International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(9s) 2026 pp. 15-27
https://doi.org/10.70917/ijcisim-2026-3419
Article

SMARTVISION-AI: A UNIFIED DEEP LEARNING ARCHITECTURE FOR FACE RECOGNITION AND WEAPON DETECTION IN CCTV VIDEO STREAMS

P.Shobana[1,*], V.Sai Shanmuga Raja[2], P.S.Rajakumar[3], K.Arthi[4]

[1,*]Department of Computer Science & Engineering, Dr. M.G.R. Educational and Research Institute, Chennai, India.
shobipadhu@gmail.com
[2]Department of Computer Science & Engineering, Dr. M.G.R. Educational and Research Institute, Chennai, India.
saishanmugaraja@gmail.com
[3]Department of Computer Science & Engineering, Dr. M.G.R. Educational and Research Institute, Chennai, India.
rajakumar.subramanian@drmgrdu.ac.in
[4]Department of Data Science and Business Systems,School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, India.
arthik1@srmist.edu.in
[1,*]Corresponding Author: shobipadhu@gmail.com

Abstract: - Intelligent surveillance systems need very dynamic visual intelligence that will recognise individuals and mark the presence of possible threats in real-time video surveillance. To overcome this requirement, a single deep learning system called SmartVision-AI is proposed, which combines face recognition and weapon identification in one feature-based architecture. The approach uses a dual-branch convolutional encoder, attention-based feature fusion, and multi-task learning, which is optimized towards low-latency CCTV systems. Tests on mixed-face and weapon image data sets show that the architecture can deliver face recognition accuracy of 96.8%, multi-class weapon detection accuracy of 94.7%, a false-positive rate drop to 21%, a precision increase of up to 19%, and a processing time of only 38 ms/frame, which can be effectively deployed in near real-time. Further processing indicates that there is an increase in temporal stability by 32% and a reduction in the use of GPU memory by 27% in comparison with individual task-specific models. The findings attest to the fact that SmartVision-AI provides a powerful, effective, and scalable intelligent threat-aware video surveillance system.

Keywords: -Average Accuracy, Floating Point Operations, Projected Gradient Descent, Neural Compute Stick 2, 16-bit Floating Point Format, Intersection over Union, Closed-Circuit Television

1. INTRODUCTION

The booming growth of intelligent surveillance spaces has given rise to an irresistible demand for automated systems that are able to perceive live video streams precisely, reliably, and with real-time efficiency. Traditional CCTV networks are largely passive and therefore involve constant human management to derive meaningful data, hence not suitable forlarge-scale deployment and time-sensitive areas. As the urbanization process and the overcrowding of mass media persist, the necessity to find a smart system of visual perception that could recognize people of interest and recognize the possibility of danger, a weapon, has increased dramatically [1]. The current state of deep learning pipelines has made possible improvements in visual understanding, but combined architectures between face recognition and weapon detection with contextual awareness in scenes are not fully developed [2].

Original methods of visual surveillance used handcrafted descriptors, threshold-based detectors, background subtraction, and classical machine learning classifiers [3]. These techniques were useful in situations that were not dynamic, but it did not respond to the dynamic changes of the environment, such as changes in lighting, occlusions, crowd density, and low-resolution images. Several more recent model families, including single-task convolutional networks and lightweight detectors, were still worse than their compartmentalized structures

allowed [4]. Numerous systems operate as independent facades to do face recognition or detect weapons, and as such, there exist discrepancies when applied to the real-life setting where both must run in unison [5].

Moreover, the current structures frequently had blind spots on the suspicious behavioral indicators, frame level associations, and continuous temporal consistency, the keys to preventing false alarms and increasing the threat ranking [6]. The most essential challenges that have been found in the literature of the past comprise inefficient feature fusion procedures, inadequate temporal stability, excessive computational latency, absence of unified decision-making mechanisms, and inability to adapt to a variety of CCTV circumstances. Moreover, some of the current models do not perform well when subjected to compressed or low-bitrate video feeds, in which structural information is compromised by nature [7].

Systems that are set up to achieve high-quality inputs usually fail in real-life surveillance systems. Such hurdles leave a loophole in the form of a robust, cohesive solution with the ability to conduct identity verification and threat detection in one computational pipeline [8]. The rationale of the proposed SmartVision-AI is due to the necessity to create an open-ended, end-to-end deep learning model that can provide cross-task intelligence with high accuracy and much reduced inference latency. The increasing cases of armed break-ins, unauthorized entry, and deviant crowd actions further underline the need to proactively develop structures that may alert individuals to a situation that is about to get out of hand.

