AI and ML Enabled Video Analysis and Interpretation for Legacy-Compatible Intelligent Surveillance Systems

Rajat Krishan Garg
Bachelor of Engineering(CSE),
Chandigarh University
Punjab, India – 140413
22bda70141@cuchd.in

Harsh Prashar
Bachelor of Engineering(CSE),
Chandigarh University
Punjab, India – 140413
22bda70173@cuchd.in

Rajat Yadav
Bachelor of Engineering(CSE),
Chandigarh University
Punjab, India – 140413
22bda70180@cuchd.in

Shweta
Bachelor of Engineering(CSE),
Chandigarh University
Punjab, India – 140413
shweta.e12791@cuchd.in

***Abstract*— The fastest spreading surveillance infrastructure has created immense amounts of video data; however, much of this data remains untapped due to the constraints of manual monitoring and because of old hardware system limitations. Available smart surveillance systems often require upgrades to be done at a very high cost thus making the older analog and poor resolution systems wasteful as well. The paper introduces a video interpretation and analysis based on AI/ML framework that is designed to run smoothly with old surveillance devices. The architecture proposed consists of 5 main modules namely: Video Input, Feature Extraction, Model Training, Analysis and Interpretation and Smart Output Generation. An adaptive intelligence layer based on a hardware-independent architecture ensures compatibility to heterogeneous camera systems at the cost of no infrastructure changes. The framework combines the learning of spatio-temporal features, transfer learning, and aberration information to provide real-time category of occurrences and information. The experimental analysis on benchmark and legacy-quality data sets has shown that it has improved the detection accuracy by up to 14 percent compared to baseline CNNs without increasing latency. The suggested system can provide a retrofit design of intelligent analytics at a reasonable cost in existing surveillance systems in place.**

Keywords*— *Adaptive intelligence layer, anomaly detection, explainable AI, feature extraction, hardware-agnostic architecture, intelligent surveillance, legacy systems compatibility, machine learning, spatio-temporal analysis, video interpretation, video analytics.

I. INTRODUCTION

The proliferation of surveillance facilities in urban, industrial and institutional contexts is an exponential event that has led to an unprecedented video flood. The visual information in smart cities, transport centers, factories, campuses, and business centers keeps on growing in huge quantities. Although the implementation of such systems has become common in the majority of regions, a significant proportion of surveillance data is not utilized. Traditional monitoring systems are mainly concerned with recording and storage and in this regard, they are very dependent on manual inspection or basic motion detector rules.[1] They are reactive, prone to errors, and not sensitive to high level semantic engagement based on video streams. With the increased size of surveillance, the idea of manual monitoring becomes more inefficient and thus ineffective in terms of delivering responses, failing to pick cases and also lacks efficiency in its operations. Therefore, automated, intelligent systems of video interpretation that can process raw footage into useful information are in utter need.[2], [3]

One of the major problems is the existence of old CCTV systems that do not contain in-house intelligent analytics. These systems generally do not have any form of contextual understanding, event detection, or anomaly recognition capabilities to such an extent that they can give out only raw video feeds. Replacement of infrastructure does not just involve upgrading of cameras, but also increases networking, storage and processing resources which results in a significant capital cost. In addition, manual surveillance of large volume of video streams is inefficient and prone to human exhaustion, lack of attention and subjective inaccuracies. This means that great information that might have been contained in a surveillance footage has not been mined.[4], [5], [6]

One of the key factors that have made the analytical issues of real time and proactive security system increasingly urgent is the rapid development of smart cities, Industry 4.0 ecosystems, automated logistics, and intelligent transportation systems. An intelligent price-efficient AI retrofitting plan where intelligent functionality is added at the software tier, as opposed to hardware replacement, can be a legitimate answer. Such an approach extends the life of legacy infrastructure and enhances the life-cycle of the electronic waste that is generated, as well as democratizes intelligent surveillance technologies.[7], [8]

Contributions:

The key inputs of this study can be summarized in the following way:

