IRSTI 28.23.15

https://doi.org/10.51889/2959-5894.2026.94.2.013

D. Yermekova [1*], A. Bykov[1], B. Tokanova[1], D. Nauryzbayev[2]

[1] International Information Technology University, Almaty, Kazakhstan
[2]Abai Kazakh National Pedagogical University, Almaty, Kazakhstan
[*]*e-mail: d.yermekova@iitu.edu.kz*

COMPARATIVE ANALYSIS OF BEHAVIORAL VIDEO ANALYTICS APPROACHES BASED ON MACHINE LEARNING

Abstract

This review systematizes and analyzes modern approaches to the intelligent detection of anomalies in human behavior based on deep learning in video surveillance systems. The work explores key methods, including hybrid architectures, generative models, and multimodal approaches. The main purpose of the study is to identify the key limitations of existing solutions and propose ways to overcome them by developing a new conceptual architecture. The analysis showed that modern models achieve high accuracy (F1-score in the range of 90-95%) on standard datasets, but face three fundamental problems: a lack of labeled anomaly data, high computational complexity that prevents real-time operation on edge devices, and low reliability with external interference. To solve these problems, a hybrid multimodal architecture is proposed that uses compressed-domain analysis to optimize the speed of inference and a Gated Cross-Attention mechanism for intelligent merging of video and audio streams. The proposed architecture demonstrates the potential for creating a reliable, scalable and proactive monitoring system.

Keywords: deep learning, video analysis, behavior analysis, computer vision, generative models, self-supervised learning, multimodal learning.

Д. Ермекова [1], А. Быков[1], Б. Токанова[1], Д. Наурызбаев[2]
[1] Халықаралық Ақпараттық Технологиялар Университеті, Алматы қ., Қазақстан
[2]Абай атындағы Қазақ ұлттық педагогикалық университеті, Алматы қ., Қазақстан

МАШИНАЛЫҚ ОҚЫТУҒА НЕГІЗДЕЛГЕН МІНЕЗ-ҚҰЛЫҚ БЕЙНЕ АНАЛИТИКАСЫНЫҢ ТӘСІЛДЕРІН САЛЫСТЫРМАЛЫ ТАЛДАУ

Аңдатпа

Бұл шолу бейнебақылау жүйелерінде терең оқыту негізінде адамның мінез-құлқындағы ауытқуларды интеллектуалды анықтаудың заманауи тәсілдерін жүйелейді және талдайды. Жұмыс гибридті архитектураларды, генеративті модельдерді және мультимодальды тәсілдерді қоса алғанда, негізгі әдістерді зерттейді. Зерттеудің негізгі мақсаты-қолданыстағы шешімдердің негізгі шектеулерін анықтау және жаңа тұжырымдамалық архитектураны әзірлеу арқылы оларды жеңу жолдарын ұсыну. Талдау көрсеткендей, қазіргі заманғы модельдер стандартты деректер жиынтығында жоғары дәлдікке (F1-score 90-95% диапазонында) жетеді, бірақ үш негізгі мәселеге тап болады: белгіленген аномалия деректерінің тапшылығы, edge құрылғыларында нақты уақыт режимінде жұмыс істеуге кедергі келтіретін жоғары есептеу күрделілігі және сыртқы кедергілерде төмен сенімділік. Осы мәселелерді шешу үшін гибридті мультимодальды архитектура ұсынылған, ол қысылған көріністегі талдауды (компрессияланған-домендік талдау) инференс жылдамдығын оңтайландыру үшін және бейне және аудио ағындарды интеллектуалды біріктіру үшін кросс-зейін механизмін (Gated Cross - Attention) қолданады. Ұсынылған архитектура сенімді, масштабталатын және белсенді бақылау жүйесін құрудың әлеуетін көрсетеді.

Түйін сөздер: терең оқыту, бейнені талдау, мінез-құлықты талдау, компьютерлік көру, генеративті модельдер, өзін-өзі басқаратын оқыту, мультимодальды оқыту.

Д. Ермекова[1], А. Быков[1], Б. Токанова[1], Д. Наурызбаев[2]
[1] Международный Университет Информационных Технологий, г.Алматы, Казахстан
[2] Казахский национальный педагогический университет имени Абая, Алматы, Казахстан

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПОДХОДОВ ПОВЕДЕНЧЕСКОЙ ВИДЕОАНАЛИТИКИ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ

Аннотация

Данный обзор систематизирует и анализирует современные подходы к интеллектуальному обнаружению аномалий в поведении человека на основе глубокого обучения в системах видеонаблюдения. Работа исследует ключевые методы, включая гибридные архитектуры, генеративные модели и мультимодальные подходы. Основная цель исследования — определить ключевые ограничения существующих решений и предложить пути их преодоления путём разработки новой концептуальной архитектуры. Анализ показал, что современные модели достигают высокой точности (F1-score в диапазоне 90–95%) на стандартных датасетах, но сталкиваются с тремя фундаментальными проблемами: дефицит размеченных данных об аномалиях, высокая вычислительная сложность, препятствующая работе в реальном времени на edge-устройствах, и низкая надёжность при внешних помехах. Для решения этих проблем предложена гибридная мультимодальная архитектура, использующая анализ в сжатом представлении (compressed-domain analysis) для оптимизации скорости инференса и механизм перекрёстного внимания (Gated Cross-Attention) для интеллектуального слияния видео- и аудиопотоков. Предложенная архитектура демонстрирует потенциал для создания надёжной, масштабируемой и проактивной системы мониторинга.

Ключевые слова: глубокое обучение, анализ видео, анализ поведения, компьютерное зрение, генеративные модели, самоконтролируемое обучение, мультимодальное обучение.

Introduction

Ensuring safety in public places, industrial facilities, transport hubs and educational institutions is one of the highest priorities of modern society. Traditional video surveillance systems based on visual control require constant human involvement to analyze potentially dangerous situations. However, with the exponential growth of video data volumes, manual monitoring is becoming not only labor-intensive, but also practically impossible. The human factor, fatigue, and inattention significantly increase the likelihood of errors and delays in responding to incidents [1-3].

