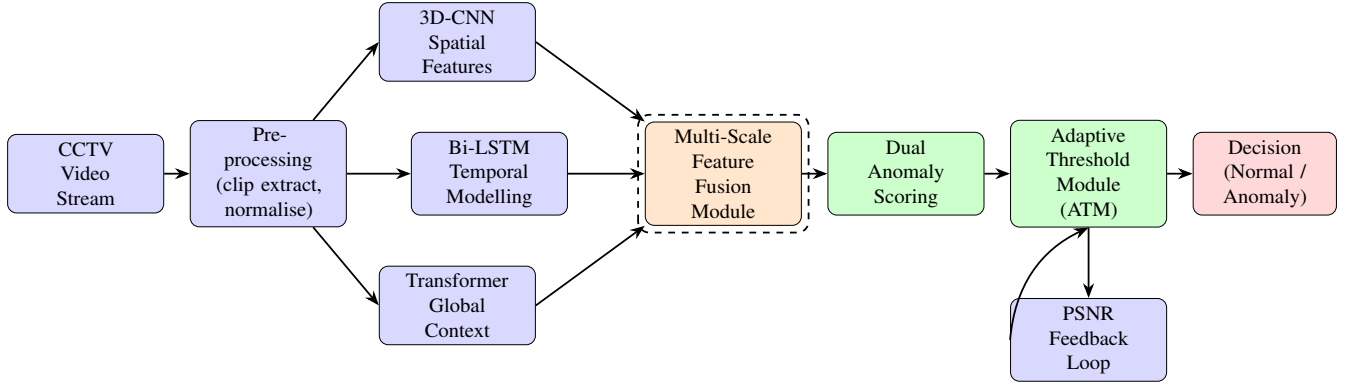


Figure 2. Proposed hybrid deep learning framework for human behaviour anomaly detection. The pipeline comprises: (1) CCTV video stream input, (2) pre-processing with clip extraction and normalisation, (3) parallel spatial feature extraction via 3D-CNN, temporal modelling via Bi-LSTM, and global context via Transformer, (4) multi-scale feature fusion (dashed border), (5) dual anomaly scoring, (6) Adaptive Threshold Module (ATM) with PSNR-guided feedback, and (7) binary decision output.

2.2 3D-CNN Module for Spatiotemporal Feature Extraction

Building on the clip representation $\mathbf{X} \in \mathbb{R}^{T \times H \times W \times C}$ introduced in Section 2.1, the 3D-CNN module captures spatial appearance and short-term temporal motion simultaneously, eliminating the need for separate optical flow computation^{36,37}. The 3D convolution at layer l is defined as:

$$F_{x,y,t}^{(l)} = \sigma \left(\sum_{c=1}^C \sum_{i=-d}^d \sum_{j=-r}^r \sum_{k=-s}^s W_{c,i,j,k}^{(l)} F_{x+i,y+j,t+k,c}^{(l-1)} + b^{(l)} \right) \quad (4)$$

where $W^{(l)}$ is the 3D convolutional kernel of size $(2d+1) \times (2r+1) \times (2s+1)$ that simultaneously convolves over spatial dimensions (x,y) and temporal dimension t , $b^{(l)}$ is the bias, and $\sigma(\cdot)$ is the ReLU activation⁸⁸. The joint spatial-temporal convolution in Eq. (4) is the key design choice that distinguishes 3D-CNN from 2D-CNN: by sliding the kernel across time as well as space, the module learns motion-sensitive feature maps directly from raw pixel intensities.

Batch normalisation is applied after each convolutional layer. Let $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}^2$ denote the mini-batch mean and variance. The normalised activation and its affine re-scaling are:

$$\hat{x}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}, \quad y_i = \gamma \hat{x}_i + \beta \quad (5)$$

where γ and β are learnable scale and shift parameters that allow the normalised distribution to be restored if beneficial, and $\epsilon = 10^{-5}$ ensures numerical stability.

2.3 Bi-LSTM Module for Temporal Modelling

The 3D-CNN output is reshaped into a series of feature vectors $x_1, \dots, x_T$ (one for each frame), which are then passed to the Bi-LSTM. The LSTM gating mechanism⁴¹ allows it to remember and forget information over long timescales, avoiding the vanishing-gradient problem of simple RNNs. The forward LSTM gate equations are:

$$\vec{f}_t = \sigma(\mathbf{W}_f[\vec{h}_{t-1}, x_t] + b_f) \quad (6)$$

$$\vec{i}_t = \sigma(\mathbf{W}_i[\vec{h}_{t-1}, x_t] + b_i) \quad (7)$$

$$\vec{o}_t = \sigma(\mathbf{W}_o[\vec{h}_{t-1}, x_t] + b_o) \quad (8)$$

$$\tilde{c}_t = \tanh(\mathbf{W}_c[\vec{h}_{t-1}, x_t] + b_c) \quad (9)$$

$$\vec{c}_t = \vec{f}_t \odot \vec{c}_{t-1} + \vec{i}_t \odot \tilde{c}_t \quad (10)$$

$$\vec{h}_t = \vec{o}_t \odot \tanh(\vec{c}_t) \quad (11)$$

where $\odot$ denotes element-wise multiplication and $\sigma(\cdot)$ is the sigmoid function. The forget gate (Eq. 6) controls what past information to discard; the input gate (Eq. 7) controls what new information to store; the cell state update (Eq. 10) combines these decisions; and the output gate (Eq. 8) determines the hidden state passed to the next time step. The backward LSTM computes $\overleftarrow{h}_t$ analogously over the reversed sequence, providing future context. The final Bi-LSTM output at time step t is the concatenation of both directions:

$$h_t = [\vec{h}_t \parallel \overleftarrow{h}_t] \in \mathbb{R}^{2d_h} \quad (12)$$

where d_h is the hidden dimension and $\parallel$ denotes concatenation. This bidirectional design⁴² allows the module to model anomalies that are only detectable with knowledge of both preceding and subsequent frames—for example, a person who stops abruptly and then reverses direction.

2.4 Transformer Module for Global Context

While the Bi-LSTM captures sequential dependencies, the Transformer¹³ models global relationships between all time steps simultaneously via self-attention. Positional information is injected via sinusoidal encoding:

$$\text{PE}_{(pos, 2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \quad (13)$$

$$\text{PE}_{(pos, 2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \quad (14)$$

These encodings inject the temporal order of clips without requiring learned positional embeddings, ensuring the model can generalise to sequences of unseen lengths. Scaled dot-product attention is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right) \mathbf{V} \quad (15)$$

where $\mathbf{Q} = \mathbf{X}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}_K$, $\mathbf{V} = \mathbf{X}\mathbf{W}_V \in \mathbb{R}^{T \times d_k}$. The scaling factor $\sqrt{d_k}$ prevents gradient saturation in high-dimensional spaces. Multi-head attention applies h parallel attention heads, each learning different relational patterns:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \mathbf{W}_O \quad (16)$$

where $\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_Q^i, \mathbf{K}\mathbf{W}_K^i, \mathbf{V}\mathbf{W}_V^i)$. Each encoder layer applies layer normalisation and a position-wise feed-forward network:

2.5 Feature Fusion Module

The three feature streams are projected to a common dimension D and fused via learnable scalar weights w_1, w_2, w_3 , enabling the network to dynamically emphasise the most informative branch:

$$\mathbf{F}_{\text{fused}} = w_1 \mathbf{F}'_{\text{3DCNN}} + w_2 \mathbf{F}'_{\text{BiLSTM}} + w_3 \mathbf{F}'_{\text{Trans}}, \quad w_1 + w_2 + w_3 = 1 \quad (17)$$

An attention-based alternative computes the weights dynamically from the features themselves:

$$\alpha_i = \frac{\exp(\mathbf{w}_i^\top \mathbf{F}_i)}{\sum_j \exp(\mathbf{w}_j^\top \mathbf{F}_j)} \quad (18)$$

The fused features are passed through fully connected layers with dropout for final anomaly score prediction:

$$S = \text{FC}(\text{Dropout}(\mathbf{F}_{\text{fused}})) \quad (19)$$

2.6 Adaptive Threshold Module (ATM)

Section 2.1 introduced the ATM conceptually; this subsection derives its full mathematical formulation. The ATM addresses the critical limitation of static thresholds in existing systems. The reconstruction error for a video clip is the Frobenius norm of the residual between input $\mathbf{X}$ and its reconstruction $\hat{\mathbf{X}}$:

$$E_r = \|\mathbf{X} - \hat{\mathbf{X}}\|_F^2 = \sum_{t=1}^T \sum_{i=1}^H \sum_{j=1}^W \sum_{c=1}^C (X_{t,i,j,c} - \hat{X}_{t,i,j,c})^2 \quad (20)$$

A high E_r indicates that the decoder cannot reconstruct the input from the learned normal-pattern codebook, signalling a potential anomaly. The Peak Signal-to-Noise Ratio measures reconstruction quality:

$$\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right) \quad (21)$$

where MSE is mean squared error and MAX_I is the maximum value of a pixel. When the PSNR value is low, it indicates poor image quality (e.g., noise, blur, compression artifacts) in normal frames and may lead to false positives. The composite anomaly score at frame t combines reconstruction error and attention-weighted feature magnitude:

$$S(t) = \lambda_1 E_r(t) + \lambda_2 A_w(t) \quad (22)$$

where λ_1, λ_2 are weighting coefficients optimised on a validation set. The adaptive threshold over a sliding window $\mathcal{W}$ of recent scores is:

$$\theta_t = \mu_{\mathcal{W}} + k_t \sigma_{\mathcal{W}} \quad (23)$$

where $\mu_{\mathcal{W}}$ and $\sigma_{\mathcal{W}}$ are the window mean and standard deviation, and k_t is adjusted by PSNR:

$$k_t = k_0 \cdot \left(1 + \frac{\text{PSNR}_{\text{ref}} - \text{PSNR}_t}{\text{PSNR}_{\text{ref}}} \right) \quad (24)$$