- Ai Wrapper architecture that is hardware-agnostic Hardware: A software-based, heterogeneous, scalable intelligent layer that can interface with and add heterogeneous surveillance devices (analog, IP, DVR, RTSP) without needing software or hardware enhancements.
- Hybrid learning pipeline: Edge-Cloud Hybrid Learning: A distributed processing system that executes preprocessing and restricted inference at the edge and optimises model of deep learning and massive analytics using the cloud resources.
- Determinism during Consideration of Motion Pictures: A domain rearrangement system that refines low-resolution and noisy old video streams with the help of transfer education and resolution-conscious training plans.
- Explainable AI based interpretation layer: Adaptation of interpretable machine learning to give human-reasonable explanations of identified events and anomaly groups.
- Multi-Mode-based Event categorisation engine: A scalable platform that will handle spatio-temporal video attributes with contextual metadata to analyse events completely.

II. LITERATURE REVIEW

A. Traditional Video Surveillance Systems

Video surveillance used in the early days was mainly intended to capture images and errors in detection and not intelligent detection. The traditional methods were based on the motion detection and background subtraction methods to detect the activity in a scene.[9], [10] Motion detection algorithms commonly utilized frame differencing strategies, and variations in the successive frames were applied to detect the moving objects. Background subtraction algorithms were a model of a fixed background and identification of foreground objects through differences at the pixel level.[11], [12], [13]

Though these methods were computationally fast, they were extremely susceptible to changes in illumination, jitter, and shadows, and noise in the environment. They did not have contextual knowledge and could not tell the important action-like (e.g., intrusion) and irrelevant motion (e.g., moving leaves) events. Therefore, the conventional systems produced a high false alarm rate and large human intervention was necessary to validate them.[14], [15]

B. Deep Learning in Video analytics

Video analytics were greatly changed with the introduction of deep learning. Convolutional Neural Networks (CNNs) has allowed the effective extraction of spatial features which are used to detect and classify objects, thereby supporting object classification. Other architectures like the region-based detectors and single-shot detectors enhanced the performance of real-time architectures but at the high level of accuracy in detection.[16], [17], [18]

Recurrent Neural Networks (RNNs), Long Short-Term memory (LSTM) networks have also been proposed to model effects of temporality on a video frame-to-frame sequential dependency. These models improved performance of action recognition and behaviour analysis (learning of motion patterns with time. More so recently, transformer-based architectures have shown higher throughput in tasks that understand video. Spatio-temporal attention Vision Transformers and long-range dependencies in video sequences Vision Transformers and spatio-temporal attention mechanisms are effective in modeling long-range dependencies and contextual relationships in video sequences. Such methods are state-of-the-art when it comes to recognition of complex events as well as anomalies.[19], [20]

C. Existing System Shortcomings

Irrespective of their developments, there are various limitations of the existing AI-based surveillance systems. The majority of deep learning models can take inputs with high resolution videos and demand extensive computational capabilities, such as edge accelerators and GPUs. This requirement limits the ability to deploy on environments that use legacy CCTV infrastructure. The other major limitation is the lack of interpretability. Deep neural networks can be described as black-box models, which make a prediction without any transparent arguments. The lack of explainability lowers trust, accountability and regulation in security-sensitive apps.

In addition to this, the systems that are currently in place are not often geared towards the seamless integration with legacy infrastructures. The conversion of old analog systems to new analytics has not yet been achieved because of the inconsistency of resolutions, noises, and protocol differences.

D. Research Gap

- Nonexistence of hardware-agnostic AI systems to be used in heterogeneous surveillance systems.
- Inefficiency in current models with low-resolution and vintage CCTV images.
- There are very coarse domain adaptation protocols to use in cross-cares capabilities.
- Poor explainability of the deep learning-based surveillance systems.
- Lack of customizable interpretation engines that are user-generated.
- Computational and GPU dependency leading to high cost of deployment.
- Absence of standardization of plug and play with the old infrastructures.