In the context of video analytics, abnormal behavior is defined as "the manifestation of atypical visual or motor characteristics, or the presence of typical visual or motor patterns that manifest themselves in spatiotemporal contexts that deviate from established norms". Anomalies by their nature are extremely rare, unpredictable, and can include both unusual actions (such as fighting) and normal actions in an unsuitable environment (such as cycling on a sidewalk) [4].

To solve these problems, deep learning and intelligent video analysis methods are actively being developed to automate the detection of abnormal behavior and significantly improve the effectiveness of monitoring [5,6].

The main research areas in this field include:

- Video preprocessing to improve the quality of frames and increase the information content of the data [7].
- The use of deep neural networks capable of analyzing complex spatiotemporal behavior [5,8].
- The use of generative methods (GAN, MNAD, AOG) to synthesize rare anomalies and improve model learning [9,10].
- Development of self - supervised learning and multimodal approaches that allow working with non-labeled data and combining information from various sources - video, audio and sensors [11,12].

The relevance of this research is supported by global digitalization strategies. The need for active implementation of artificial intelligence and machine learning technologies in all spheres of activity is emphasized at the highest state level in different countries, as this is a key factor for building a "smart state" and ensuring the safety of citizens [1,13].

In particular, at the state level in Kazakhstan, as in many other countries, it is officially recognized that the development of artificial intelligence and machine learning is a strategic priority for ensuring national security and economic progress. Thus, the systematization of existing approaches in the intelligent monitoring of abnormal human behavior is of particular strategic importance.

The purpose of the study is to systematize existing approaches to intelligent monitoring of abnormal human behavior, identify their strengths and weaknesses, and propose an architecture for recognizing abnormal actions with the highest accuracy.

Research methodology

Studying and detecting abnormal human behavior in a video stream is one of the most urgent tasks in the field of computer vision and deep learning. An analysis of the existing literature and methods shows that research is actively developing in several key areas.

Evaluation methods, datasets, and metrics

Standardized CUHK-Avenue datasets [5] and metrics are used to evaluate the effectiveness of the models. The estimation is based on the Accuracy, Precision, Recall, and F1-score metrics, with the F1-score considered the most balanced metric for tasks with unbalanced data. Their definitions are based on the basic indicators (Table 1) of true/false positive and negative results (TP, TN, FP, FN) and are common to all classification tasks in machine learning (Pedregosa et al., 2011). Metrics are calculated at the episode level, not individual frames: an episode is counted as detected if at least one trigger falls within its time limits.

Table 1. Key metrics for determining the effectiveness of models [14]

Metric	*Formula*	*Interpretation for analyzing people's behavior on video*
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	*The percentage of correctly classified observation units. If there is a pronounced imbalance of classes, the information content is limited. For rare anomalies, Recall and F1 should always be accompanied for correct interpretation*
Precision	$\frac{TP}{TP + FP}$	*The proportion of truly positive episodes among all detections of the system; high values correspond to a low frequency of false positives*
Recall	$\frac{TP}{TP + FN}$	*The proportion of truly positive episodes among all annotated anomalous episodes; high values correspond to a low frequency of false negative omissions.*
F1-score	$2 * \frac{Precision * Recall}{Precision + Recall}$	*Harmonic mean of Precision and Recall; characterizes the balance between false positive and false negative detections on unbalanced samples*

When evaluating binary classification, especially in conditions where data is highly unbalanced (as in the case of anomalies), the traditional Accuracy metric can be deceptive. Instead, the Area Under the Receiver Operating Characteristic Curve (AUC) is often used for an objective assessment.

The Receiver Operating Characteristic (ROC) curve shows a compromise between the proportion of correctly detected True Positive Rate (TPR) anomalies and the proportion of False Positive Rate (FPR) false positives. AUC is the area under this curve, which represents the probability that the model correctly ranks a randomly selected positive example higher than a negative one.

AUC (ROC) is a key metric in anomaly detection, as it:

- It is resistant to unbalanced data. It is not affected by the ratio of normal and abnormal events.
- Independent of the classification threshold, which allows you to evaluate the overall performance of the model.

Research shows (Jing Li, 2025) that AUC provides the most stable and consistent assessment of model quality, making it an ideal choice for tasks where anomalies are extremely rare [15].

Methods based on deep neural networks

Hybrid architectures combining CNN and recurrent neural networks (RNN) or Long Short-Term Memory (LSTM) are the basis of modern anomaly detection systems. In such models, CNNs are responsible for extracting spatial features, while RNNs or LSTMs analyze them in a temporal sequence, revealing dependencies and non-standard movement patterns. These approaches demonstrate high performance. For example, in Saket et al. (2025), the hybrid model showed an accuracy of 90% to 94%, and Wang et al. (2025), they achieved an F1 score from 92% to 95%, which confirms the high efficiency of spatiotemporal analysis. As its shown in Figure 1 the use of architectures such as MRCNN+LSTM in the study by Manju et al. (2025) made it possible to achieve an accuracy of 93.6% on the UCF Crime dataset, which makes them suitable for early detection of anomalies [2,6,8].