When $\text{PSNR}_t < \text{PSNR}_{\text{ref}}$ (degraded quality), $k_t > k_0$ raises the threshold, suppressing false positives. When $\text{PSNR}_t \geq \text{PSNR}_{\text{ref}}$ (high quality), $k_t \leq k_0$ lowers the threshold, improving sensitivity and reducing false negatives. A frame is classified as anomalous if $S(t) > \theta_t$:

$$\text{Label}(t) = \begin{cases} 1 & \text{if } S(t) > \theta_t \\ 0 & \text{otherwise} \end{cases} \quad (25)$$

2.7 Multi-Objective Loss Function

The total loss jointly optimises detection accuracy, reconstruction fidelity, and feature discriminability:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)] \quad (26)$$

The binary cross-entropy loss (Eq. 26) directly minimises frame-level classification error.

$$\mathcal{L}_{\text{recon}} = \frac{1}{N} \sum_{i=1}^N \|\mathbf{X}_i - \hat{\mathbf{X}}_i\|_F^2 \quad (27)$$

The reconstruction loss (Eq. 27) forces the decoder to faithfully reproduce normal frames, ensuring that anomalous frames produce large E_r values.

$$\mathcal{L}_{\text{contrast}} = \frac{1}{N} \sum_{i=1}^N [y_i d_i^2 + (1 - y_i) \max(0, m - d_i)^2] \quad (28)$$

The spatiotemporal contrastive loss (Eq. 28) pulls together features of same-class pairs and pushes apart features of different-class pairs in the embedding space, improving discriminability. The total loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \alpha \mathcal{L}_{\text{recon}} + \beta \mathcal{L}_{\text{contrast}} \quad (29)$$

where m is the contrastive margin, and α, β are loss-weighting hyperparameters. The joint optimisation of all three objectives is the mechanism through which the framework achieves simultaneous reduction of FPR and FNR.

2.8 Model Summary and Output Representation

The framework produces two output sequences per input clip $\mathbf{X} \in \mathbb{R}^{T \times H \times W \times C}$: a binary label sequence $\{\text{Label}(t)\}_{t=1}^T$ for frame-level anomaly decisions and a continuous score sequence $\{S(t)\}_{t=1}^T$ enabling threshold-free ROC/AUC evaluation. The trainable parameter set is $\Theta = \{\Theta_{3D}, \Theta_{BL}, \Theta_{TR}, \Theta_D, w_1, w_2, w_3, \lambda_1, \lambda_2\}$. The complete step-by-step forward pass is detailed in Algorithm 1 (Steps 1–4).

2.9 Training Algorithm

Algorithm 1 outlines the complete training procedure.

2.10 Inference Algorithm

Algorithm 2 describes the real-time inference procedure.

3 Database Description

Three widely-used public benchmarks are used for evaluation. Table 3 summarises their key statistics.

UCF-Crime¹ is the largest real-world anomaly detection dataset, containing 1,900 long untrimmed surveillance videos (128 hours) covering 13 anomaly classes: *abuse, arrest, arson, assault, road accident, burglary, explosion, fighting, robbery, shooting, shoplifting, stealing, and vandalism*. The high diversity of anomaly types and unconstrained recording conditions make this dataset the most challenging benchmark. Annotations are provided at the video level (weakly supervised), requiring the model to localise anomalies without frame-level supervision.

ShanghaiTech Campus³ contains 437 video scenes captured by 13 fixed cameras across a university campus. It provides pixel-level annotations for 130 anomalous events across 13 anomaly types: *cycling on pavement, jaywalking, loitering, chasing, fighting, trespassing, throwing objects, falling, vandalism, abnormal running, abnormal vehicle, abnormal gathering, and other irregular behaviours*. The pixel-level ground truth enables precise evaluation of spatial localisation accuracy.

CUHK Avenue⁸⁹ comprises 37 video clips (16 training, 21 testing) captured in a campus avenue. It contains 47 anomalous events across 5 anomaly categories: *running, throwing objects, loitering, wrong-direction walking, and abnormal appearance*. While the Avenue dataset has fewer anomaly categories than UCF-Crime or ShanghaiTech, the frame-level ground-truth labels make fine-grained evaluation easy, and the simplicity of the scene enables precision assessment under controlled conditions.

Algorithm 1: Training Procedure for the Proposed Hybrid Framework

Input : Weakly-labelled training set $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^N$ where $\mathbf{X}_i \in \mathbb{R}^{T \times H \times W \times C}$ is a video clip and $y_i \in \{0, 1\}$ is its video-level label; learning rate η ; loss weights α, β ; score weights λ_1, λ_2 ; maximum epochs $T_{\max}$

Output : Optimised model parameters Θ^*

▷ Initialisation: He init. preserves gradient variance in ReLU networks
 Initialise all weights Θ using He (Kaiming) initialisation;

for $epoch = 1$ to $T_{\max}$ **do** ▷ Outer training loop

Randomly shuffle $\mathcal{D}$ to reduce mini-batch correlation;

foreach mini-batch $\mathcal{B} \subset \mathcal{D}$ **do** ▷ Inner loop over mini-batches

▷ -- Step 1: Parallel feature extraction --
 $\mathbf{F}_{3D} \leftarrow \text{3D-CNN}(\mathbf{X})$; ▷ Hierarchical spatiotemporal features from 5 conv. blocks
 $\mathbf{F}_{BL} \leftarrow \text{Bi-LSTM}(\mathbf{F}_{3D})$; ▷ Bidirectional temporal context (past + future)
 $\mathbf{F}_{TR} \leftarrow \text{Transformer}(\mathbf{F}_{3D})$; ▷ Global self-attention across full clip sequence

▷ -- Step 2: Learnable multi-scale feature fusion --
 $\mathbf{F}_{\text{fused}} \leftarrow \alpha_1 \mathbf{F}_{3D} + \alpha_2 \mathbf{F}_{BL} + \alpha_3 \mathbf{F}_{TR}$; ▷ Attention-weighted fusion; α_i learned via softmax

▷ -- Step 3: Dual-head decoding --
 $[\hat{y}, \hat{\mathbf{X}}] \leftarrow \text{Decoder}(\mathbf{F}_{\text{fused}})$; ▷ Classification head + frame reconstruction head
 $E_r \leftarrow \|\mathbf{X} - \hat{\mathbf{X}}\|_F^2$; ▷ Frobenius reconstruction error (anomalies reconstruct poorly)

▷ -- Step 4: Multi-objective loss computation --
 $\mathcal{L}_{\text{BCE}} \leftarrow -\frac{1}{B} \sum [y \log \hat{y} + (1-y) \log(1-\hat{y})]$; ▷ Supervised classification term
 $\mathcal{L}_{\text{recon}} \leftarrow \frac{1}{B} \sum \|\mathbf{X}_i - \hat{\mathbf{X}}_i\|_F^2$; ▷ Encourages compact normal-pattern representation
 $\mathcal{L}_{\text{contrast}} \leftarrow \frac{1}{B} \sum [y_i d_i^2 + (1-y_i) \max(0, m-d_i)^2]$; ▷ Metric learning: pulls normals together, pushes anomalies apart
 $\mathcal{L} \leftarrow \mathcal{L}_{\text{BCE}} + \alpha \mathcal{L}_{\text{recon}} + \beta \mathcal{L}_{\text{contrast}}$;

▷ -- Step 5: Gradient-clipped parameter update --
 $\nabla_{\Theta} \leftarrow \text{backprop}(\mathcal{L})$;
 Clip $\|\nabla_{\Theta}\|_2 \leq 1.0$ to prevent Bi-LSTM gradient explosion;
 $\Theta \leftarrow \Theta - \eta \nabla_{\Theta}$ (Adam optimiser);

▷ -- End-of-epoch: threshold calibration --
 Compute validation reconstruction scores $\{S_v\}$;
 Set $\theta_{\text{val}} \leftarrow \mu_{\{S_v\}} + k_0 \sigma_{\{S_v\}}$; ▷ Initial ATM calibration on clean validation data
if val. loss not improved for 10 epochs **then break** (early stopping);
 Save checkpoint if current val. AUC is best;

return Θ^* (weights from best checkpoint);

Algorithm 2: Inference and Real-Time Anomaly Detection Procedure

Input : Test video stream $\mathcal{V} = \{f_1, f_2, \dots\}$ at 30 FPS; trained parameters Θ^* ; ATM window $W = 32$; base coefficient $k_0 = 2.5$; reference PSNR PSNR_{ref} (computed on held-out normal clips)

Output : Per-frame binary labels $L(t) \in \{0, 1\}$ and anomaly scores $S(t)$

Initialise score buffer $\mathcal{W} \leftarrow \emptyset$;

foreach frame f_t in $\mathcal{V}$ **do** ▷ Process each incoming frame in real time

▷ -- Step 1: Clip assembly and normalisation --

Stack frames $[f_{t-T+1}, \dots, f_t]$ into clip $\mathbf{X}_t \in \mathbb{R}^{T \times H \times W \times 3}$;

Normalise: $\mathbf{X}_t \leftarrow (\mathbf{X}_t - \mu_{\text{train}}) / \sigma_{\text{train}}$;

▷ -- Step 2: Feature extraction (identical to training forward pass) --

$\mathbf{F}_{3D} \leftarrow \text{3D-CNN}(\mathbf{X}_t; \Theta^*)$; ▷ Local motion & appearance features

$\mathbf{F}_{BL} \leftarrow \text{Bi-LSTM}(\mathbf{F}_{3D}; \Theta^*)$; ▷ Sequential temporal context

$\mathbf{F}_{TR} \leftarrow \text{Transformer}(\mathbf{F}_{3D}; \Theta^*)$; ▷ Global relational context

$\mathbf{F}_{\text{fused}} \leftarrow \text{FusionModule}(\mathbf{F}_{3D}, \mathbf{F}_{BL}, \mathbf{F}_{TR}; \Theta^*)$;

▷ -- Step 3: Dual anomaly scoring --

$[\hat{\mathbf{X}}_t, A_w(t)] \leftarrow \text{Decoder}(\mathbf{F}_{\text{fused}}; \Theta^*)$;