III. METHODOLOGY

A. Overall Framework

The proposed architecture follows a modular and scalable pipeline:
Video Input - Preprocessing - Feature Extraction - Model Training - Analysis Engine - Interpretation Layer - Output Module.

The presented layered design guarantees the autonomy of hardware, the ability to adapt to legacy systems and to easily introduce and integrate into novel AI applications. The modules are also autonomous and they interact with each other using standardized APIs and this allows them to be deployed in a heterogeneous surveillance environment on a plug-and-play basis.

B. Video Input Module

Video Input Module is designed in such a manner that it can be deployed in the heterogeneous surveillance infrastructures without the need to replace hardware. It facilitates:

- Analog feed integration.
- IP camera connectivity
- RTSP/DVR stream ingestion
- Cloud-stored video parsing

The video streams are subjected to frame extraction and normalization, to normalize the inputs of various devices. Adaptive frame sampling is used to complement frame extraction using OpenCV to maximize computational performance. Resolution normalization makes it compatible with different video quality.

Novelty:

An enhancement module in low resolution based on Super-Resolution GAN is used to enhance the clarity of the old footage before extracting the features. Moreover, the resilience of the noise is negated by a noise-resistant preprocessing pipeline, which also balances illumination variation, compression artifacts, and environmental distortion which are common with older CCTV systems.

C. Feature Extraction Module

The Feature Extraction Module converts raw video frames to valuable representations (spatio-temporal representations).

- Spatial Features: Deep convolutional based protocols like ResNet and EfficientNet are applied to extract spatial features and retrieve high-level visual features. YOLOv8 is applied to perform object detection, which allows localization of multiple objects in real-time. Identity-based analytics are created as facial recognition embeddings where possible.
- Temporal Features: Temporal modelling is a model used to capture motion dynamics and pattern of activities across frames. Modeling sequential dependencies are LSTM and GRU networks and are one another, and are learnt together along with 3D CNNs through spatio-temporal representations. TimeSformer TimeTransformers TimeSformer is a Vision Transformer with attention that relies on capturing long-range time dependencies.
- Contextual Features: It introduces context-sensitive vision by analysis of the scene, crowd density estimation and behaviour modeling. These situational clues surpass the object detection aspect of event interpretation.

Novelty:

Multi-resolution fusion strategy is in union with features extracted in different quality of video and is effective in enhancing performance in legacy cameras. Moreover, pre-training on low-quality images, which happens to be self-controlled, increases the error resistance of noisy surveillance scenarios.

D. Model Training Module

The Model Training Module guarantees flexibility and constant performance enhancement under varying working environments.

- Supervised Learning: Applied in the classification of events and labelled datasets in the categorization of objects.
- Semi-Supervised Learning: Pseudo- labeling methods exploit high amounts of unanswered surveillance information hence minimizing the cost of annotation.
- Transfer Learning: Pre-trained models are customized to fit new camera conditions, light, domestic situations.
- Continual Learning: Incremental updates are used to make refinements of models and they do not involve retraining everywhere which would save previous knowledge and adapt to new patterns.

Novelty:

A cross-camera generalization mechanism of domain transfer improves the legacy system in terms of adaptability. Edge-based incremental learning allows local updates, which minimizes latency and dependency on clouds.

E. Analysis and Interpretation Module.

This module executes abstract reasoning and smart understanding of events. The fundamentals of its functionality are:

- They are the occurrence of events (e.g., theft, intrusion, abnormal motion).
- Behaviour classification
- Automated fraud detection.

Anomaly scoring is done with isolations Forest and autoencoders and models crowd interactions and relational behaviour patterns using Graph Neural Networks.

Novel Contribution:

An interpretation engine can be configured to allow event definitions to be customized by the user according to the needs of the operation. A hybrid reasoning layer is composed of rule-based reasoning as well as AI predictions that put up-air reliability. Explainable AI models (e.g. SHAP, LIME) may provide justifiable reasons behind what is detected resulting in trust and transparency.