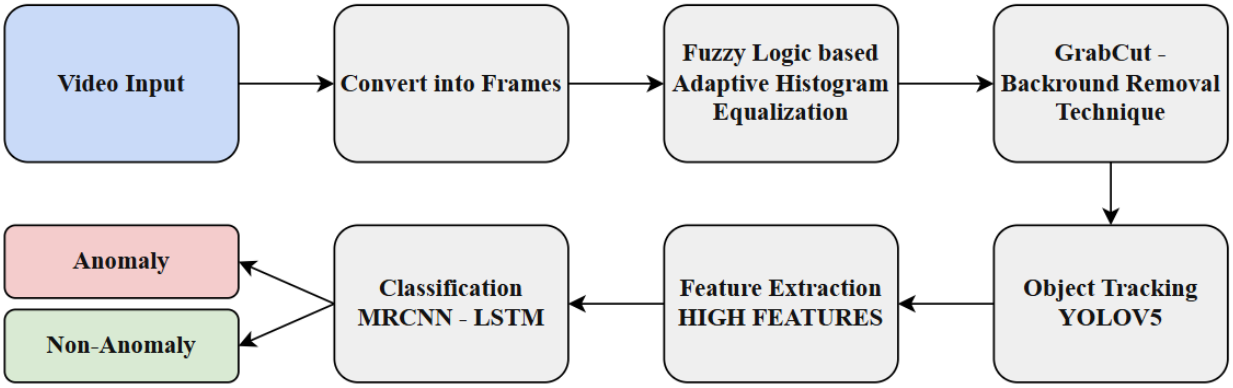

Figure 1. Schematic representation of the MRCNN-LSTM hybrid architecture for early anomaly detection [2]

Classical generative adversarial networks

One of the main problems in detecting anomalies is their rarity. Generative models such as Generative Adversarial Networks (GANS) solve this problem by synthesizing artificial data. The review by Saeed et al. (2025) shows that GAN models can achieve accuracy from 88% to 94% on the ShanghaiTech and CUHK datasets. Figure 2 shows the classic GAN architecture, where the generator creates "fake" anomalies, and the discriminator tries to distinguish them from the real ones [9].

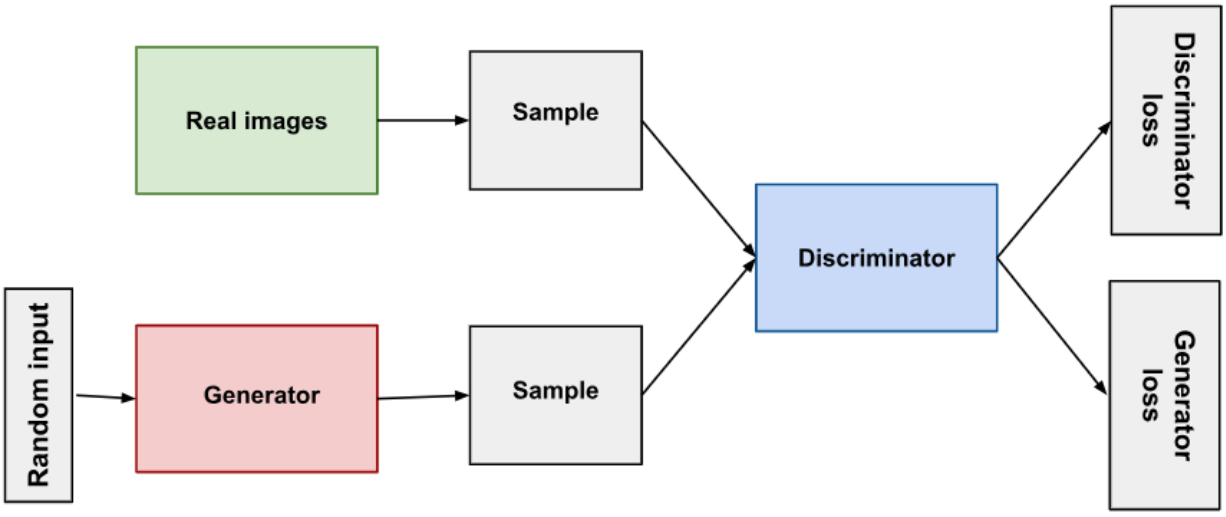

Figure 2. Classical Generative Adversarial Network (GAN) architecture for Anomaly Synthesis [16]

FutureGAN Architecture

GANs can be applied not only for anomaly detection. Their core principle is based on predicting future video frames. The model is trained on a large dataset containing only normal behavior and learns to forecast the subsequent frames in a sequence. A notable example of this approach is the FutureGAN model introduced by Aigner and Körner (2018). This architecture employs spatiotemporal 3D convolutions to capture both the static features of each frame and the temporal dynamics of motion. When an abnormal event occurs in a test video, such as a fall or a fight, the model fails to accurately predict it because it deviates from the learned "normal" patterns. In such cases, a high prediction error serves as a key indicator of anomaly (Figure 3).

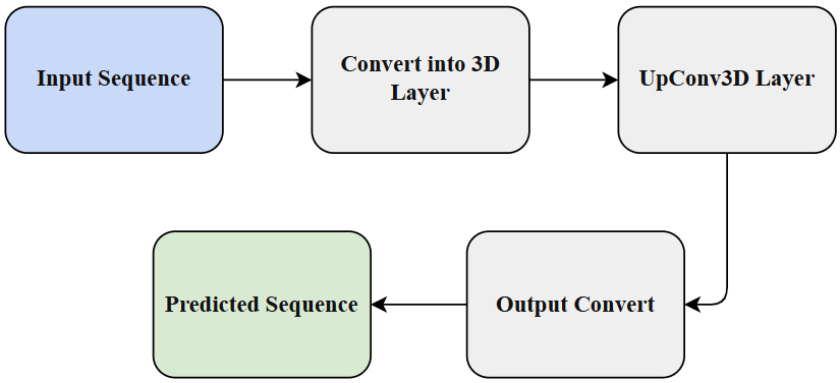

Figure 3. FutureGAN generator during training [17]

In addition, there are specialized GAN architectures designed specifically for anomaly detection, such as AnoGAN and its faster variant f-AnoGAN. These models are trained on normal data and use the learned network to identify events that cannot be accurately generated or mapped into the latent space.

Variational Autoencoders (VAEs)

As an alternative to GANs, Variational Autoencoders (VAEs) are often used for anomaly detection. They are trained to compress “normal” data into a latent space and reconstruct it. When an anomalous event is fed into the model, it fails to accurately reconstruct it, and a high reconstruction error serves as a signal of anomaly. This approach has been extensively discussed in several studies [17]. The application of VAEs is based on the principle of reconstruction. A VAE consists of an encoder, which compresses the data into a latent representation, and a decoder, which reconstructs it.