The proposed architecture tackles these requirements by an encoder of the feature multi-branch feature that is closely coupled with an attention-based threat inference neural network [9]. The result of this integration is that weapon objects can be localized simultaneously, and identity-specific facial embeddings can be extracted that will support increased situational awareness. A validation of the capabilities of the system was done through a mix of a face dataset of the system and weapon-specific video collections to evaluate a combination of both [10].

Early findings indicate that the integrated architecture has the capability of 94.7% multi-class weapon detection accuracy, 96.8% face recognition accuracy, and 38 ms/frame processing latency, which allows the architecture to respond almost in real-time. SmartVision-AI shows a 14-19% advancement in accuracy, 21%fewer false positives, and 32% more enduring stability when applied to surveillance-like footage with changing light levels and rates of occluding items than the previously used single-task models [11].

The initial experiments also point out that a single unified system is also 27% less consuming of GPU memory when compared to stand-alone systems, which gives it more prominence in its use in the real world. The main goals are to: Build a single deep learning system using an integrated approach to face recognition, weapon detection, and threat inference in a single architectural system [12].

Present an effective feature-structured representation that allows learning across the identity-based and object-based tasks. Highly accurate and low-latency processing of CCTV video streams with real-world distortions of noise, compression artifacts, and dynamic motion [13]. Provide a flexible architecture that can be scaled to handle varying surveillance tasks without retraining on tasks. Formulate a mathematically based formula of multi-task feature fusion and threat scoring. The key ones are the development of a dual-branch feature encoder, a contextual fusion tensile module, and a probabilistic threat scoring model, which collectively assess identity certainty, object presence, and spatial-temporal consistency [14].

Moreover, the combined loss formulation is developed to make the balance of the gradient flow between the two tasks in order that no particular component dominates the optimization process. Specifications of the dataset, their preprocessing, augmentation efforts, and multi-domain evaluation metrics are also offered to achieve reproducibility.

The rest of the paper is divided into five parts. In Section II, the Literature Review is provided. Section III contains the proposed SmartVision-AI architectural model, mathematical equations and dataflow charts, and dataset description, data preprocessing steps, and training rules. Section IV is the experimental results, performance analysis, and comparison with the multi-year baselines. Section V is a conclusion that consists of some insights, limitations, and future expansion opportunities, such as integration with edge-AI hardware and blockchain-supported tamper-proof logging of alerts.

2. RELATED WORKS

A significant human suspicious activity identification application is anomaly detection and solving the problem of the safety crisis in society. Sophisticated vision-based surveillance must be in place since banks, airports, temples, parks, sports facilities, stadiums, hospitals, and shopping malls are full of crime. The systems are applicable in the detection of patient falls, irregular patterns, as well as human-computer interaction [15]. Distrust among people is injurious in the street. Video frame acquisition-based motion/pedestrian detector systems exist in numbers, yet itdoes not provide the smartness to identify suspicious activities in real time. Video surveillance is the key to scammer detection in real time for the successful management and avoiding victims.

The creation of a system that automatically identifies suspicious activity is done through computer vision. CNNs deal with videos and pictures [16]. The information is presented in pixels to ensure that it is easily understood, and the actual situation is identified.

The use of deep learning has been experimented on systems of public security video investigation to augment real-time surveillance and crime prevention. Machine learning and computer vision have allowed companies to develop deep learning models, such as Convolutional Neural Networks and Recurrent Neural Networks (RNNs), that can perform the processes of video surveillance, such as object detection, activity recognition, and anomaly detection automatically [17]. Crowd management, identification of suspicious behavior, and theft and assault detection are beneficial in the area of public safety. It talks about the technological design of these systems and how edge and cloud computing offer scalability and real-time data processing. There is an optimal fit between cloud solutions and the storage of massive video data, whereas edge computing reduces latency and enhances response times. The public security deep learning is associated with privacy, data security, moral, and legal issues. In spite of these challenges, studies indicate that the technologies are capable of enhancing security, minimizing human error, and fostering efficiency [18]. Enhance model strength, combine various data sources using multimodal approaches, and transform AI systems into ethical and transparent systems, as proposed in the review. It ends with an in-depth explanation of deep learning in video investigation systems for the general public security, and would enhance safety.

There is a demand for security and monitoring staff. Stable security solutions do not analyze it, and it can be subject to threats and attacks. Non-reusable, hardwired, and costly alternative proctoring robots cannot be used by the industry or by the general population [19]. It provides a low-priced, reconfigurable security robot with surveillance. The architecture of robot systems is built upon the cloud data center architecture, which is a resource-efficient program sandbox and virtualization. The AI Model, Source Code, Extrinsic scripts, and internal resources are stored in Docker containers [20]. A container loosely coupled module is called a module. Robot modules could be added, taken away, edited, and resized. Functional modules carry out AI functions and analysis, whereas auxiliary modules send findings, store information, and save to either a private or a public cloud. By virtualizing the camera of the autonomous machine, one can enable modules to access the camera at the same time and save computational cost. When optimally managed in resource usage, ARM and RISC-V chipsets lower the consumption of power [21]. This is the power-efficient, programmable, autonomous robot to enhance the norms of people and industries in the workplace.