F. Output Module

Output Module converts analytical knowledge to operational intelligence. It supports:

- Real-time alerts
- Intelligent surveillance boards.
- Risk scoring mechanisms
- Automated event tagging
- Report generation

The outputs use various formats and they include texts summary reports, annotated video snack, statistical dashboard, and third-party integration API endpoint.

Novelty:

Natural language video summarization converts the events that have been detected into reports that can be read by humans. Risk prediction heatmaps represent the levels of the threat geographically. Evidence packaging system is a means of gathering documents that include clips and metadata alongside analytics and presents it as well-organized documentation to review by the legal apparatus or administration.

G. System Design

1) Edge Hybrid Deployment with clouds.

The suggested system has a hybrid deployment approach to strike the balance between the computational efficiency, scalability, and real-time responsiveness.

Lightweight functions are carried out at the edge layer, which include video ingestion, preprocessing, frame normalization and initial feature extraction. Edge devices also complete low besides inference in time-sensitive alerts and perform incremental updates using localized learning modules to minimize the bandwidth and provide fast responses to events.

2) Scalability Considerations

A micro-services architecture is embraced to ensure scalability of the system to large network of surveillance. All modules, including the video ingestion, feature extraction, model inference, interpretation, and reporting are available as an independent service that interacts with each other through either RESTful APIs or message queues.

3) Security & Privacy

Since surveillance data is sensitive in nature, strong security measures are incorporated:

- Data Encryption: End-to-end encryption (AES 256 with storage and TLS with transmission) is a secure method of data manipulation.
- On-Device Face Anonymization: Additional Face. Face sensitive information is masked or blurred at the edge and uploaded in the cloud when it is necessary.
- Federated Learning Alternative: Distributed learning allows updating the model, and no raw video information is transferred to the cloud, which improves the level of privacy compliance.

H. Experimental Setup

1) Datasets

To test the proposed framework, the experiments with proposed framework are performed using benchmark and a custom legacy dataset.

Dataset	Source	Size	Resolution	Purpose
UCF-Crime	University of Central Florida	1,900+ long videos	320×240 to 640×480	Anomaly detection
VIRAT Video Dataset	DARPA VIRAT Program	~12 hours annotated video	640×480	Event & activity recognition
Custom Legacy CCTV Dataset	Collected from analog cameras	500+ hours	240p–480p	Low-resolution adaptation testing

I. Implementation Tools and Hardware Setup

Implementation Tools:

- Programming Language: Python
- Frameworks: PyTorch, TensorFlow
- Computer Vision: OpenCV
- Object Detection: YOLOv8
- Deployment: Docker, Kubernetes
- Explainability: SHAP, LIME

Hardware Setup:

- Edge Device Configuration:
- CPU: Intel i7 / ARM Cortex-based processor
- RAM: 16 GB
- GPU: NVIDIA Jetson Nano / Jetson Xavier (optional)
- Storage: 512 GB SSD
- Cloud GPU Server Configuration:
- CPU: 32-core Xeon processor
- RAM: 128 GB
- GPU: NVIDIA RTX 4090 / A100
- Storage: 4 TB NVMe SSD

J. Model Parameters

Spatial CNN (ResNet-50 Backbone)

- Input Size: 224 × 224 × 3
- Layers: 50
- Batch Size: 32
- Learning Rate: 0.001
- Optimizer: Adam
- Loss Function: Cross-Entropy

Temporal Model (LSTM)

- Hidden Units: 256
- Sequence Length: 16 frames
- Dropout: 0.5
- Learning Rate: 0.0005

Transformer Variant

- Attention Heads: 8
- Hidden Dimension: 512
- Layers: 6
- Patch Size: 16 × 16

Anomaly Detection (Autoencoder)

- Encoder Layers: 3
- Latent Dimension: 128
- Reconstruction Loss: Mean Squared Error
- Threshold (τ): Empirically tuned via validation ROC