- For data synthesis: VAEs can generate new, realistic frames by sampling random vectors from the latent space and reconstructing them through the decoder.
- For anomaly detection: The model is trained exclusively on normal data. During testing, it attempts to reconstruct a new frame. If the frame is normal, the reconstruction error remains low. However, if the frame contains an anomalous event, the VAE fails to accurately reconstruct it, resulting in a significantly higher reconstruction error. This error serves as an indicator for anomaly detection (Figure 4).

Thus, generative models provide an effective solution to the problem of limited labeled data, offering powerful and versatile tools for anomaly detection based on the concepts of prediction and reconstruction (Kingma & Welling, 2019) [18].

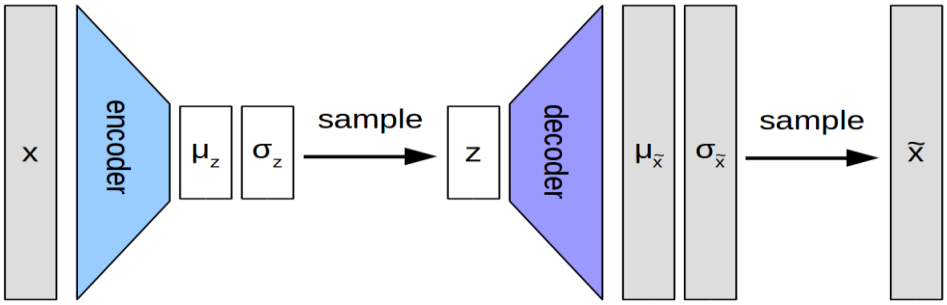

Figure 4. VAE Architecture

Memory-Augmented Autoencoders

The models proposed by Chu et al. (2025), which utilize Memory-Augmented Autoencoders (MNAD) and And-Or Graphs (AOG), achieved an F1-score ranging from 91% to 93% on the Ped2 and Avenue datasets. This demonstrates that, despite their computational complexity, generative methods effectively enhance the robustness of models to novel and rare scenarios, which is critically important for security systems [9, 10].

Unlike the classical autoencoder, the architecture of the Memory-Augmented Autoencoder (MemAE) includes a memory block that stores prototypes of normal behavior. When an anomalous frame is input to the network, it does not match any of these prototypes, resulting in a high reconstruction error (Figure 5). This mechanism enables effective differentiation between anomalies and normal events [10].

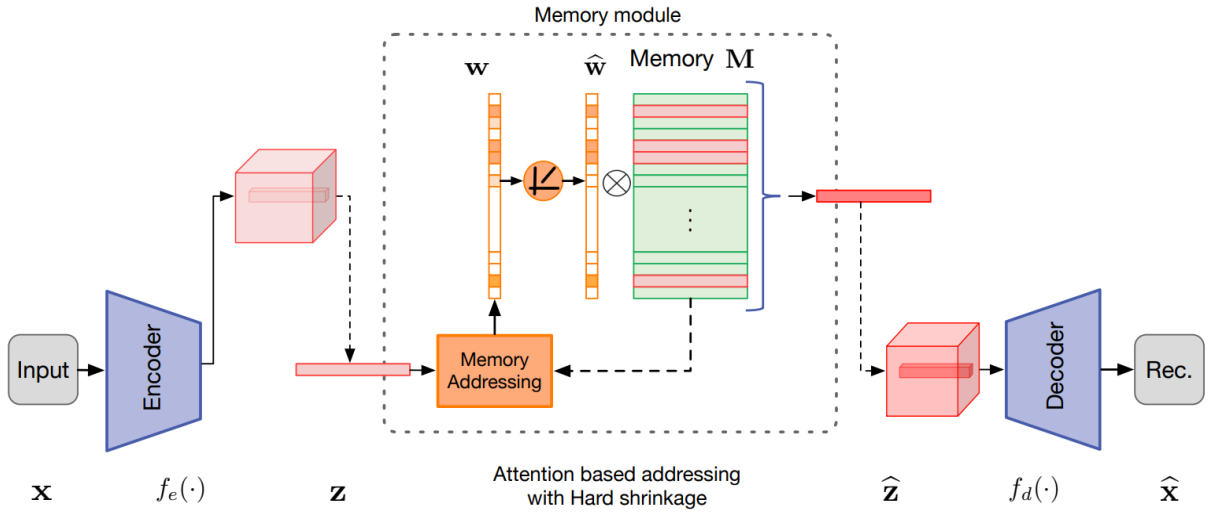

Figure 5. Memory-Augmented Autoencoder Architecture Scheme for Anomaly Detection [9]

Self-supervised and multimodal learning

Self-supervised learning enables models to learn from unlabeled data by using temporal and contrastive features as “surrogate tasks” [5, 11]. This approach demonstrates high effectiveness, as evidenced by an F1-score ranging from 91% to 92% in the study by Liu et al. (2025). Another promising direction is the use of multimodal approaches, which combine information from multiple sources, such as video and audio. In the work of Barbosa et al. (2025), this approach achieved an F1-score between 88% and 90% on the CUHK dataset, improving system reliability by leveraging independent signals [11, 12]. As illustrated in Figure 6, such multi-stream architectures further enhance the robustness and reliability of the system [19].