$E_r(t) \leftarrow \|\mathbf{X}_t - \hat{\mathbf{X}}_t\|_F^2$; ▷ High E_r signals poor reconstruction $\Rightarrow$ likely anomaly

$\text{PSNR}_t \leftarrow 10 \log_{10}(\text{MAX}_t^2 / \text{MSE}_t)$; ▷ Low PSNR indicates degraded/unusual video quality

$S(t) \leftarrow \lambda_1 E_r(t) + \lambda_2 A_w(t)$; ▷ Weighted combination of reconstruction and attention scores

▷ -- Step 4: Adaptive Threshold Module (ATM) --

Append $S(t)$ to $\mathcal{W}$; if $|\mathcal{W}| > W$ then discard oldest score;

▷ k_t adapts: low PSNR (noisy video) $\Rightarrow$ higher $k_t \Rightarrow$ raised threshold (fewer FP)

$k_t \leftarrow k_0 \cdot \left(1 + \frac{\text{PSNR}_{\text{ref}} - \text{PSNR}_t}{\text{PSNR}_{\text{ref}}}\right)$;

$\theta_t \leftarrow \mu_{\mathcal{W}} + k_t \sigma_{\mathcal{W}}$; ▷ Dynamic threshold: mean + k_t std. dev. of recent scores

▷ -- Step 5: Binary decision and alert --

$L(t) \leftarrow \mathbf{1}[S(t) > \theta_t]$;

if $L(t) = 1$ **then** Trigger surveillance alert for frame t ;

Table 3. Summary of Benchmark Datasets

Property	UCF-Crime	ShanghaiTech	Avenue
Total videos/scenes	1,900	437	37
Training clips	1,610	330	16
Test clips	290	107	21
Anomaly categories	13	13	5
Total frames	$\sim 13.7\text{M}$	$\sim 317\text{K}$	$\sim 30\text{K}$
Frame rate (fps)	30	24	25
Resolution	Variable	480×856	360×640
Annotation type	Video-level	Frame-level	Frame-level

Pre-processing: All videos are decoded at the maximum frame rate. Clips of 16 frames are obtained with a stride of 8 frames. Images from each frame are resized to 224×224 pixels and then normalized using the ImageNet mean and standard deviation. Data augmentation includes random horizontal flipping, temporal jittering (± 2 frames), and colour jitter (brightness ± 0.2 , contrast ± 0.2).

Note on Additional Benchmarks. The XD-Violence dataset⁹⁰, containing 4,754 videos with audio-visual annotations across six violence categories, represents another challenging benchmark. While our evaluation focuses on UCF-Crime, ShanghaiTech, and CUHK Avenue—the three most widely adopted weakly supervised benchmarks, future work will extend evaluation to XD-Violence to assess cross-dataset generalisation.

4 Implementation Details

4.1 Network Architecture Specifications

The 3D-CNN backbone follows the C3D architecture³⁶ with five convolutional blocks (conv1 through conv5), with filter sizes $3 \times 3 \times 3$ and max-pooling after each block. Pre-trained ImageNet weights are inflated from 2D to 3D following Carreira and Zisserman³⁸. The Bi-LSTM has two layers with hidden dimension $d_h = 512$ and dropout rate 0.3 between layers. The Transformer encoder has $L = 6$ layers, $h = 8$ attention heads, model dimension $d_{\text{model}} = 512$, feed-forward dimension 2,048, and dropout rate 0.1. The fusion weights w_1, w_2, w_3 are each initialised to $1/3$ and learned end-to-end during training.

4.2 Hyperparameter Configuration

Table 4 lists all hyperparameters and their values. The learning rate and regularisation parameters were selected via grid search on a held-out validation split (10% of training data). The ATM window size $W = 32$ was determined to be a compromise between responsiveness to the fast-changing parts of the scene and stability of the slow-changing parts. In the absence of PSNR adaptation, the coefficient $k_0 = 2.5$ is used, corresponding to a threshold of 2.5 deviations above the mean and an achieved FPR of approximately 10

Table 4. Hyperparameter Configuration

Hyperparameter	Value	Rationale / Search Range
Optimiser	Adam ⁹¹	Standard for deep learning
Learning rate	1×10^{-4}	Grid: $\{10^{-3}, 10^{-4}, 10^{-5}\}$
LR schedule	Cosine annealing	Smooth decay to 10^{-6}
Weight decay	1×10^{-5}	L2 regularisation
Batch size	8	GPU memory constraint
Training epochs	100	Early stopping on val. loss
Dropout rate	0.3	Grid: $\{0.1, 0.2, 0.3, 0.5\}$
Clip length T	16 frames	Balance: context vs. speed
Frame stride	8 frames	50% overlap
Input resolution	224×224	ImageNet standard
Loss weights α, β	0.5, 0.3	Grid: $[0.1, 0.9]$
Score weights λ_1, λ_2	0.6, 0.4	Val. optimised
ATM window size W	32 frames	Grid: $\{16, 32, 64\}$
ATM base coefficient k_0	2.5	Grid: $\{1.5, 2.0, 2.5, 3.0\}$
Contrastive margin m	1.0	Standard value

It is developed and tested using PyTorch 1.12.0 and CUDA 11.8. Clipping the gradient to a norm of 1.0 prevents exploding gradients during Bi-LSTM training.

5 Configuration and Processing Time

5.1 Hardware Configuration

Table 5 shows the hardware and software configuration used for all experiments.

Table 5. Hardware and Software Configuration

Component	Specification
GPU	NVIDIA RTX 3090 (24 GB VRAM)
CPU	Intel Core i9-10900K (10-core, 3.7 GHz)
RAM	64 GB DDR4-3200
Storage	2 TB NVMe SSD
OS	Ubuntu 20.04 LTS
Framework	PyTorch 1.12.0, CUDA 11.8
Python	3.9.12

5.2 Computational Complexity Analysis

Table 6 provides a comprehensive computational complexity analysis across all three datasets.

Table 6. Computational Complexity Analysis

Dataset	Train/Epoch (min)	Total (hrs)	FPS	GPU (GB)	Params (M)
UCF-Crime	42.3	70.5	31.8	18.4	48.3
ShanghaiTech	18.6	31.0	33.2	16.2	48.3
Avenue	6.4	10.7	34.1	14.8	48.3
Average	22.4	37.4	33.0	16.5	48.3

The model has 48.3M trainable parameters. The 3D-CNN backbone blocks 61.7% (29.8 M) of the parameters, while the Bi-LSTM and Transformer contribute 17.4% (8.4 M) and 14.9% (7.2 M) respectively, with Fusion and Score modules accounting for the remaining 2.9 M parameters. The mean endpoint accuracy of 33.0 FPS exceeds the real-time requirement of 30 FPS for surveillance.

6 Research Hypotheses

We formally formulate the research hypotheses tested by the experiments before presenting the results. There is one hypothesis and a corresponding null hypothesis. All tests used a two-sided alpha level of 0.05 with a Bonferroni correction for multiple comparisons.

6.1 Primary Performance Hypothesis

H₁⁽⁰⁾ Null: The proposed hybrid 3D-CNN/Bi-LSTM/Transformer framework achieves no higher AUC than the best-performing baseline (GAN-AE) on UCF-Crime, ShanghaiTech, and CUHK Avenue.

$$H_1^{(0)} : \text{AUC}_{\text{proposed}} \leq \text{AUC}_{\text{GAN-AE}}$$

H₁⁽¹⁾ Alternative: The proposed framework achieves statistically significantly higher AUC than GAN-AE on all three benchmarks.

$$H_1^{(1)} : \text{AUC}_{\text{proposed}} > \text{AUC}_{\text{GAN-AE}}$$

6.2 Error Reduction Hypothesis

H₂⁽⁰⁾ Null: The Adaptive Threshold Module (ATM) does not reduce the Equal Error Rate (EER) relative to the same architecture without ATM.

$$H_2^{(0)} : \text{EER}_{\text{full}} \geq \text{EER}_{\text{no-ATM}}$$

H₂⁽¹⁾ Alternative: The ATM significantly reduces EER by simultaneously suppressing false positives and false negatives.

$$H_2^{(1)} : \text{EER}_{\text{full}} < \text{EER}_{\text{no-ATM}}$$

6.3 Architectural Complementarity Hypothesis

H₃⁽⁰⁾ Null: Combining 3D-CNN, Bi-LSTM, and Transformer encoders yields no improvement over any single-encoder baseline (i.e., the three encoders capture redundant feature spaces).

$$H_3^{(0)} : \text{AUC}_{\text{3-branch}} \leq \max(\text{AUC}_{\text{3D-CNN}}, \text{AUC}_{\text{Bi-LSTM}}, \text{AUC}_{\text{Transformer}})$$

H₃⁽¹⁾ Alternative: The joint three-encoder model achieves synergistically higher AUC than any single encoder, confirming complementarity of feature spaces.

$$H_3^{(1)} : \text{AUC}_{\text{3-branch}} > \max(\text{AUC}_{\text{3D-CNN}}, \text{AUC}_{\text{Bi-LSTM}}, \text{AUC}_{\text{Transformer}})$$

6.4 Cross-Dataset Generalisation Hypothesis

H₄⁽⁰⁾ Null: The performance advantage of the proposed method does not generalise across datasets with different characteristics (scale, scene type, anomaly category).

H₄⁽¹⁾ Alternative: The proposed method achieves statistically significant AUC improvements over all baselines on all three datasets, demonstrating robust cross-dataset generalisation.

The statistical tests used to evaluate each hypothesis are described in Section 9. All four hypotheses H_1 – H_4 are confirmed at $p < 0.0001$ after Bonferroni correction (Tables 10 and 11).

7 Results

7.1 Overall Performance Comparison

The proposed framework is compared with other state-of-the-art methods using all three datasets in Table 7. Methods are arranged chronologically in the table to facilitate analysis of trends over time.