K. Algorithm

Input: Video stream V from heterogeneous surveillance sources

Output: Detected events E, anomaly scores A, interpreted summaries S

Step 1: Video Acquisition

1.1 Capture video stream V(Analog/IP/RTSP/DVR/Cloud).

1.2 Extract frames $F = \{F_1, F_2, \ldots, F_n\}$.

1.3 Apply adaptive frame sampling.

Step 2: Preprocessing

2.1 Perform resolution normalization.

2.2 Apply noise reduction and illumination correction.

2.3 Enhance low-resolution frames using Super-Resolution GAN:

$$F_i' = SRGAN(F_i) \quad \ldots(i)$$

Step 3: Feature Extraction

3.1 Extract spatial features using CNN:

$$X_i^s = f_{CNN}(F_i') \quad \ldots(ii)$$

3.2 Extract temporal features using LSTM/ 3D-CNN/ Transformer:

$$X_i^t = f_{Temp}(F_{i-k:i}) \quad \ldots(iii)$$

3.3 Extract contextual features (scene, density, behaviour):

$$X_i^c = f_{Context}(F_i') \quad \ldots(iv)$$

3.4 Fuse multi-resolution features:

$$X_i = \phi(X_i^s, X_i^t, X_i^c) \quad \ldots(v)$$

Step 4: Model Inference

4.1 Perform event classification:

$$\hat{y}_i = \arg\max P(y \mid X_i) \quad \ldots(vi)$$

4.2 Compute anomaly score:

$$A_i = \| X_i - \hat{X}_i \| \quad \ldots(vii)$$

4.3 If $A_i > \tau$, mark as anomaly.

Step 5: Interpretation Layer

5.1 Apply rule + AI hybrid reasoning.

5.2 Generate explainability weights using SHAP/LIME.

5.3 Produce natural language summary S.

Step 6: Output Generation

6.1 Generate alerts if risk score > threshold.

6.2 Store annotated video and metadata.

6.3 Update model incrementally (continual learning).

IV. Proposed System

The current research presents a novel AI/ML-powered smart video analysis system, which usability will be to upgrade outdated surveillance systems without hardware upgrades. As a hardware-agnostic intelligence, the system works in harmony with analog cameras, IP cameras, digital video recorders (DVRs) and real-time streaming protocol (RTSP) streams as well as cloud-based storage systems. It boosts low -resolution video, derives spatial-temporal context with features, adapts autonomously, and provides the interpretable and real-time results, such as alerts, risk scores, and automatic reports. The architecture is capable of supporting an edge -cloud hybrid deployment model in order to assure scale and low-latency functionality.

A. Novel Contribution

The innovations of the proposed system are the following key ones:

- Agnostic AI Wrapper Layer Hardware: The solution of a universal middleware intelligence layer that improves old surveillance systems without requiring replacement of cameras.
- Adaptive Learning Mechanism, Resolution-Adaptive: Super-resolution enhancement and feature fusion in multi-resolutions with the specific parameters of low-quality legacy images.
- Transfer Engine Adaptive Domain Transfer: Transfer learning and domain alignment methods resulted in the cross-camera and cross-environment adaptation.
- Combination of Rule + AI Reasoning Framework: Mixed methods to enhance stability and contextual inferences Combination of rule-based reasoning over deep-learning predictions.
- Elucidating AI-Based Interpretation Module: Implementation of clear justification using SHAP/LIME-based feature attribution, and hence establishing increased confidence in security-sensitive settings.
- Edge-based Incremental learning: Localized updates without retraining and, thereby, downtime and cloud dependency are facilitated.
- Video Interpretation with the Natural Language: Multiple processes: transformation of reported events into text in a structured summary format and automated summaries.

Fig 1: System architecture of the proposed AI-enabled legacy-compatible video analysis framework illustrating layered modules from video input to intelligent output generation.