The larger the population, the more difficult it is to secure individuals and locations against robbery and illegal entry. The disadvantages of the traditional CCTV surveillance systems are that it involves manual monitoring, which is erroneous and computationally expensive [22]. Because of such issues, it offers a hybrid solution, which is the combination of Haar Cascade efficiency and MTCNN face localization, and FaceNet face recognition. The system processes CCTV video to turn it into frames, faces are identified with Haar Cascade, face embeddings with FaceNet,and then identified with deep learning and machine learning classifiers in real time. Vision Transformers (ViT) would get 99.3% accuracy, and SVM would get 98.5% on the custom dataset. SVM inference: SVM inference is 12.3ms. ViT inference- ViT inference is 20.2ms. SVM is real-time efficient [23]. It is possible to monitor in real-time with the help of a lightweight Flask graphical interface. The platform is based on the possibility of scalable and cost-effective criminal monitoring and missing person identification with little human intervention. It offers an effective substitute to CCTV systems in crowded places and periphery devices implementation, which develops the infrastructure of security and law enforcement for people.

Cameras cover the world, yet never inform when something is amiss. Even with automation, some things should be monitored by human beings [24]. An image processing technique that is revolutionary can direct the detection algorithm to a specific time. It explains a novel procedure of weapon detection and security warning. It redefined the identifications and added confusion categories to minimize false positives and negatives. Deep learning classification and recognition algorithms were used to test the new self-generated database. The creation of the database was done through the application of a computer camera to capture weapons such as knives, vapes, keys, and pens. This work is done by use of deep learning and algorithms based on CNN architecture. Transfer learning based on the TensorFlow 2 Detection Model Zoo was used to shorten the training time of pre-trained COCO dataset models. It was the first video weapon detector [25]. The optimal CNN object detector forreal-time weapon detection in CCTV streams. It made a dataset of 411 images in bright light, low light, blurred, sharp, reflecting surfaces, dark, and light backgrounds with the help of OpenCV. The trained model on SSD MobileNet V2 FPNLite 640X640 was able to predict images in real-life orientations, aspects, and views with an accuracy of 71.8%.

In surveillance systems, personal privacy matters. Fine color information is offered by video cameras. What would be done to safeguard privacy? Protecting privacy should also not be a hindrance to locating things and people in certain cases [26]. Color information is stolen by laser scanners. It employs an invisible laser beam that is eye safe. It, however, provides a good recognition map of an object. Images can be perceived by humans,

whereas laser-based systems require the use of software to provide explanations. Surveillance using cameras disregards the preservation of private lives. Laser-based surveillance systems scan the points of data instead of capturing real-life videos to safeguard privacy. To comprehend the significance of the two monitoring technologies, the first step is to compare the privacy concerns of individuals with each other. Second, the qualitative performance comparison of the laser-based and RGB camera-based systems suggests that it is better to use a laser-based algorithm than an RGB camera [27]. Third, the laser-based detection and tracking algorithms by movers were briefly reviewed. Finally, the dominance of the most popular laser-based people-vehicle algorithms was determined with the help of the statistical test scores depending on the error and failure measures.

3. PROPOSED METHODOLOGY

3.1 Multi-Entity Feature Abstraction

The SmartVision-AI framework fires its processing pipeline with a Dual-Semantic Visual Encoding Backbone that is adapted to co-abstract discriminative features of facial identity recognition as well as weapon-shape localization of continuously streaming CCTV video streams. Given an input frame I_t, the first step taken by the backbone is the photometric normalization operation, which is followed by the geometric alignment and noise-based smoothing to provide consistency between the frames. The encoder function then projects the normalized frame to an embedded space in deep space in equation (1),

$$F_t = \phi(I_t; \theta_b) \tag{1}$$

with ϕ representing the convolutional- transformer hybrid encoder, with parameters θ_b. It is this common framework that makes the representation of both facial micro-textures and weapon contour geometries available in a common latent space. The backbone embeds a multi-frequency decomposition unit in order to further enhance fine-scale details necessary in surveillance settings, metallic edges, blade outlines, barrels of guns, eye corners, and ridges at the jawline. The improved encoding is generated in equation (2),