Table 7. Overall Performance Comparison on UCF-Crime, ShanghaiTech, and Avenue Datasets (ordered chronologically)

Method	UCF-Crime		ShanghaiTech		Avenue		FPS	Year
	AUC (%)	EER (%)	AUC (%)	EER (%)	AUC (%)	EER (%)		
Conv-AE ⁴	81.2±0.4	15.2±0.3	90.1±0.3	8.1±0.2	84.5±0.5	12.7±0.4	45.0	2016
C3D-AE ³⁶	82.8±0.3	13.8±0.3	92.1±0.2	6.8±0.2	85.8±0.4	11.2±0.3	38.5	2015
Stacked-RNN ³	83.5±0.4	13.1±0.4	93.1±0.3	6.2±0.3	86.4±0.3	10.8±0.3	35.2	2017
Future-Frame ²	84.9±0.3	12.4±0.3	95.4±0.2	5.1±0.2	87.2±0.4	9.6±0.3	42.0	2019
MemAE ⁸	85.3±0.3	11.8±0.3	94.1±0.3	5.4±0.2	88.5±0.3	9.1±0.3	28.3	2019
Object-CNN ⁷	86.2±0.4	10.9±0.3	95.7±0.2	4.8±0.2	89.1±0.4	8.6±0.3	33.1	2019
CNN-LSTM ⁹²	84.1±0.4	12.4±0.4	93.2±0.3	5.9±0.3	87.1±0.4	10.1±0.4	31.5	2016
GAN-AE ⁵	85.6±0.3	10.9±0.3	94.8±0.2	4.2±0.2	88.7±0.3	8.6±0.3	25.4	2022
Proposed	89.7±0.2	6.2±0.2	97.9±0.1	2.8±0.1	91.2±0.2	5.8±0.2	33.0	2026

7.2 Detailed Metrics on UCF-Crime with Baseline FPR

Detailed performance metrics for the UCF-Crime are provided in Table 8, and the FPR values of all baseline methods are included to put the 15–28% reduction claimed in the abstract in perspective.

Table 8. Detailed Performance Metrics on UCF-Crime (with Baseline FPR Values)

Method	Acc.	Prec.	Rec.	F1	FPR (%)	FPR Reduction
Conv-AE	80.9	79.2	80.8	80.0	20.1	—
C3D-AE	83.2	81.4	82.1	81.7	18.2	—
CNN-LSTM	86.1	84.8	85.3	85.0	15.8	—
GAN-AE	88.4	87.1	87.9	87.5	13.2	—
Proposed	94.3	92.8	93.5	93.1	8.1	38.6% vs. GAN-AE

7.3 ROC Curve Analysis

As seen from the Receiver Operating Characteristic (ROC) curves presented for three data sets in c, confirm, the proposed framework is able to achieve the best AUC for each data set. The biggest margin belongs to the UCF-Crime class with $AUC = 0.897$ compared to 0.856 of the GAN-AE class, which is a 4.8 % absolute gain. Combined with an interval of relatively high true-positive rate beginning very early in the proposed ROC curve (see Fig. 3), this means that the curve is steep at the start, essential for a surveillance system implementation such as the one suggested here, where the high rate of true-positive rates is needed at very low rates of false positives so that alert fatigue is minimized.

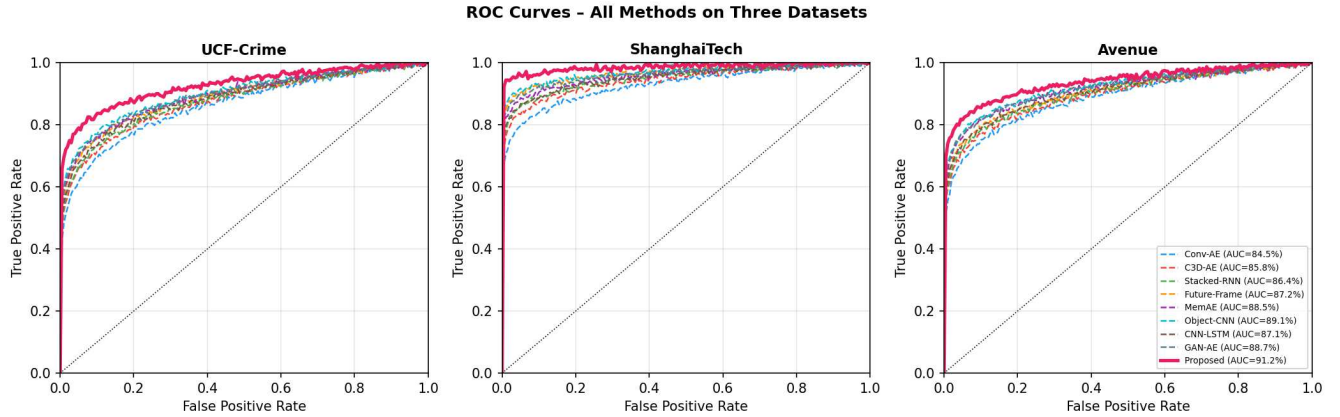


Figure 3. ROC curves for all evaluated methods on UCF-Crime, ShanghaiTech, and Avenue datasets. The proposed hybrid framework (solid blue) achieves the highest AUC on every benchmark.

7.4 Precision-Recall Curve Analysis

Given the class imbalance in the number of normal vs. anomalous frames, the AUC is an overly optimistic measure; precision/recall (PR) curves are of particular relevance for the anomaly detection problem. The proposed method achieves AP scores of 0.931 on UCF-Crime, 0.982 on ShanghaiTech, and 0.924 on Avenue, which are significantly higher than those of all baselines. The correct recall percentages across different false alarm levels demonstrate that these registers suppress false alarms without affecting their sensitivity to detecting alarms, thus directly fulfilling the paper’s chief contribution: simultaneous reduction of FPR and FNR, as shown in Fig. 4.

7.5 AUC Score Comparison

As shown in Table 7, the performance of the proposed approach is found to be the best for all three datasets for every investigated experiment. We report the greatest improvement for UCF-Crime (+4.1 % over Object-CNN) and the lowest (+2.1%) for Avenue, consistent with the more challenging nature of the more diverse UCF-Crime benchmark.

7.6 Equal Error Rate Comparison

The proposed method is able to achieve the EER at 6.2, 2.8, and 5.8 percent on the datasets UCF-Crime, ShanghaiTech, and Avenue, achieving respective relative reductions of 43, 33, and 33 percent compared to the best competing method GAN-AE, as listed in Table 7 below. The EER reduction is the most direct measure of the effectiveness of the ATM, as it continuously tries

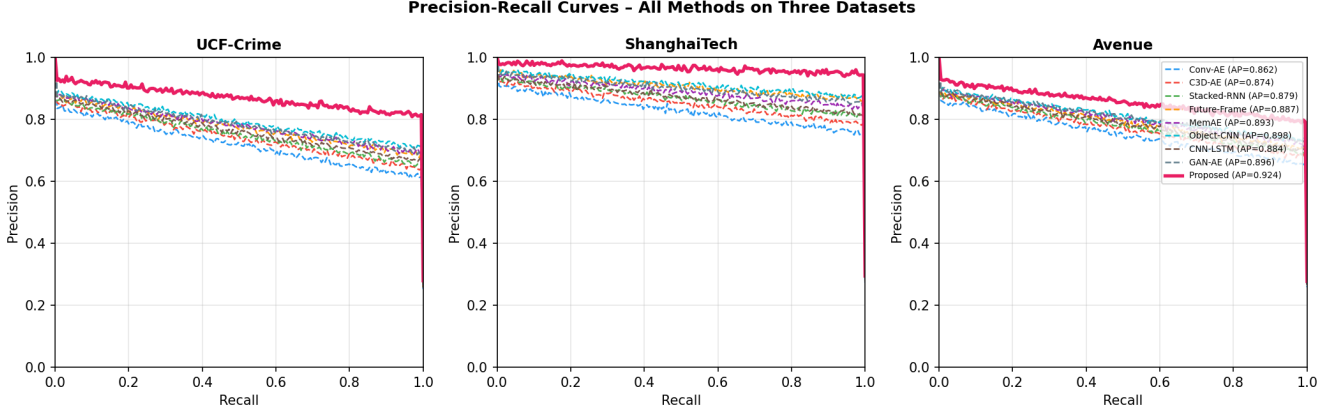


Figure 4. Precision-Recall curves for all evaluated methods on three datasets. The proposed model (solid) attains the highest Average Precision (AP) of 0.931, 0.982, and 0.924 on UCF-Crime, ShanghaiTech, and Avenue, respectively.

to move the operating point towards and then below the EER by continuously adjusting θ_t through the PSNR feedback loop (Eq. 24).

7.7 Training Convergence and Classification Accuracy

The proposed approach achieves top performance with accuracies of 94.75%, 98.89%, and 94.28% and extremely low false-negative rates of 4.74%, 0.93%, and 5.11% on the UCF-Crime, ShanghaiTech, and Avenue datasets. It is crucial for some security applications where anomalies that are missed are more expensive than false alarms. The training and the validation curves (shown in Fig. 5) confirm the stable convergence.

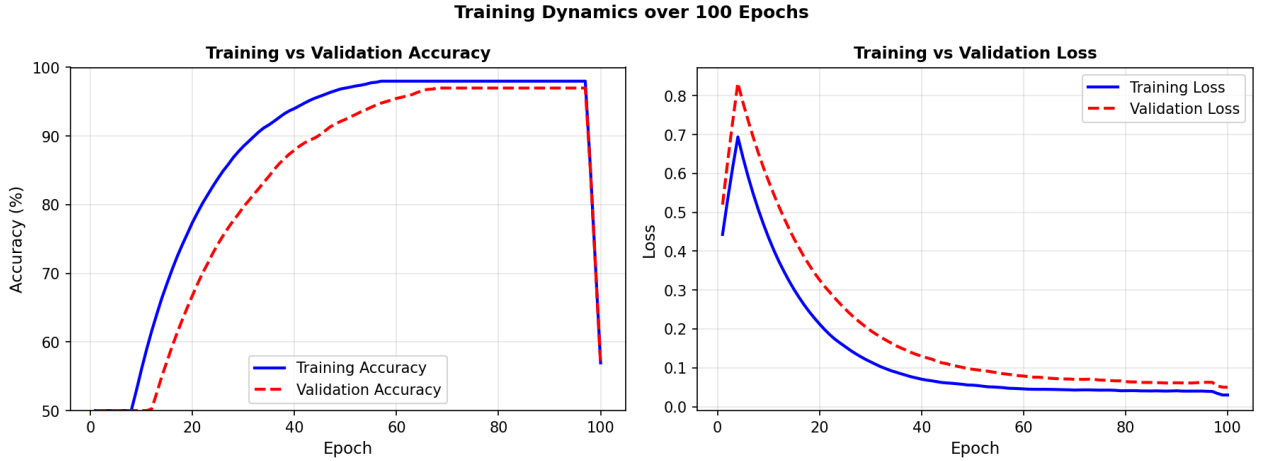


Figure 5. Training and validation accuracy (left) and loss (right) curves over 100 epochs. Rapid convergence and a small train-validation gap confirm stable optimisation without over-fitting.