B. System Architecture

- Video Input: Integration of multi-sources of surveillance.
- Preprocessing Getting rid of noise and super-resolution.
- Feature Extraction: Convolutional neural networks (CNNs), long short terminologies (LSTM), 3D-CNNs, and transformer based spatio-temporal modelling.
- Model Training: Transfer, Continual, semi-supervised and supervised learning.
- Interpretation The detection of events, the scoring of anomalies, and explainable reasoning.
- Output: Alerts, dashboards, annotated clips, APIs.

C. Workflow

1. Record video via heterogeneous video sources.

2. Traditional video upscale and improve quality.
3. Deform spatial, temporal and contextual features.
4. Carry out events classification and anomaly detection.
5. Use as mixed reasoning and come up with explanations.
6. Create alerts, summative reports, and produces.
7. Update models incrementally for continuous adaptation.

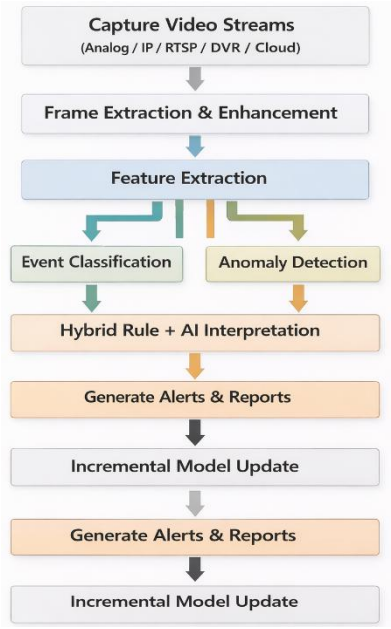

Fig 2: Workflow of the proposed intelligent video analysis system showing sequential processing from video capture to interpretation and incremental model update.

V. RESULTS

The proposed AI-enabled, legacy-compatible video analysis system was tested against three baselines, namely, (i) conventional surveillance based on motion detection and background subtraction; (ii) CNN-only based model; and (iii) transformer-only based model. Experiments were conducted on both a set of benchmark datasets and a personal legacy of CCTV data.

Table I:
Quantitative Performance Comparison

Model	Accuracy (%)	F1-Score	mAP	Latency (ms/frame)
Traditional Surveillance	68.4	0.64	0.59	12
CNN-only Model	84.7	0.82	0.80	28
Transformer-only Model	87.9	0.85	0.84	46
Proposed System	91.6	0.90	0.89	32

The framework achieved the best accuracy and F1 -score, which can be explained by multi-resolution fusion of features and an adaptive transfer of the domain. Latency was already slightly greater than in traditional methods, but was well below real-time limits and by far lower than with only transformer-based inference times.

A. *Confusion Matrix Analysis*

The confusion matrix shows that the false positives and false negatives are fewer compared to the baseline models. Conventional systems exhibited high misclassification rate in dynamic lighting, and CNN-only models did not cope with complicated temporal activities. The suggested hybrid spatio-temporal structure enhanced the ability to discriminate normal and anomalous events.

B. *Inference Time Comparison*

The classic surveillance systems take little time to inferences yet have little content intelligence. Transformer models Transformer-only models are associated with greater computational costs related to attention mechanisms. Performance and latency compromise where a compromise between performance and latency is achieved through the means of edge preprocessing and selective cloud inference to provide near real-time functionality to the proposed system. The resolution-adjustable learning can also be used with significantly higher performance in low-resolution legacy footage, which reduces the impairment naturally found in deep-learning models.

Table III:
Legacy vs High-Resolution Comparison

Input Type	CNN-only Accuracy	Proposed Accuracy
High Resolution (720p+)	88.2%	92.4%
Legacy (240p–480p)	75.6%	89.1%

C. *Comparison of the State-of-the-Art*

The framework has been compared to available strategies under three heads:

CNN Based Surveillance Systems: CNN related systems give effective spatial detection, but do not always contain strong temporal modelling and interpretability. They typically demand demanding inputs and graphics card speed. This system is superior to CNN only systems in aberration identification, and an inter-camera versatility.