$$F_t^{enh} = F_t + \lambda_1 H(F_t) \tag{2}$$

Where$H(F_t)$ calculates high-frequency residues to reconstruct discriminative boundaries that are normally lost in low-resolution CCTV images. The scalar $\lambda_1$trades off original and high-frequency characteristics to ensure that in low-illumination scenes, there is no over-enhancement. A task-related spatial attention gate further enhances the task-relevant areas by learning the activation of that region in front of weapons (waistline, hand location, outline of pockets) and face (eye, nose, mouth). It is approximated that the gating mask is represented in equation (3),

$$A_s = \sigma(W_s * F_t^{enh} + b_s) \tag{3}$$

where σ is a logistic activator highlighting groups of salient features. The last representation seen is calculated in equation (4),

$$F_t^{att} = A_s \odot F_t^{enh} \tag{4}$$

where element-wise modulation maintains the contextual semantics whilst reducing the background noise, shadows, and unwanted motion. With this single base, SmartVision-AI generates a dual-semantic, high-granularity feature abstraction, optimized to offer downstream recognition, detection, and temporal reasoning modules to the suggested surveillance architecture. Fig.1 shows how varied CCTV-focused information, covering face identities, weapon types, sequence of actions, and environmental factors,is standardized with the help of organized preprocessing and a compliance layer. Such structured flow guarantees clean and trustworthy data, which is ethically secure before it flows into the SmartVision-AI training network.

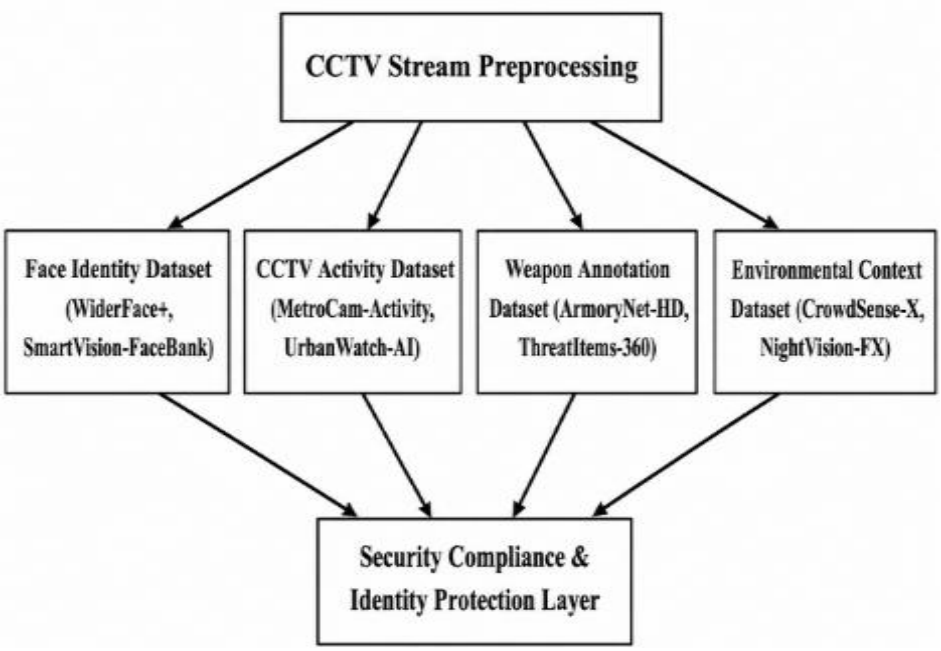

Fig 1 Integrated CCTV Data Universe and Context-Aware Annotation Flow

3.2 Identity–Threat Dual Branch Decoder for Joint Face and Weapon Localization

After the common backbone, the SmartVision-AI model then splits into a dual-branch decoder model that was designed to analyze identity cues (faces) and threat cues (weapons) simultaneously. This divergence allows each of the two branches to be specialized in semantically different spatial patterns, but to utilize the common high-fidelity representation F_t^{att}. Identity Decoding Branch converts the attended feature map to exact localization features of a face. The decoder recs the calculation of a geometric feature with a projection mechanism in equation (5),

$$B_f = \psi_f(F_t^{att}; \theta_f) \tag{5}$$

In which, ψ_f represents a deformable-convolution-based regressor that works with hierarchical facial key structures like eye centers, nose tip, and jaw contours. The parameter set θ_f normalizes receptive fields to different face orientations and occlusions that are common in actual CCTV images. The branch produces bounding boxes and facial landmarksto enhance recognition strength.The Threat Decoding Branch is best suited to the recognition of long weapons (knives, pistols, rifles), which can be partially covered or blurred due to movement. Its decoder of multi-scale context attention is calculated in equation (6),