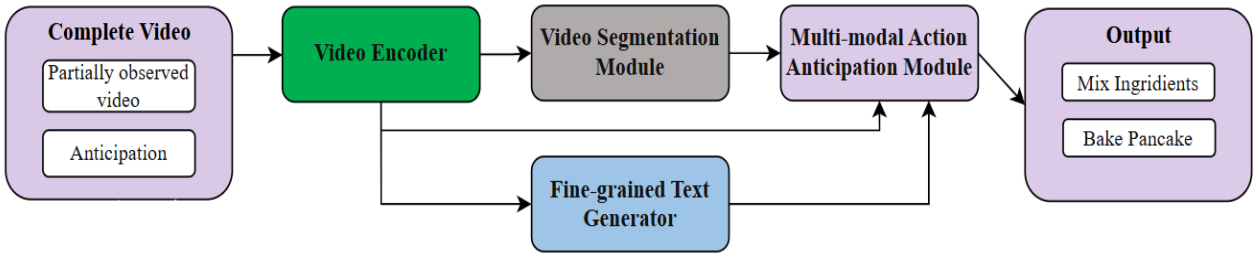

Figure 6. Generalized scheme of multithreaded architecture for video stream analysis [19]

Recent studies, such as those involving large multimodal models, show that video stream analysis becomes even more effective when integrating information from various sources, including audio. This allows models not only to recognize visual patterns but also to understand the context of the situation, which significantly increases the accuracy of anomaly detection [20].

For correct multimodal integration of features (video and audio) in anomaly detection systems, it is critically important to ensure their precise temporal alignment (synchronization). The need for this procedure is caused by the following technical differences.

Difference in sampling rate (Rate Disparity). The video stream is processed at a low sampling rate measured in frames per second (FPS), usually 25–30 Hz. The audio stream has a significantly higher sampling rate measured in kilohertz (kHz), for example, 44.1 kHz or 48 kHz (44,100 or 48,000 samples per second).

Consequence: Without synchronization, one video frame may correspond to thousands of audio samples. If their timestamps do not match, the model will incorrectly associate visual and sound events.

Thus, synchronization is a necessary prerequisite, without which the Gated Cross-Attention and Temporal Alignment mechanisms in the integration module cannot operate correctly, which would lead to a significant decrease in reliability and an increase in false detections.

In addition to self-supervised learning, the approach of weakly supervised learning is actively developing, which uses a minimal amount of labeled data to generate so-called pseudo-labels. As shown in the work of Zhang et al. (2023), the use of these pseudo-labels allows training models more effectively, significantly reducing the need for manual labeling and improving the accuracy of anomaly detection, which is especially relevant for imbalanced datasets [21].

Comparative Analysis of Methods

For a clearer representation of the analysis results, Table 2 presents a comparative analysis of the aforementioned methods (deep neural networks, hybrid neural networks, generative, self-supervised, and multimodal methods). The comparison is based on their key principles, main advantages and disadvantages, as well as optimal application areas, which enables a well-founded choice for the development of a new architecture.

Table 2. Comparative Analysis of Intelligent Monitoring Methods for Anomalous Behavior

Method Type	*Key point*	*Main advantages*	*Main disadvantages*	*Optimal usage*
Deep Neural Networks (Hybrid CNN+ LSTM/RNN)	*Spatio-temporal analysis: CNN extracts spatial features, while LSTM/RNN analyzes their temporal dynamics [2].*	*High accuracy in recognizing specific actions; Capability for early anomaly detection [4]*	*Require large and well-annotated datasets; Low interpretability of results [2]*	*Detection and classification of known anomalies (fights, falls) in real time [4]*

Generative Methods (GAN, MemAE)	*Synthesis of artificial anomalous data for training or anomaly detection by measuring deviations from "normality" [9], [10].*	*Effective solution to the data scarcity problem; Improved model robustness to new and rare types of anomalies [10]*	*High computational complexity; May be sensitive to changes in background and lighting [9]*	*Data augmentation for model training; Detection of anomalies not represented in the training dataset [22].*
Self-Supervised and Multimodal Methods (Multi-stream, Self-supervised)	*Training on unlabeled data; Fusion of information from multiple sources (video, audio, sensors) [12]*	*Reduced dependence on manual annotation; Increased system reliability and versatility [23].*	*Limited testing on real-world streams; Data privacy issues (especially with audio) [5]*	*Training on large unlabeled datasets; Development of fault-tolerant monitoring systems [5]*

The analysis of the table shows that each of the methods considered has its significant limitations:

- High dependence on labeled data. This is the main drawback of hybrid architectures (CNN+RNN), which require enormous amounts of manually annotated data to achieve high accuracy. This is costly, time-consuming, and not always feasible, as anomalous events are extremely rare.
- Computational complexity. Both hybrid and generative models, such as GANs, demand powerful hardware and high computational resources. This makes them unsuitable for real-time operation on "smart" cameras (edge devices), where fast data processing is required.
- Limited generalization and contextual dependency. Many models trained on specialized datasets (e.g., custom campus datasets) perform poorly in other environments (outdoors, in transport), as they do not account for variations in surroundings, lighting, or motion.
- Insufficient utilization of information. Traditional models focus solely on video data, ignoring other potentially useful signals such as audio. This makes them vulnerable to errors and false positives

The drawbacks of different methods are often interconnected. For example, the data scarcity issue typical of hybrid models can be addressed using generative methods, while the lack of manual annotation can be mitigated through self-supervised learning.

Results of the study

The analysis of existing solutions (Table 2) revealed their key limitations, including high dependence on labeled data and computational complexity. The main idea is not to oppose these methods but to combine them into a unified architecture. The proposed system is based on this principle and offers the following improvements:

1. Integration of generative models. To address the data scarcity problem, generative models will be used to synthesize pseudo-anomalies. This allows training the hybrid part of the architecture on a more balanced and diverse dataset, increasing its accuracy and generalization without the need for manual annotation.
2. Multimodal fusion. This approach involves using not only video but also other information sources, such as audio (e.g., screams). This makes the system more robust, as it relies on multiple independent signals.

Thus, the limitations of each existing method become the starting point for creating a more advanced and comprehensive solution that overcomes these constraints.