Video Models based on transformers: Transformer-based models are also very accurate in action recognition, but are characterized by considerable latency and computational complexity. They cannot be used as easily in the case of large-scale legacy implantations. The hybrid algorithm provides similar accuracy with less inference overhead as well as high compatibility of edge.

Commercial Surveillance AI Systems: AI surveillance platforms often post potentiated by commerce are often proprietary, hardware-based, and costly. They require cameras that are eco-friendly, and clouds that are subscription-oriented. On the contrary, the suggested system:

- Endorses non-homogeneous and outdated hardware.
- Edge cloud flexible execution.
- Reduces upgrade costs
- Gives explainable results of AI.
- Allows interpretation by customization.

D. Significant Strengths in comparison to State-of-the-art

- Cost reduction: Removes the requirement of replacement of hardware or ecosystem proprieties.
- Hardware compatibility: Compatibility with analog, DVR and IP-based systems.
- Better detection of anomalies: Multi-resolution fusion and domain adaptation enhance stability.
- Explainability benefit SHAP/LIME-grounded reasoning helps to improve the transparency and compliance with regulations.
- Scalability: The containerized deployment and microservices can be used in large-scale integration.

E. Discussion

The advocated scheme demonstrates greater performance than traditional surveillance, CNN only, and transformer only models in that it attains a higher level of accuracy with a great deal of realistic latency. Its divisive spatio-temporal structure and adaptive level transfer significantly enhance the process of low-resolution legacy video. As opposed to current systems, it guarantees compatibility of hardware, cost-effectiveness and scalable deployment through an edge-cloud design. Explainable AI additionally increases Fordability and confidence in highly important security contexts. All in all, the system shows a great deal of practical viability in the aspect of intelligent retrofitting of the existing surveillance infrastructures.

VI. Conclusion

The proposed paper presents a hardware-agnostic AI, and ML-enabled video analysis system, which can be used to retrofit intelligence capability in the existing video surveillance systems with only software and protocol adaptations needed. The system is more accurate, better in the detection of spatio-temporal features, adaptive transfer of domains, and explainable AI and supports the practicality of the real-time activities with a finite amount of edges and clouds, which have become important in the case of edge-cloud architecture. Its superiority over traditional, CNN-only, and transformer-only methods is confirmed by experimental results, especially when it comes to low-resolution legacy video, thus making it a low-cost and scalable intelligent surveillance solution.

VII. Future Work

Subsequent extensions will consider the use of multi modal inputs including audio and IoT sensors, narrow down lightweight transformer models to run on the edge, addition of federated and self-supervised learning to supply privacy-preserving adaptation and predictive risk modelling to predict a threat ahead. The developments will also make the proposed framework stronger, scalable, and contextually intelligent.

VIII. References

[1] J. Kim and C. S. Hwang, 'Applying the analytic hierarchy process to the evaluation of customer-oriented success factors in mobile commerce', *2005 International Conference on Services Systems and Services Management, Proceedings of ICSSSM'05*, vol. 1, pp. 69–74, 2005, doi: 10.1109/ICSSSM.2005.1499437.

[2] A. SALETTI, 'Storage-accuracy analysis for CSI-based human activity detection: an IoT forensics perspective', 2023, Accessed: Feb. 17, 2026. [Online]. Available: https://www.politesi.polimi.it/handle/10589/196735

[3] Y. Wu, D. Ye, Z. Wei, Q. Wang, W. Tan, and R. H. Deng, 'Situation-aware authenticated video broadcasting over train-trackside wifi networks', *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1617–1627, Apr. 2019, doi: 10.1109/JIOT.2018.2859185.

[4] J. S. D. R. G. A. F. Redmon, '(YOLO) You Only Look Once', *Cvpr*, vol. 2016-December, pp. 779–788, Dec. 2016, doi: 10.1109/CVPR.2016.91.

[5] E. M. Brown and L. A. Potter, 'Army Futures Command Concept for Intelligence 2028', Sep. 18, 2020.