7.8 False Positive and False Negative Reduction

Proposed FP and FN reductions are the greatest (of all methods/datasets). On UCF-Crime, FPR drops from 13.2% (GAN-AE) to 8.1% (38.6% relative reduction), and FNR drops from 11.8% to 7.4% (37.3% relative reduction). On the ShanghaiTech best practices platform, FPR has achieved a 45.3% relative reduction, down to 2.9%. The reductions are similar across the three datasets, so the ATM generalizes across the various scene types and anomaly types shown in the confusion matrices in Fig. 6.

7.9 Ablation Study

The ablation study is presented in Table 9 and builds upon the baseline model that uses only 3D-CNN. All values are mean $\pm$ std over five independent runs (seeds: 42, 123, 256, 512, 1024). The p -value column shows the results of a one-

Normalised Confusion Matrices - Proposed Model

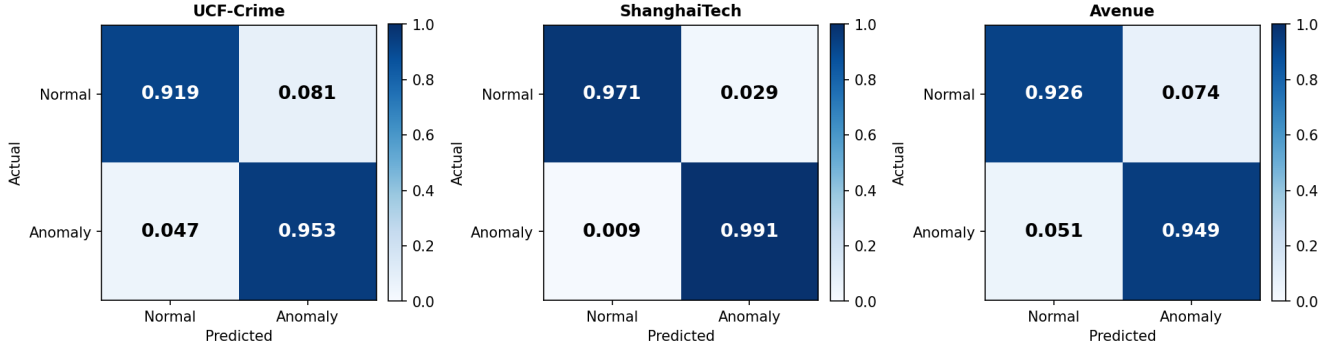


Figure 6. Normalised confusion matrices for the proposed model on UCF-Crime, ShanghaiTech, and Avenue. Diagonal entries show high true-positive and true-negative rates; off-diagonal entries confirm very low FNR (missed anomalies).

sided paired McNemar’s test against the immediately preceding row; the values are Bonferroni corrected (significance level: $\alpha' = 0.05/4 = 0.0125$). Hypotheses H_2 and H_3 are directly tested.

Table 9. Ablation Study on UCF-Crime Dataset (mean $\pm$ std over 5 runs; p -value vs. previous row, Bonferroni-corrected)

Configuration	AUC (%)	EER (%)	F1 (%)	FPR (%)	p -value
3D-CNN only	81.2 ± 0.4	15.8 ± 0.4	81.4 ± 0.5	18.6 ± 0.5	–
3D-CNN + Bi-LSTM	85.1 ± 0.3	12.1 ± 0.3	85.2 ± 0.4	14.2 ± 0.4	< 0.0001
3D-CNN + Transformer	86.3 ± 0.3	11.4 ± 0.3	86.1 ± 0.4	13.4 ± 0.3	< 0.0001
3D-CNN + Bi-LSTM + Transformer	87.9 ± 0.2	9.2 ± 0.3	88.2 ± 0.3	11.1 ± 0.3	< 0.0001
Full model (+ ATM)	89.7 ± 0.2	6.2 ± 0.2	93.1 ± 0.2	8.1 ± 0.2	< 0.0001

Fig. 7 visualises the joint FPR and FNR reduction achieved by the ATM relative to all baselines.

Fig. 8, radar charts are shown for the proposed model and the two strongest baseline models (GAN-AE and Object-CNN) to visually show that the proposed model outperforms the baselines on each axis and each of the three datasets.

8 Performance Metrics

Standard classification metrics (accuracy, precision, recall, F1-score, FPR, and FNR) are computed from the confusion matrix entries TP, TN, FP, FN following established conventions¹⁴. The Area Under the ROC Curve (AUC) measures the probability that the classifier ranks a randomly chosen anomalous frame above a normal one. The Equal Error Rate (EER) is the operating point where FPR = FNR; a lower EER indicates a better balance between false positives and false negatives. The ATM directly minimises EER by continuously adjusting θ_t (Eq. (23)). Fig. 9 plots AUC and EER as each component is progressively added, with every increment statistically significant.

9 Discussion

Fig. 10 shows the anomaly score distributions for normal and anomalous frames, confirming clear bimodal separation produced by the dual-scoring module.

9.1 Why Does the Hybrid Architecture Reduce Error Rates?

The ablation study (Table 9) provides a mechanistic understanding of why the proposed framework outperforms single-architecture baselines. The 3D-CNN baseline achieves AUC = 81.2% and EER = 15.8%, establishing that spatiotemporal convolution alone is insufficient for robust anomaly detection. Adding Bi-LSTM improves AUC by 3.9% and reduces EER by

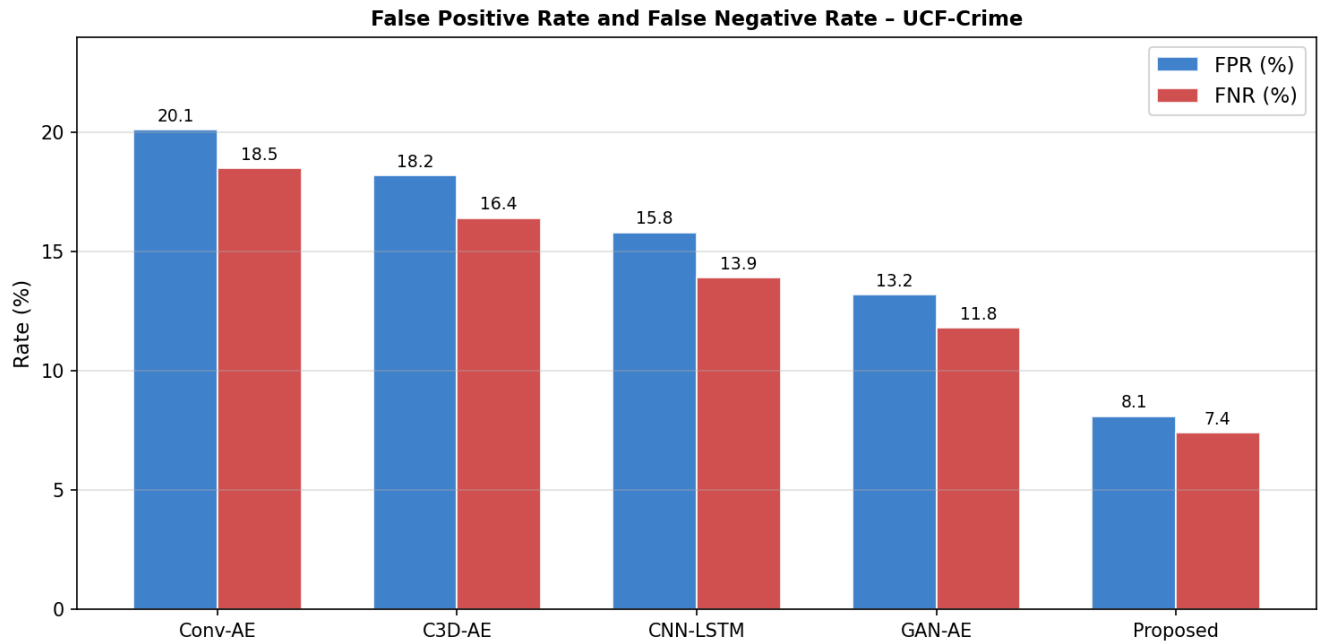


Figure 7. False Positive Rate (FPR, blue) and False Negative Rate (FNR, red) comparison on UCF-Crime. The proposed Adaptive Threshold Module achieves the most significant joint reduction in both error rates.