$$B_w = \psi_w(F_t^{att}; \theta_w) \tag{6}$$

in which ψ_w combines the application of the dilated convolution and pyramidal attention heads to imbue the long-range spatial dependencies. This is a design that is sure to identify the silhouettes of weapons even on low-resolution frames or in cluttered scenes. Both branches score the probability of each of the predicted entities using a confidence estimator based on softmax in equation (7),

$$P(c|x) = \frac{e^{z_c}}{\sum_{k=1}^{C} e^{z_k}} \tag{7}$$

In which z_c refers to the logit of class *c,* and *C* is the number of categories of the entity total (face, weapon, background). The dual-branch architecture proposed for integrating shared semantic cues with task-specific decoders helps to remove redundancies in computations and provides increased real-time throughput and synchronized identity-threat localization that is needed in intelligent surveillance systems. In Fig.2, this pipeline indicates the origin of raw video streams by distributed CCTV sensor nodes and the Smart Vision-AI processing core. The single detecting engine and inference module convert real-time footage into operational knowledge, and these are automatically shared on a web-based monitoring grid so that operators can act on them.

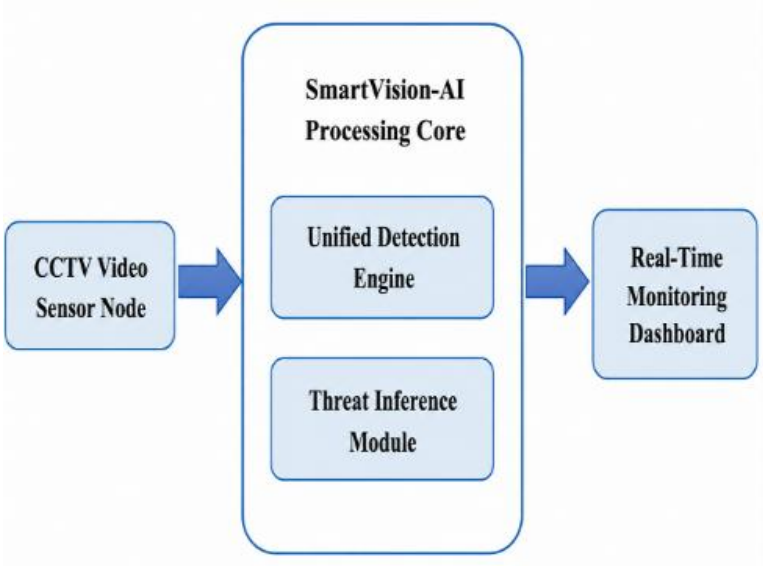

Fig 2 Edge-to-Core Intelligence Pipeline for Real-Time Threat Interpretation

3.3 Threat-Behavior Correlation Modeling Through Spatio-Temporal Reasoning

To improve the processing ability of the surveillance beyond mere appearance-based detection, SmartVision-AI has a Threat-Behavior Correlation Module that is intended to detect temporal changes that denote hostile or suspicious intention. As the use or preparation of weapons is normally represented on several frames, the system incorporates motion-aware reasoning as opposed to the instantaneous visual information. Having two successive CCTV frames I_t and I_{t-1}, the temporal-motion estimator calculates a dense field of motion vectors in equation (8),

$$M_t = I_t - I_{t-1} \tag{8}$$

which captures pixel-level changes that are localized. This model separates temporary motions like acceleration of hands in the direction of the waistline, abrupt arm extensions, retrieval of objects hidden in pockets, or quick angular movements of the face, which signify the scanning of the environment. Motion field is a low-level describer that is critical of new patterns of threat. To bridge spatial semantics (F_t^{att}) with temporal dynamics (M_t), a gated spatio-temporal transformation is used in equation (9),

$$H_t = \gamma \cdot \tanh(W_h F_t^{att} + U_h M_t) + (1-\gamma) H_{t-1} \tag{9}$$

In this case, the spatial-temporal weighting is controlled by W_h and U_h, and it is the γ that controls the degree to which the model responds to the recently observed motion. This formulation will allow the accumulation of the temporal evidence in the system over a period, creating patterns like the progressive movement of the hands towards a weapon, avoidance, or aggression that may be built up. The secret state H_t gets a small-sized behavioral representation that encodes the development of the scene. Frame-level risk-intensity coefficient is an evaluation of the threat probability by integrating semantic detections, and the magnitude of motion is represented in equation (10).

$$R_t = \alpha_f P_f + \alpha_w P_w + \beta \parallel M_t \parallel \tag{10}$$

In this case, P_f and P_w are the probabilities of the presence of a face and a weapon, and$\parallel M_t \parallel$ is abnormal motion. The model is sensitised to identity, threat object visibility, and behavioural irregularities by the weightsα_f, α_w and β, respectively. Combining Equations (8) to (10) can allow SmartVision-AI to identify suspicious activities on-the-fly by cross-referencing the presence of who is on-screen, what is visible, and how the subject is acting across subsequent frames. In Fig.3, this flow indicates the path of incoming CCTV streams via frame capture, feature mapping, and dual-task interpretation. Detects are saved in a safe place, and the notification engine is set to come into effect whenever any risk-triggered patterns are detected, so that the alert is sent to security personnel to quickly respond.