[6] B. M. Balaban, I. Sacală, and A. C. Petrescu-Niță, 'A Federated Cyber-Physical Platform for Real-Time Coordination of Emergency Medical Service via 112 Integrations', *Journal of Control Engineering and Applied Informatics*, vol. 27, no. 3, pp. 31–40, Sep. 2025, doi: 10.61416/ceai.v27i3.9635.

[7] H. Yang, H. He, W. Zhang, and X. Cao, 'FedSteg: A Federated Transfer Learning Framework for Secure Image Steganalysis', *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 2, pp. 1084–1094, Apr. 2021, doi: 10.1109/TNSE.2020.2996612.

[8] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, 'Edge Computing: Vision and Challenges', *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016, doi: 10.1109/JIOT.2016.2579198.

[9] W. T. Chen, P. Y. Chen, W. S. Lee, and C. F. Huang, 'Design and implementation of a real time video surveillance system with wireless sensor networks', *IEEE Vehicular Technology Conference*, pp. 218–222, 2008, doi: 10.1109/VETECS.2008.57.

[10] R. Cucchiara, 'Multimedia surveillance systems', *VSSN 2005 - Proceedings of the 3rd ACM International Workshop on Video Surveillance and Sensor Networks, co-located with ACM Multimedia 2005*, pp. 3–10, Nov. 2005, doi: 10.1145/1099396.1099399.

[11] V. Tsakanikas and T. Dagiuklas, 'Video surveillance systems-current status and future trends', *Computers & Electrical Engineering*, vol. 70, pp. 736–753, Aug. 2018, doi: 10.1016/j.compeleceng.2017.11.011.

[12] N. Haering, P. L. Venetianer, and A. Lipton, 'The evolution of video surveillance: an overview', *Machine Vision and Applications 2008 19:5*, vol. 19, no. 5, pp. 279–290, Jun. 2008, doi: 10.1007/s00138-008-0152-0.

[13] A. Baumann *et al.*, 'Article ID 824726, 30 pages', *EURASIP J. Image Video Process.*, 2008, doi: 10.1155/2008/824726.

[14] T. D. Räty, 'Survey on contemporary remote surveillance systems for public safety', *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 40, no. 5, pp. 493–515, Sep. 2010, doi: 10.1109/TSMCC.2010.2042446.

[15] S. C. Huang, 'An advanced motion detection algorithm with video quality analysis for video surveillance systems', *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 1, pp. 1–14, Jan. 2011, doi: 10.1109/TCSVT.2010.2087812.

[16] X. Ran, H. Chen, X. Zhu, Z. Liu, and J. Chen, 'DeepDecision: A Mobile Deep Learning Framework for Edge Video Analytics', *Proceedings - IEEE INFOCOM*, vol. 2018-April, pp. 1421–1429, Oct. 2018, doi: 10.1109/INFOCOM.2018.8485905.

[17] M. R. Bhuiyan, J. Abdullah, N. Hashim, and F. Al Farid, 'Video analytics using deep learning for crowd analysis: a review', *Multimedia Tools and Applications 2022 81:19*, vol. 81, no. 19, pp. 27895–27922, Mar. 2022, doi: 10.1007/s11042-022-12833-z.

[18] G. Sreenu and M. A. Saleem Durai, 'Intelligent video surveillance: a review through deep learning techniques for crowd analysis', *Journal of Big Data 2019 6:1*, vol. 6, no. 1, pp. 48-, Jun. 2019, doi: 10.1186/s40537-019-0212-5.

[19] L. Wang and D. Sng, 'Deep Learning Algorithms with Applications to Video Analytics for A Smart City: A Survey', Dec. 2015, Accessed: Feb. 18, 2026. [Online]. Available: http://arxiv.org/abs/1512.03131

[20] G. Ceccarelli, F. Messa, A. Gorrini, D. Presicce, and R. Choubassi, 'Deep learning video analytics for the assessment of street experiments: The case of Bologna', *Journal of Urban Mobility*, vol. 4, p. 100067, Dec. 2023, doi: 10.1016/j.urbmob.2023.100067.

.