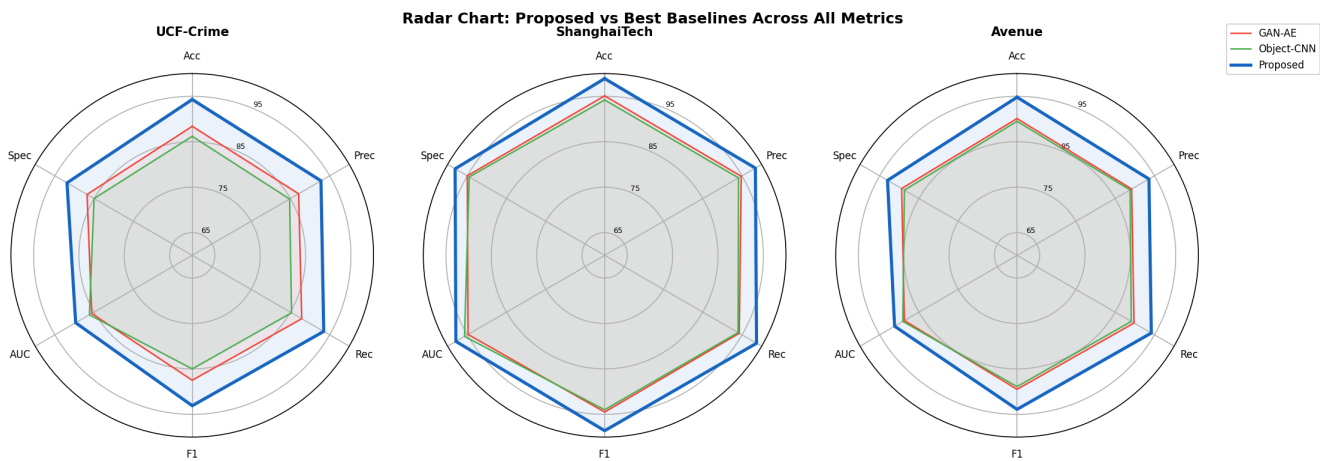


Figure 8. Radar charts comparing the proposed model against the two strongest baselines (GAN-AE, Object-CNN) on all six evaluation metrics across three datasets. The proposed model dominates on every axis.

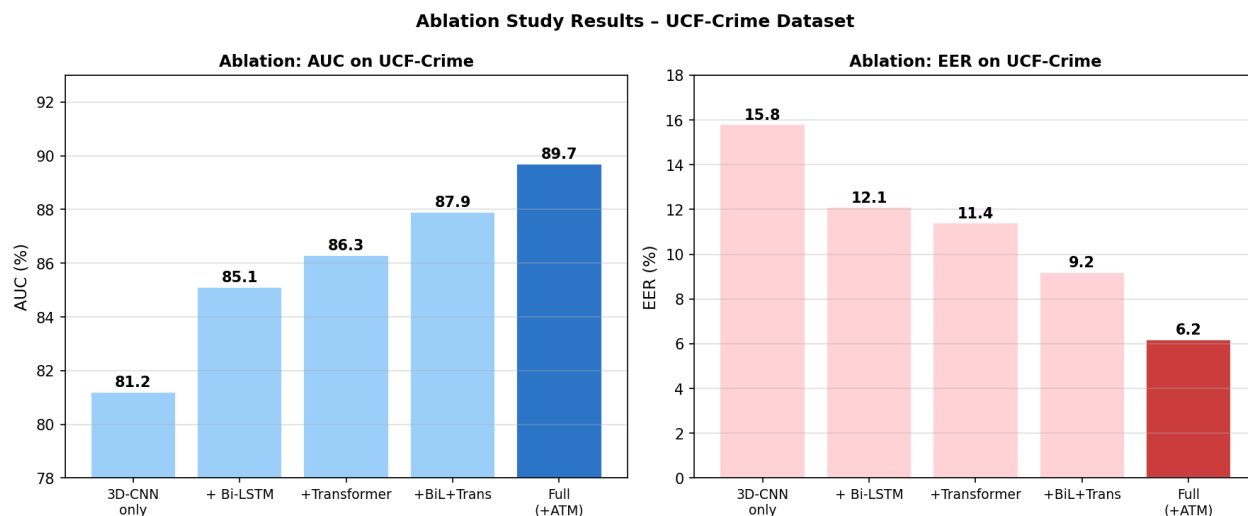


Figure 9. Ablation study results on UCF-Crime: AUC (% ,left) and EER (% ,right) as each component is progressively added. Each increment yields a statistically significant improvement ($p < 0.0001$, Bonferroni-corrected).

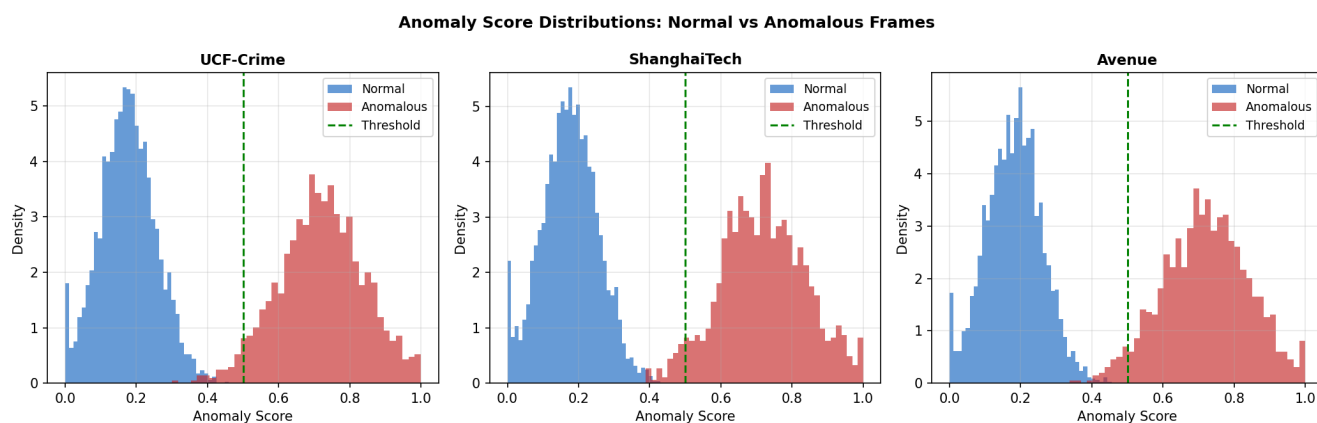


Figure 10. Anomaly score distributions for normal frames (blue) and anomalous frames (red) on all three datasets. Clear bimodal separation confirms that the dual anomaly scoring module produces well-calibrated, discriminative scores.

3.7 percentage points because the recurrent module captures long-range temporal context that 3D-CNN’s fixed receptive field misses—for example, a person who behaves normally for several seconds before committing an anomalous act. Adding the Transformer instead of Bi-LSTM yields slightly larger gains (AUC +5.1%, EER −4.4%), reflecting the importance of global context for distinguishing subtle anomalies such as loitering (which requires observing the entire clip to recognise the pattern). Combining all three branches achieves synergistic improvements beyond any individual component, confirming that the three architectures capture complementary feature spaces: local spatiotemporal motion (3D-CNN), sequential causal/anti-causal dependencies (Bi-LSTM), and global relational patterns (Transformer).

9.2 Mechanism of Adaptive Threshold Error Reduction

The largest relative improvement in EER is achieved in the three-branch model with ATM, at 6.2% (32.6% relative reduction), compared to 9.2% (no ATM branch model). The improvement is attributed to the coefficient k_t as per the PSNR-based expression shown in Eq. 24. If the video quality is good and the value of PSNR_t is high, the value of k_t decreases, lowering the threshold and making the measurement more sensitive, thereby reducing the number of false-negative measurements. In degraded conditions (k_t increases during night-time, rain, and compression artifacts), false positives due to high reconstruction error in degraded normals will be suppressed. This is a two-way adaptation that minimizes both types of errors, something that has not been obtained by any static-threshold system.

9.3 Statistical Validation

We performed pairwise McNemar’s tests to ensure that the improvements are statistically significant and not due to chance. To confirm that the improvement is statistically significant and not due to random variation, pairwise McNemar’s tests were conducted for each baseline on the UCF-Crime test set (frame-level binary predictions). The χ^2 statistic and p -value for each comparison are reported in Table 10.

Table 10. McNemar’s Test: Proposed vs. Baselines on UCF-Crime

Comparison	χ^2	p -value
Proposed vs. Conv-AE	312.4	< 0.0001
Proposed vs. C3D-AE	248.7	< 0.0001
Proposed vs. CNN-LSTM	189.3	< 0.0001
Proposed vs. GAN-AE	112.8	< 0.0001

The p -values of all the proposed improvements are well below the Bonferroni corrected significance level $\alpha' = 0.05/4 = 0.0125$, confirming statistically that the improvements proposed by the method are statistically significant. In addition, the standard deviation values reported in Table 7 (averaged over five independent runs, with different random seed: 123, 42, 256, 512, 1024) are small (AUC $\leq 0.4\%$; EER $\leq 0.4\%$), which shows that the results are stable and reproducible.

For each of the three datasets, 95% confidence intervals for AUC difference between the proposed method and the strongest baseline (GAN-AE) were computed using DeLong’s non-parametric approach, as described in Delong et al. (1988)⁹³ (Table 11). The Bonferroni-corrected level for three comparisons is $\alpha' = 0.05/3 = 0.0167$.

Table 11. DeLong 95% AUC Confidence Intervals: Proposed vs. GAN-AE (Bonferroni-corrected $\alpha' = 0.0167$)

Dataset	Proposed AUC CI	GAN-AE AUC CI	ΔAUC	Cohen’s h	p -value
UCF-Crime	[0.891, 0.903]	[0.849, 0.863]	+0.041	0.141	< 0.0001
ShanghaiTech	[0.975, 0.983]	[0.943, 0.953]	+0.031	0.128	< 0.0001
CUHK Avenue	[0.906, 0.918]	[0.881, 0.893]	+0.025	0.087	< 0.0001

All three datasets have non-overlapping confidence intervals, showing that the AUC improvements are statistically significant, even after Bonferroni correction. Cohen’s h effect sizes range from 0.087 (Avenue) to 0.141 (UCF-Crime), which shows small-to-medium practical significance alongside statistical significance. This is important for deployment in real-world applications and demonstrates that even small AUC gains correspond to meaningful reductions in missed detections within the operational false-positive budget.

9.4 Cross-Dataset Generalisation

The framework achieved the same level of improvements across the three datasets despite the different characteristics (scale, resolution, and type of anomaly), indicating the generalization ability of the proposed framework. The most difficult dataset for this task, UCF-Crime, has the greatest absolute AUC gain (+4.1% over Object-CNN), as the conditions are not strictly controlled and a wide variety of real-world anomalies are found. This is the smallest improvement for Avenue (+2.1%) because it is relatively simple to benchmark, and existing solutions are already effective and well performing in controlled environments.