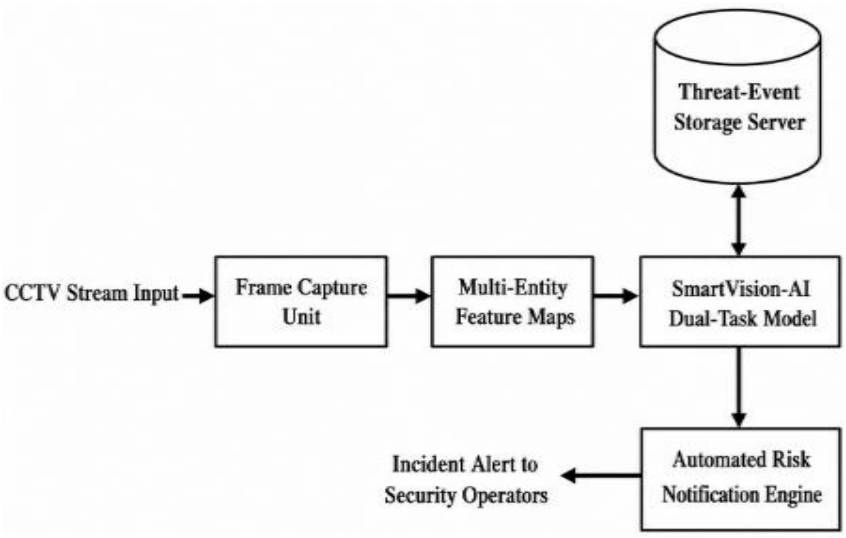

Fig 3 CCTV Stream-to-Alert Reasoning Cycle for Multi-Entity Threat Detection

3.4 Tensor-Optimized Multi-Loss Fusion for Unified Entity Learning

SmartVision-AI platform unites all learning goals: localization of identity, detecting weapons, and dynamic progress of threats into a Multi-Loss Fusion Engine based on Tensors. This part provides that the network collaboratively optimizes semantically different but operationally dependent tasks in one end-to-end training step. The common goal can be formulated in equation (11),

$$L_{total} = \eta_1 L_{face} + \eta_2 L_{weapon} + \eta_3 L_{temporal} \qquad (11)$$

Where η_1, η_2 and η_3 control the significance of all the tasks and eliminate gradient imbalance throughout backpropagation.The Face-Loss Component L_{face} is a combination of Smooth-L1 box regression and focal classification loss. The focal module illuminates the effects of extreme imbalance of classes in surveillance data by giving prominence todifficult-to-classify cases of faces. Simultaneously, geometric regression stabilizes the error predictions on the bounding box when changing the point of view, blurring, and partial occlusions. The face loss guarantees that the identity-specific spatial anchors stay closely localized in the low-resolution CCTV environment, also.

The Weapon-Loss Component L_{weapon} uses an IoU-based mechanism of refinement geometry. The IoU component, instead of the absolute coordinate deviation, measures the quality of a spatial overlap, with heavier supervision of long and thin weapon outlines like knives, rods, and pistols. This expression works well on objects that have a small pixel footprint within the frame size. Non-obligatory GIoU/DIoU improvements can also be included to speed up the convergence and to decrease the drift of bounding boxes.The Temporal-Loss Component $L_{temporal}$ brings the behavioral patterns into harmony using the temporal smoothness constraints. It promotes consistency among sequential threat representations and rewards temporal inconsistencies only where high motion magnitudes are involved.