Proposed Hybrid Multimodal Architecture

Based on the identified limitations from the comparative analysis, a hybrid multimodal architecture is proposed for intelligent monitoring of anomalous behavior. It combines best practices from existing solutions and includes the following key components:

1. Data Augmentation Module (Generative Methods): Used exclusively during the training phase. Generative models (MemAE) create pseudo-anomalies to overcome data scarcity, improving model robustness to rare scenarios. MemAE was chosen because it is expected to achieve high F1 scores compared to other autoencoders. Thus, using generative models effectively addresses the lack of labeled data.

2. Preprocessing & Synchronization Module: Performs basic data processing (normalization, noise reduction) and crucial temporal alignment (A/V Timestamp Sync) to ensure proper fusion of data streams.

3. Multi-stream Analysis Module: To enable real-time operation, compressed-domain analysis is used instead of computationally expensive optical flow. This allows extraction of motion features (e.g., H.264 motion vectors) without full decoding, providing high inference speed.

- In H.264 video, frames are encoded at the macroblock level (typically 16×16 pixels), containing built-in motion and texture information usable without decoding all frames.
- Compressed-domain analysis leverages existing H.264 features – motion vectors, macroblock types, and residual coding errors – which reflect scene dynamics and object movement.
- Unlike optical flow, these features do not require full decoding or heavy computation per frame, reducing time and resource costs while maintaining motion information.
- This approach ensures high inference speed, lower energy consumption, and sufficient accuracy for anomaly detection and behavior analysis.

4. Multi-stream Integration Module: Responsible for intelligent feature fusion:

- Temporal Alignment: Aligns features across time.
- Gated Cross-Attention: An attention mechanism that allows streams to influence each other selectively.
- Feature Fusion / Projection: Final merging and projection of features into a unified multimodal space.

5. Decision-Making Module: Based on an LSTM classifier, this component analyzes the fused features and produces the final anomaly decision. Additionally, reconstruction error from MemAE can be used as a complementary evaluation method. This architecture leverages generative models to handle data scarcity, compressed-domain analysis for efficiency, and multimodal fusion to improve robustness, creating a comprehensive solution for real-time anomaly monitoring.

Visualization and Explanation of the Architecture. Figure 7 shows the diagram of the proposed hybrid multimodal architecture. It is designed to overcome the key limitations identified during the review: data scarcity, computational complexity, and insufficient reliability.

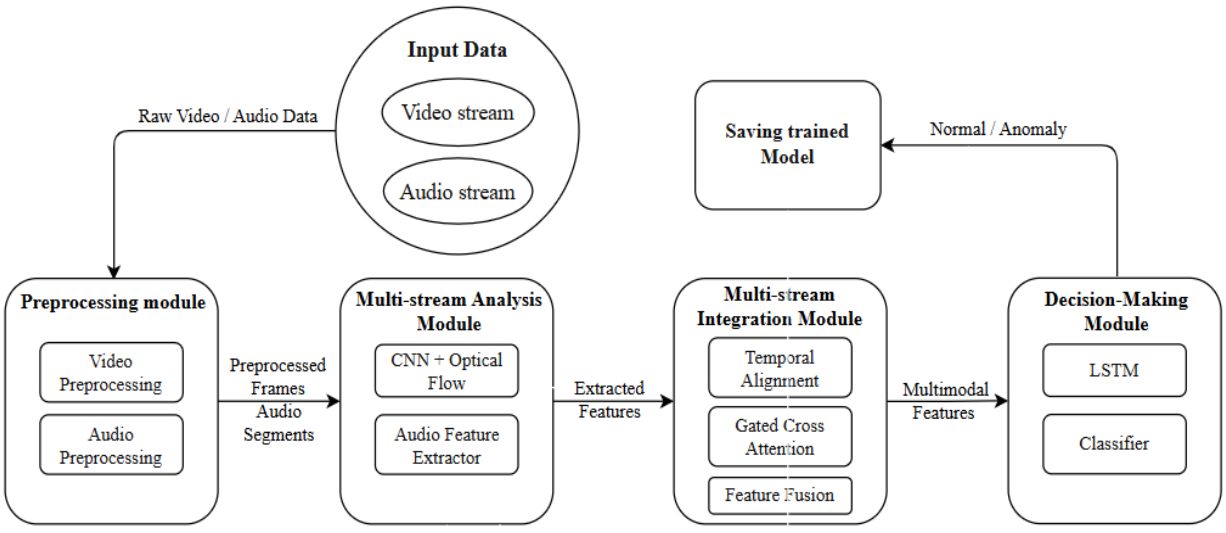

Figure 7. Model Training Architecture

The architecture operates by separating the logic between the training and inference stages, while also implementing optimized multi-stream analysis:

1. Preprocessing and Synchronization

This module ensures that the data is ready for fusion:

- A/V Timestamp Sync: Critical synchronization of video and audio timestamps is performed, which serves as the basis for subsequent accurate multimodal integration.

2. Multi-Stream Analysis Module

To enable real-time processing (a requirement for edge devices), optimization is implemented in this module:

- Compressed-domain Analysis: Instead of computationally expensive full-stream decoding and optical flow calculation, the model uses compressed-domain analysis (for instance, motion vectors from H.264 containers). This significantly reduces computational costs and increases inference speed.
- Audio Feature Extractor: Extracts high-level features from the audio stream.

3. Multimodal Integration Module

This module performs intelligent feature fusion to enhance system reliability:

- Temporal Alignment: Aligns features coming from different frame rates (video and audio) to ensure accurate temporal correspondence of events.
- Gated Cross-Attention: An attention mechanism that allows each stream (for instance, audio) to influence the processing of another (e.g., video) in a weighted manner. This helps focus on the most informative segments.
- Feature Fusion / Projection: Combines the aligned and weighted features into a unified multimodal space, which is then fed to the classifier.

4. Decision-Making Module

- LSTM & Classifier: The component, based on an LSTM network, analyzes the fused multimodal features over the temporal sequence and produces the final anomaly decision.