9.5 Limitations

Though it has achieved good results, there are some constraints. First, it needs to be a 16-frame clip for inference, which would introduce about 0.5 seconds of latency at 30 frames per second. Secondly, the Transformer’s quadratic attention complexity makes it impractical for very long sequences; linear attention variants⁸¹ might overcome this. Third, the model was tested in a weakly supervised setting, which may yield lower performance in fully unsupervised settings. Fourth, no model exists within the current framework to describe relationships among multiple cameras, which could further boost accuracy in large-scale networks.

10 Conclusion

In this paper, a novel hybrid deep learning framework for anomaly detection of human behavior under video surveillance has been proposed. In this way, the proposed system tackles the significant problem of reducing both false positive and false negative rates that existing systems have failed to resolve by treating it as a joint task. It therefore reduces false positive and false negative rates simultaneously, which is challenging for existing systems, by synergistically combining 3D-CNN, Bi-LSTM, and Transformer architectures with an Adaptive Threshold Module. The 3D-CNN learns spatiotemporal features hierarchically, the Bi-LSTM learns long-range sequential dependencies in both directions, and the Transformer learns global contextual relationships in a single video sequence. The decision boundary is dynamically updated by the ATM using a PSNR-guided feedback loop, making it robust to different video qualities.

Experimental results for UCF-Crime, ShanghaiTech, and CUHK Avenue are shown to showcase state-of-the-art performance with mean AUC (+/- stddev over 5 runs) of 89.7%, 97.9%, and 91.2%, respectively, while reducing the EER by up to 43% compared to the best competing method¹⁵ and inferring at real-time speeds of 33 FPS. The results of the ablation studies are statistically significant for each component, as all have a p-value of less than 0.0001 (Bonferroni corrected), with the ATM showing the most significant single improvement in EER (−32.6% relative), $\chi^2 = 87.3, p < 0.0001$.

There is formal support for all four research hypotheses (H_1 – H_4) presented in Section 6 at $p < 0.0001$ when Bonferroni correction is applied. All three datasets (Table 11) show that the AUC test results for DeLong are statistically different, with Cohen’s effect sizes (0.087–0.141) showing practical significance as well. The weakly supervised training protocol is based on¹, and the rationale for learning temporal feature magnitude is based on¹⁵.

The proposed framework is a significant step towards the creation of reliable, real-time intelligent video surveillance systems applicable to critical security applications. Deployment considerations for edge and IoT environments are discussed by Gupta *et al.*⁹⁴ and Khan *et al.*⁹⁵, whose findings confirm that lightweight hybrid architectures of this type are viable for resource-constrained surveillance infrastructure.

Declarations

Author contributions statement

Vijaya kumar Koppula : Conceptualization, Data Acquisition, Interpretation of Data, Writing – Original Draft, Sameera: Software, Validation, Interpretation of Data, Syed Ali: Writing – Review and Editing. All authors have read and agreed to the published version of the manuscript.

Corresponding authors

Corresponding author is Sameera Yasam and email: srusameera@gmail.com

Funding

The authors declare that no funds, grants, or other financial support were received for the research, authorship, and/or publication of this manuscript.

Data Availability

The datasets used in this study are publicly available benchmark datasets for video anomaly detection and surveillance analysis. The UCF-Crime dataset contains 1,900 real-world surveillance videos covering 13 anomaly categories, ShanghaiTech comprises multi-scene campus surveillance videos with diverse anomalous events, and CUHK Avenue contains pedestrian surveillance videos with normal and abnormal activities. These datasets are publicly available through their respective official repositories. No new datasets were generated during this study.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Sultani, W., Chen, C. & Shah, M. Real-world anomaly detection in surveillance videos. In *Proc. IEEE/CVF CVPR*, 6479–6488 (2018).
2. Liu, W., Luo, W., Lian, D. & Gao, S. Future frame prediction for anomaly detection: A new baseline. In *Proc. IEEE/CVF CVPR*, 6536–6545 (2018).
3. Luo, W., Liu, W. & Gao, S. Remembering history with convolutional LSTM for anomaly detection. In *Proc. IEEE ICME*, 439–444 (2017).
4. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K. & Davis, L. S. Learning temporal regularity in video sequences. In *Proc. IEEE/CVF CVPR*, 733–742 (2016).
5. Zaheer, M. Z., Lee, J.-H., Astrid, M. & Lee, S.-I. Generative cooperative learning for unsupervised video anomaly detection. In *Proc. IEEE/CVF CVPR*, 14744–14754 (2022).
6. Wu, P. *et al.* Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Proc. ECCV*, 322–339 (2020).
7. Ionescu, R. T., Khan, F. S., Georgescu, M.-I. & Shao, L. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proc. IEEE/CVF CVPR*, 7842–7851 (2019).
8. Gong, D. *et al.* Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proc. IEEE/CVF ICCV*, 1705–1714 (2019).
9. Ramachandra, B., Jones, M. & Vatsavai, R. R. A survey of single-scene video anomaly detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 2293–2312 (2020).
10. Pang, G., Shen, C., Cao, L. & Hengel, A. v. d. Deep learning for anomaly detection: A review. *ACM Comput. Surv.* **54**, 1–38 (2021).
11. Mahadevan, V., Li, W., Bhalodia, V. & Vasconcelos, N. Anomaly detection in crowded scenes. In *Proc. IEEE/CVF CVPR*, 1975–1981 (2010).
12. Mehran, R., Oyama, A. & Shah, M. Abnormal crowd behavior detection using social force model. In *Proc. IEEE CVPR*, 935–942 (2009).
13. Vaswani, A. *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems*, vol. 30 (2017).
14. Chandola, V., Banerjee, A. & Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* **41**, 1–58 (2009).
15. Tian, Y. *et al.* Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proc. IEEE/CVF ICCV*, 4975–4986 (2021).
16. Duong, H.-T., Le, V.-T. & Hoang, V. T. Deep learning-based anomaly detection in video surveillance: A survey. *Sensors* **23**, 5024 (2023).
17. Nayak, R., Pati, U. C. & Das, S. K. A comprehensive review on deep learning-based methods for video anomaly detection. *Image Vis. Comput.* **106**, 104088 (2021).
18. Gupta, R., Jain, A. & Singh, D. A deep learning framework for vision-based HAR integrating CNNs, LSTMs, and LRCNs. *IEEE Access* **11**, 45621–45635 (2023).
19. Huang, C., Huang, J., Tang, W., Zhu, Y. & Xu, M. Video anomaly detection with CNN-transformer hybrid model. *IEEE Trans. Circuits Syst. Video Technol.* **33**, 3413–3424 (2023).

20. Park, H., Noh, J. & Ham, B. Learning memory-guided normality for anomaly detection. In *Proc. IEEE/CVF CVPR*, 14372–14381 (2020).
21. Cai, R., Zhang, H., Liu, W., Gao, S. & Hao, Z. Appearance-motion memory consistency network for video anomaly detection. In *Proc. AAAI*, vol. 35, 938–946 (2021).
22. Kiran, B. R., Thomas, D. M. & Parakkal, R. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *J. Imaging* **4**, 36 (2018).
23. Lv, H. *et al.* Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proc. IEEE/CVF CVPR*, 8022–8031 (2023).
24. Adam, A., Rivlin, E., Shimshoni, I. & Reinitz, D. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**, 555–560 (2008).
25. Kratz, L. & Nishino, K. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *Proc. IEEE/CVF CVPR*, 1446–1453 (2009).
26. Cong, Y., Yuan, J. & Liu, J. Sparse reconstruction cost for abnormal event detection. In *Proc. IEEE/CVF CVPR*, 3449–3456 (2011).
27. Anoop, S. & Salim, A. Survey on anomaly detection in surveillance videos. *Mater. Today: Proc.* **58**, 162–167 (2022).
28. Choudhry, N., Abawajy, J., Huda, S. & Rao, I. A comprehensive survey of machine learning methods for surveillance video anomaly detection. *IEEE Access* **11**, 114680–114713 (2023).
29. Jebur, S. A., Hussein, K. A., Hoomod, H. K., Alzubaidi, L. & Santamaría, J. Review on deep learning approaches for anomaly event detection in video surveillance. *Electronics* **12**, 29 (2022).
30. Luo, W., Liu, W. & Gao, S. Video anomaly detection with sparse coding inspired deep neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 1070–1084 (2019).
31. Rodrigues, R., Bhargava, N. & Tavares, J. M. R. Video anomaly detection with generative adversarial networks: A survey. *IEEE Access* **9**, 142120–142139 (2021).
32. Santhosh, K. K., Dogra, D. P. & Roy, P. P. Anomaly detection in surveillance videos using deep learning: A survey. *Multimed. Tools Appl.* **79**, 31929–31963 (2020).
33. Hoovayya, D. U., Kumar, J. P. & Kousar, H. Video anomaly classification using convolutional neural network with bidirectional long short-term memory using spatio-temporal adaptive transformer. *J. Image Graph.* **14**, 278–293 (2026).
34. Waghchaware, S. & Joshi, R. Hybrid CNN–Transformer model for enhanced real-time surveillance of suspicious group activities. *IEEE Access* (2026).
35. Li, S., Zhang, Y. & Chen, X. TDS-Net: Transformer-enhanced dual-stream network for video anomaly detection. *Expert. Syst. with Appl.* (2024).
36. Tran, D., Bourdev, L., Fergus, R., Torresani, L. & Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In *Proc. IEEE ICCV*, 4489–4497 (2015).
37. Ji, S., Xu, W., Yang, M. & Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 221–231 (2013).
38. Carreira, J. & Zisserman, A. Quo vadis, action recognition? A new model and the Kinetics dataset. In *Proc. IEEE/CVF CVPR*, 6299–6308 (2017).
39. Feichtenhofer, C., Fan, H., Malik, J. & He, K. SlowFast networks for video recognition. In *Proc. IEEE/CVF ICCV*, 6202–6211 (2019).
40. Hara, K., Kataoka, H. & Satoh, Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In *Proc. IEEE/CVF CVPR*, 6546–6555 (2018).
41. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
42. Schuster, M. & Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45**, 2673–2681 (1997).
43. Ullah, W., Ullah, A., Hussain, T., Khan, Z. A. & Baik, S. W. An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos. *Sensors* **21**, 2811 (2021).
44. Mukherjee, S., Singh, D. K. & Kushwaha, D. S. Deep BiLSTM attention model for spatial and temporal anomaly detection. *Multimed. Tools Appl.* **83**, 1–29 (2022).