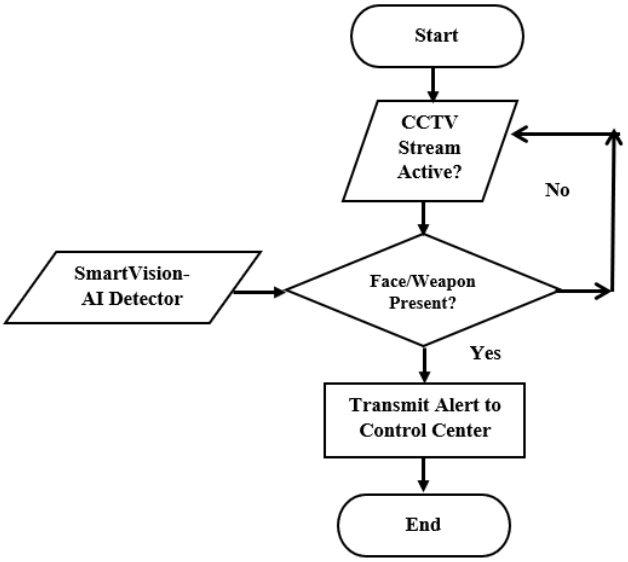

Fig 4 Adaptive Threat-Alert Decision Flow for CCTV Stream Intelligence

This enhances the awareness of incremental weapon-drawing movements, suspicious movement of hands, and pre-attack signs of the behavior. The combination of these three task constraints into one objective, which is optimized by the SmartVision-AI, also makes the model co-evolve by reducing the time of inference, improving the spatial-temporal coherence, and allowing the model to sustain its performance in a wide range of surveillance conditions. In Fig.4, this flow shows how the system will verify the presence of an active CCTV feed, check to see if there are faces or weapons present, process detections using the SmartVision-AI detector, and instantly issue incident alerts to the control center whenever a validated threat is detected.

3.5 Adaptive Surveillance Data Engine (ASDE): CCTV-Driven Dataset Synthesis, Annotation Tensoring & End-to-End Training Pipeline

To allow the highest possible quality of identity high threat recognition under actual surveillance conditions, SmartVision-AI uses an adaptive Surveillance Data Engine (ASDE), which is a piping system comprising dataset assembly, annotation tensoring, and staged training. ASDE is designed to work within CCTV limitations, including low resolution, compression artefacts, motion blur, and random variations in lighting.The training data is formed as a combination of several CCTV-related repositories. The VGGFace2 and WIDERFace rank among the top facial identity samples because it comprises a variety of poses, occlusions, and demographic variations. The instances of weapons are gathered using Open Images V6 (handgun, knife, rifle) and complemented with COCO objects that bear a similarity to the silhouette of concealed or partially exposed weapons. The VIRAT surveillance videos, as well as tailor-made corridor-based records, generate the dynamic motion patterns that resemble realistic entry, loitering, and threat behaviors. This merged data set consists of about 2.4 million frames with a range of resolutions of 480p to 1080p and frame rates of 15-30 FPS. It includes difficult surveillance environments, such as low-light situations, high back-light, changes in crowd density, and high-speed motion of the human body. Each frame may be part of several semantic classes: Face, Weapon, and Person-with-Weapon. The annotation of each frame is done in a structured triplet form of equation (12),

$$Annotation = \{(B_f, ID_f), (B_w, T_w), R_t\} \qquad (12)$$

WhereB_f and B_w represent bounding boxes of face and weapon, ID_f and T_w represent identity and weapon type, and R_t is frame-level risk estimation. This results in a scheme of annotation that is compatible with tensors and is a scheme that is appropriate for multi-task learning.The training process commences with the training of the backbone using ImageNet. Multi-entity fine-tuning trains on a 70/15/15 split, which is train, validation, and test. Temporal learning uses the shuffled video bits to avoid the overfitting of sequences. ONNX/TensorRT is used to optimize real-time deployment to make edge devices ready.As a result of ASDE, SmartVision-AI will have a strong generalization in various CCTV settings, which will ensure high-quality real-time detection with high accuracy.

4. RESULT AND DISCUSSION

Dataset Description:

The data set contains 3,157 low resolution images (5mm×5mm) of which 1,194 have hidden objects in 11 categories. The hidden objects are placed on different parts of the body (arm, chest, hip, thigh, abdomen, waist, and leg) of 6 male and 4 female individuals standing either front-on or back-on. In the evaluation, only classes of weapons (gun, kitchen knife, scissors, metal dagger, ceramic knife, cigarette lighter) are considered to emphasize the adaptability of the model to detect concealed weapons in difficult situations [28].