To evaluate the performance and applicability of the proposed architecture in real-world conditions, an optimized inference pipeline was developed, shown in Figure 8. This diagram reflects the system's operation after the training phase and generative data augmentation.

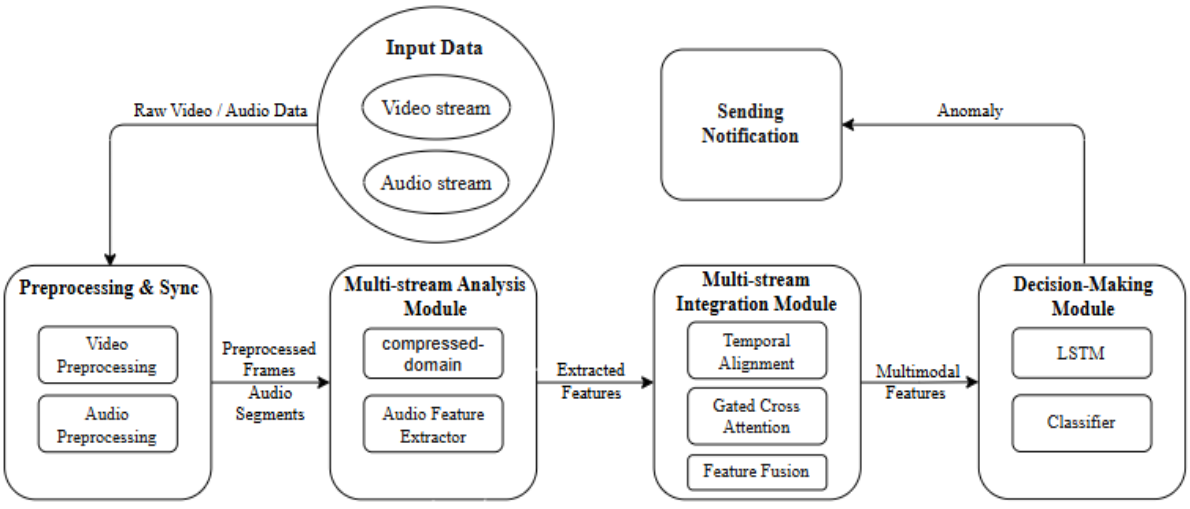

Figure 8. Optimized Inference Pipeline of the Multimodal Architecture

A key requirement for the inference architecture is real-time data processing on low-power devices (edge devices). To achieve this, special optimization methods were applied in the preprocessing and analysis modules. Video and audio stream synchronization was also implemented.

1. Preprocessing and Synchronization

This module is responsible for preparing raw multimodal data and performing temporal alignment:

- Video Preprocessing: Performs standard operations such as normalization, resizing, and denoising.
- Audio Preprocessing: Provides filtering and preparation of audio data for feature extraction.
- A/V Timestamp Sync: A critical component that synchronizes the timestamps of video and audio streams. This ensures that multimodal integration matches features corresponding to the same moment in time, significantly enhancing reliability.

2. Multi-stream Analysis Module

This module provides fast and efficient analysis without excessive computational overhead:

- Compressed-domain Analysis: It is a key element for real-time optimization. Instead of computationally expensive full video stream decoding, the model uses a compressed data representation. It works directly with existing features, such as motion vectors from H.264 containers. This avoids full decoding and significantly saves computational resources and memory, making the system suitable for edge devices.
- Audio Feature Extractor: Extracts low-level or high-level features from the prepared audio segments.

3. Multi-stream Integration Module

This module is responsible for the intelligent and reliable fusion of the extracted features:

- Temporal Alignment: Ensures precise temporal alignment of features coming from different frequencies (for instance, video frames and audio features).
- Gated Cross-Attention: Allows one stream (for instance, audio capturing a scream) to influence another stream (video) in a weighted manner, highlighting the features most important for detecting anomalies.
- Feature Fusion: Combines the aligned and weighted features into a single multimodal feature vector.

4. Decision-Making Module

- LSTM & Classifier: The LSTM network analyzes the sequence of fused multimodal features and produces the final decision. The classifier determines whether an event is normal or anomalous, taking the context into account.

5. Sending Notification

In case an anomaly is detected, the system immediately generates a signal (Anomaly), which triggers an alert to the operator or the relevant services.

Overall, the proposed inference architecture demonstrates a balance between high reliability, provided by multimodality and attention mechanisms, and practical applicability, achieved through compressed-domain analysis and optimization for edge devices.

Discussions

The proposed hybrid multimodal architecture is designed to overcome the key limitations of traditional systems, emphasizing real-time efficiency and increased reliability.

The main difference from standard hybrid architectures (CNN + Optical Flow) lies in the implementation of two main groups of innovations. First, there is the optimization of the data pipeline for edge devices. Instead of computationally expensive full video stream decoding, the system uses compressed-domain analysis, directly extracting motion vectors from the compressed container (H.264). This measure significantly reduces computational complexity and latency, which is critical for real-time processing.

Second, the system employs intelligent multimodal integration to enhance reliability. During preprocessing, a mandatory step of synchronizing the timestamps of video and audio streams is introduced. Additionally, the integration module includes a Gated Cross-Attention mechanism. This mechanism allows data streams to influence each other (for example, an unusual sound increases attention to the corresponding visual segment), providing mutual verification of signals and significantly improving resilience to false positives.

As a result, a significant increase in inference speed and reliability (F1-score) is expected. However, the architecture has limitations: its performance depends on the quality of motion vectors provided by the codec. Moreover, the addition of attention mechanisms makes the decision-making process less interpretable, which may complicate error diagnosis.

Conclusion

In this work, a conceptual architecture was proposed and described, aimed at improving the efficiency of anomaly detection in video surveillance systems. The main feature of our approach lies in combining two key directions: optimization and multimodal analysis.

The use of the generative MemAE method addresses the problem of scarce-labeled anomalous data. Models trained on "normal" behavior can effectively detect rare and previously unknown types of anomalies.