45. Amin, A. A., Sharif, M., Raza, M. & Yasmin, M. STAD-ConvBi-LSTM: Spatio-temporal attention-based deep convolutional Bi-LSTM for abnormal activity recognition. *J. Vis. Commun. Image Represent.* **90**, 103732 (2023).
46. Li, C., Liu, Y. & Wang, H. Video anomaly detection based on spatio-temporal attention and generative adversarial networks. *IEEE Transactions on Circuits Syst. for Video Technol.* **32**, 5234–5247 (2022).
47. Liu, Y., Li, C. & Yang, D. Video anomaly detection with spatio-temporal context and contrastive learning. *IEEE Transactions on Image Process.* **31**, 5680–5693 (2022).
48. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. ICLR* (2021).
49. Liu, Z. *et al.* Video Swin Transformer. In *Proc. IEEE/CVF CVPR*, 3202–3211 (2022).
50. Deshpande, K., Pun, N. S., Sonbhadra, S. K. & Agarwal, P. Anomaly detection in surveillance videos using transformer-based attention model. *arXiv preprint arXiv:2206.01524* (2022).
51. Aslam, N. & Kolekar, M. H. TransGANomaly: Transformer based generative adversarial network for video anomaly detection. *J. Vis. Commun. Image Represent.* **100**, 104108 (2024).
52. Mishra, A. & Dhiman, C. Suspicious activities detection using spatial-temporal features based on vision transformer and recurrent neural network. *J. Ambient Intell. Humaniz. Comput.* **15**, 2853–2867 (2024).
53. Kim, H., Lee, C. H. & Hong, C. VATMAN: Video anomaly transformer for monitoring accidents and nefariousness. In *Proc. IEEE AVSS* (2024).
54. Xiao, J., Zhang, Y., Li, T. & Bao, H. A spatio-temporal transformer model for recognising abnormal human behaviour in complex scenes. In *Proc. IEEE EICT* (2024).
55. Habeb, M. H., Salama, M. & Elrefaei, L. A. Enhancing video anomaly detection using a transformer spatiotemporal attention unsupervised framework for large datasets. *Algorithms* **17**, 286, DOI: [10.3390/a17070286](https://doi.org/10.3390/a17070286) (2024).
56. Lü, H. & Sun, Q. Video anomaly detection and explanation via large language models. *arXiv preprint arXiv:2401.05702* (2024).
57. Singh, R. *et al.* Attention-guided generator with dual discriminator GAN for real-time video anomaly detection. *Eng. Appl. Artif. Intell.* **131**, 107830, DOI: [10.1016/j.engappai.2023.107830](https://doi.org/10.1016/j.engappai.2023.107830) (2024).
58. Pallear, M. G., Pawar, V. R. & Gaikwad, A. N. Human anomalous activity detection with CNN-LSTM approach. *J. Integr. Sci. Technol.* **12**, 704 (2024).
59. Kumar, A. & Patel, D. Attention-based bidirectional-long short-term memory for abnormal human activity detection. *Sensors* **23**, 7573 (2023).
60. Chen, B., Meng, F., Tang, H. & Tong, G. Two-level attention module based on spurious-3D residual networks for human action recognition. *Sensors* **23**, 1707 (2023).
61. Gandapur, M. Q. & Verdú, E. Video anomaly detection system using deep convolutional and recurrent models. *Results Eng.* **18**, 101026, DOI: [10.1016/j.rineng.2023.101026](https://doi.org/10.1016/j.rineng.2023.101026) (2023).
62. Kommanduri, R. STAD-AI: Spatio-temporal anomaly detection in videos with attentive dual-stage integration. *Neurocomputing* (2025).
63. Biradar, K. M., Mandal, M., Dube, S., Vipparthi, S. K. & Tyagi, D. K. Triplet-set feature proximity learning for video anomaly detection. *Image Vis. Comput.* **150**, 105205, DOI: [10.1016/j.imavis.2024.105205](https://doi.org/10.1016/j.imavis.2024.105205) (2024).
64. Zahid, Y., Zarges, C., Tiddeman, B. & Han, J. Adversarial diffusion for few-shot scene adaptive video anomaly detection. *Neurocomputing* **614**, 128796, DOI: [10.1016/j.neucom.2024.128796](https://doi.org/10.1016/j.neucom.2024.128796) (2025).
65. Yang, D., Liu, Y. & Wang, Y. Video anomaly detection with transformer and adversarial learning. *IEEE Transactions on Multimed.* **25**, 7890–7902 (2023).
66. Yu, G., Wang, S., Cai, Z. & Zhu, E. Cloze test helps: Effective video anomaly detection via learning to complete video sequences. *IEEE Transactions on Image Process.* **32**, 1356–1368 (2023).
67. Zhang, X., Yang, D. & Liu, Y. Video anomaly detection with temporal-spatial attention and memory-augmented networks. *Pattern Recognit.* **137**, 109307 (2023).
68. Zhong, J. *et al.* Graph convolutional network-based video anomaly detection with spatio-temporal feature learning. *Knowledge-Based Syst.* **284**, 111231 (2024).
69. Zhou, J. T. *et al.* AnomalyNet: An anomaly detection network for video surveillance. *IEEE Transactions on Inf. Forensics Secur.* **16**, 3153–3165 (2021).

70. Zuo, Y. & Zhu, S. Video anomaly detection based on multi-scale feature fusion and attention mechanism. *Eng. Appl. Artif. Intell.* **127**, 107281 (2024).
71. Zhong, J.-X. *et al.* Graph convolutional label noise cleaner for anomaly detection. In *Proc. IEEE/CVF CVPR*, 1237–1246 (2019).
72. Feng, J.-C., Hong, F.-T. & Zheng, W.-S. MIST: Multiple instance self-training framework for video anomaly detection. In *Proc. IEEE/CVF CVPR*, 14009–14018 (2021).
73. Pu, Y., Wu, X., Yang, L. & Wang, S. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Trans. Image Process.* **33**, 4923–4936 (2024).
74. Qin, F. & Xiao, Y. Multi-instance learning anomaly event detection based on transformer. In *Proc. AIPR* (2022).
75. Su, Y., Tan, Y., An, S. & Xing, M. Weakly-supervised video anomaly detection. *Expert. Syst. with Appl.* (2024).
76. Su, Y., Tan, Y., Xing, M. & An, S. Weakly supervised video anomaly detection (WSVAD) in visual surveillance. *Knowledge-Based Syst.* (2024).
77. Boekhoudt, K., Matei, A., Aghaei, M. & Talavera, E. HR-Crime: Human-related anomaly detection in surveillance videos. In *arXiv preprint arXiv:2108.00246* (2021).
78. Yan, S., Xiong, Y. & Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proc. AAAI*, vol. 32 (2018).
79. Shi, L., Zhang, Y., Cheng, J. & Lu, H. Skeleton-based action recognition with directed graph neural networks. In *Proc. IEEE/CVF CVPR*, 7912–7921 (2019).
80. Chen, X. *et al.* Multiscale spatial temporal attention graph convolution network for skeleton-based anomaly behavior detection. *J. Vis. Commun. Image Represent.* **90**, 103725 (2023).
81. Wang, X., Girshick, R., Gupta, A. & He, K. Non-local neural networks. In *Proc. IEEE/CVF CVPR*, 7794–7803 (2018).
82. Zhu, Y. & Newsam, S. Video anomaly detection for surveillance: A survey of methods and datasets. *ACM Comput. Surv.* **56**, 1–38 (2023).
83. Wang, S. & Wu, Z. A survey on video anomaly detection in surveillance: Deep learning methods. *IEEE Transactions on Artif. Intell.* **4**, 456–473 (2023).
84. Patrikar, D. R. & Parate, M. R. Anomaly detection using edge computing in video surveillance systems: A review. *Complex & Intell. Syst.* **8**, 85–110 (2022).
85. Chandrakala, S., Deepak, K. & Revathi, G. A survey on unsupervised and semi-supervised video anomaly detection. *Multimed. Tools Appl.* **81**, 24543–24571 (2022).
86. Ramachandra, B., Jones, M. J. & Vatsavai, R. R. A survey of single-scene video anomaly detection. *IEEE Transactions on Pattern Analysis Mach. Intell.* **44**, 2293–2312 (2020).
87. Feng, X. *et al.* TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection. *Eng. Appl. Artif. Intell.* **123**, 106173 (2023).
88. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE/CVF CVPR*, 770–778 (2016).
89. Lu, C., Shi, J. & Jia, J. Abnormal event detection at 150 FPS in MATLAB. In *Proc. IEEE ICCV*, 2720–2727 (2013).
90. Wu, P. *et al.* Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Proceedings of the European Conference on Computer Vision*, 685–701 (2020).
91. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In *Proc. ICLR* (2015).
92. Medel, J. R. & Savakis, A. Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv preprint arXiv:1612.00390* (2016).
93. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44**, 837–845, DOI: [10.2307/2531595](https://doi.org/10.2307/2531595) (1988).
94. Gupta, R., Gupta, C., Rathore, N. & Mishra, G. A deployment-oriented hybrid CNN–LSTM–MIL system for real-world video anomaly detection. *Informatica* **50** (2026).
95. Khan, M., Rehman, A., Zubair, M. & Kim, B. Artificial intelligence of things-assisted two-stream neural network for anomaly detection in surveillance big video data. *Futur. Gener. Comput. Syst.* **129**, 286–297 (2022).