Furthermore, the integration of data from other sources, such as audio, makes the system more reliable and robust to noise. This significantly expands the range of detectable anomalies, allowing the system to rely not only on visual signals but also on auditory and other modalities.

The proposed concept opens new possibilities for creating reliable and universal anomaly detection systems capable of operating in complex environments and adapting to new types of threats.

Future directions will focus on scaling and enhancing the system's predictive capabilities. In particular, the integration of multi-view (multi-stream) data of a single action captured from different angles is planned. As demonstrated in the study on anomaly detection using spatio-temporal graph convolutional networks (ScienceDirect, 2024) [24], using data from multiple viewpoints allows modeling complex object interactions with more complete information, significantly improving overall system accuracy.

The most promising direction is the transition from reactive detection to proactive prediction of critical events. This requires the implementation of models capable of predicting Time-to-Collision (TTC), as discussed in Alhashim & Khan, 2019 [25]. TTC analysis necessitates the development of models that can forecast object trajectories in the near future, enabling the system to issue alerts before an actual anomaly occurs.

References

[1] Fu, H., Zhang, L., & Chen, Y. (2025). Analysis and warning of safety behavior of smart campus personnel using deep learning. Frontiers in Artificial Intelligence, 8, 112–127. https://doi.org/10.3233/FAIA231414

[2] Manju, R., Singh, P., & Kumar, A. (2025). Early anomalous action detection using MRCNN-LSTM. ETASR, 10656, 1–15. https://doi.org/10.48084/etasr.10656

[3] Chen, Y., Li, F., & Wang, H. (2025). Intelligent surveillance and anomaly detection in public spaces. Computers, Materials & Continua, 82(1), 345–362. http://doi.org/10.3778/j.issn.1002-8331.2408-0230

[4] Alinezhad Noghre, G., Danesh Pazho, A., & Tabkhi, H. (2025). A Survey on Video Anomaly Detection via Deep Learning: Human, Vehicle, and Environment. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(8), 1-21.

[5] Liu, W., Chen, Q., & Li, H. (2025). Self-supervised learning video anomaly detection: A survey. Pattern Recognition, 133, 109–125. https://doi.org/10.1016/j.patcog.2025.109125

[6] Saket, R., Patel, M., & Sharma, V. (2025). Deep learning applied for abnormal human behavior recognition. Springer, 2025, 1–17. https://link.springer.com/article/10.1007/s00779-021-01586-5

[7] Manuel J. C. S. Reis. (2025). Low-light image enhancement using deep learning. Applied Sciences, 15(11), 6330. https://doi.org/10.3390/app15116330

[8] Wang, T., Zhao, J., & Li, K. (2025). Video anomaly detection with spatio-temporal representation learning. Expert Systems, 42(5), 1–16.

[9] Saeed, A., Khan, S., & Ahmed, S. (2025). GAN-enabled anomaly detection: A comprehensive review. Expert Systems with Applications, 224, 120–135. https://doi.org/10.1016/j.eswa.2025.120135

[10] Chu, X., Li, Y., & Wang, J. (2025). Enhancing video anomaly detection: Object-based pseudo anomalies and memory-augmented autoencoders. In Advances in Intelligent Systems and Computing (Vol. 10345, pp. 223–237). Springer. https://doi.org/10.1007/978-3-031-98164-7_7

[11] Xie, Y., Sun, H., & Zhang, L. (2025). Abnormal behavior detection using multi-modal weakly supervised learning. In Advances in Intelligent Systems and Computing (pp. 105–120). Springer. https://doi.org/10.1007/978-981-99-6706-3_16

[12] Barbosa, R., Silva, F., & Oliveira, P. (2025). Multi-modal and weakly supervised video anomaly detection: A review. Neurocomputing, 525, 213–230. https://doi.org/10.1504/IJDATS.2023.136681

[13] Omarov, B., Omarov, N., Sarbasova, A., Sapakova, S., Koishiyeva, T., Kaliyev, N., Koishibayev, A., & Omarov, B. (2019). Face recognition in video surveillance systems as an application of smart cities. Journal of Theoretical and Applied Information Technology, 97(9), 2497–2509.

[14] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, V., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[15] Li, J. (2025). Area under the ROC Curve has the most consistent evaluation for binary classification. PLoS ONE, 20(4), e0316019. https://doi.org/10.1371/journal.pone.0316019

[16] Google Developers. (n.d.). GAN architecture. Retrieved October 9, 2025, from https://developers.google.com/machine-learning/gan/gan_structure?hl=ru

[17] Backus, M., & Jiang, Y. (2022). Video Frame Prediction with Deep Learning. CS231n Stanford University. https://cs231n.stanford.edu/reports/2022/pdfs/29.pdf

[18] Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. Foundations and Trends® in Machine Learning, 12(4), 307–392.

[19] Kim, J., Lee, H., & Park, S. (2025). m&m-Ant: Multi-level and Multi-modal Anticipation for Fine-grained Human Action Prediction. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[20] Kim, S., & Park, J. (2025). A survey on abnormal human behavior detection using video analytics. Expert Systems, 42(3), 1–15. https://www.sciencedirect.com/science/article/abs/pii/S0031320325009495

[21] Zhang, R., Ding, Q., & Zhou, X. (2023). Exploiting Completeness and Uncertainty of Pseudo Labels for Weakly Supervised Video Anomaly Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[22] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning Temporal Regularity in Video Sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[23] Zhou, P., & Liu, Q. (2025). Video anomaly detection based on deep neural networks: A comprehensive review. Expert Systems with Applications, 229, 120–139.

[24] Alhashim, I., & Khan, I. (2019). Forecasting Time-to-Collision from Monocular Video: Feasibility, Dataset, and Challenges. Presented at the International Conference on Computer Vision Workshops (ICCVW).

[25] ScienceDirect. (2024). Video anomaly detection using a multi-view perspective and spatial–temporal graph convolutional networks. DOI:10.1109/IROS40897.2019.8967730