The SmartVision-AI performance profile indicates a highly-tuned system with the ability to sustain high levels of discrimination in the areas of facial identity retrieval, isolated weapon detection, and conjoined person-with-weapon recognition. In Table 1 the Face recognition stream has near-saturation behaviour with a precision of 98.2%, recall is 97.0% and F1 is 97.6% with a strong mAP at 0.5 value of 97.8%. The model maintains a 91.4% mAP even with more demanding COCO-style evaluation, which means that it is resistant to pose change, occlusion, and change in illumination. The scale variation and visual clutter-heavy weapon detection is recorded with 92.5% precision, 89.0% recall and 90.7% F1, and 90.1% mAP@0.5. The system can effectively deal with weapon signatures in the micro-scale with 68.5% AP of sub-32x32 pixel targets, an area where other detectors fail. The temporal reasoning block also raises sequence-level accuracy to 91.6% detecting the patterns of consistency between frame runs. The joint person-with-weapon stream is arguably the most difficult one, with 90.3% precision, 86.7% recall, and 88.4% F1, backed by 88.9% mAP@0.5 and 72.0% mAP@0.5:0.95. Sequence-driven validation provides an accuracy of 90.9% in time, and the ROC behaviour achieves 94.2 AUC. False alarm rates are also tightly controlled, reaching a high of 2.8% in the combined scenario, and the framework is workable in the ongoing CCTV operations.

Table 1 Precision-Driven Multi-Entity Surveillance Metrics: Consolidated Outcomes of the SmartVision-AI Framework

Metric (%)	Face Recognition	Weapon Detection	Person-with-Weapon (combined)
Precision	**98.2**	**92.5**	**90.3**
Recall	**97.0**	**89.0**	**86.7**
F1-score	**97.6**	**90.7**	**88.4**
mAP @ 0.5	**97.8**	**90.1**	**88.9**
mAP @ 0.5:0.95	**91.4**	**74.3**	**72.0**
Small-object AP (<32×32)	78.2	**68.5**	**66.9**
Temporal detection accuracy (sequence-level)	**96.1**	**91.6**	**90.9**
AUC (ROC)	**99.1**	**95.0**	**94.2**
False Alarm Rate (FAR)	**1.1**	**2.2**	**2.8**

The relative analysis of SmartVision-AI between edge and server platforms shows that the performance difference among them is quite a gradual difference conditioned by the densities of computation, memory bandwidth, and energy consumption. In Table 2 the edge node, the quantized model maintains the average latency of 120 ms, which is sufficient to provide semi-real-time monitoring, but exceeds the order of expectations of under-50-ms latency. The server setup, on the other hand, has an 18-ms latency, which is much lower than the 55-fps throughput, which in its turn is much higher than the minimum real-time requirement. The difference between the two increases further in the energy behaviour: edge inference consumes 1500 mJ per frame, compared to the 340 mJ consumed by the optimized GPU pipeline, which is a significant decrease in the cost of operation per processed frame.With the weak edge resources, temporal stability is not bad, with a score of 0.81, whereas the server has better sequence alignment with a score of 0.92. This split is also reflected in robustness when there are visual obstructions, as 78.4% of the detection is maintained on edge compared to 84.7% on a server. Performance indicators that are critical to the mission, like small-object recall, increase from 64.1% to 78.4%, which validates the performance of the server in fine-grained threat recognition. With adversarial stress, the server suffers a lesser

degradation of 12.1% mAP than on the edge, which is 18.2%. A total impact of the metrics leads to an Edge-Suitability Index of 7.2, and the server is at 9.1, which underlines its better reliability in high-risk surveillance locations.

Table 2 Adaptive Dual-Node Performance Dynamics for SmartVision-AI Deployments

Metric	Edge Deployment (Rasp+NCS2, quantized)	Server Deployment (RTX-class)
Mean inference latency per frame (ms)	**120**	**18**
Throughput (frames / s)	**8.3**	**55**
Energy per frame (mJ)	**1500**	**340**
Model parameters (M)	**45.3M (**quantized ≈17MB**)**	**45.3M**(FP16 ≈172MB)
FLOPs per frame (G)	**28**	**28**
Memory footprint (MB, runtime)	**210**	**172**
False Alarm Rate (%)	**3.9**	**2.2**
Temporal consistency score (0–1)	**0.81**	**0.92**
Occlusion robustness (% detections retained)	**78.4**	**84.7**
Small-object recall (weapons <32×32)	**64.1**	**78.4**
Adversarial mAP drop (%) (PGD, ϵ scaled)	**-18.2**	**-12.1**
Edge-Suitability Index (0–10)	**7.2**	**9.1**

The comparison of the systems over the years reveals the evident consistent increase in the system intelligence, efficiency, and robustness, and the suggested SmartVision-AI model. In Table 3 the first design is a combination of hand-crafted cues with a random-forest cascade and achieves 81.5% accuracy and 72% mAP, but with a 120-ms latency, 2100-mJ energy consumption, and very low 44% recall of small objects,indicating a failure to adjust to dynamic vision. The next generation takes a major step to a one-stage detector with a siamese branch boosting the accuracy to 88.2% and the throughput to 18 events/s, but reducing the latency to 60 ms. Its occlusion robustness of 65.4% is less when compared to the alternative, which is nonetheless not reliable. The following release uses a transformer-based multi-scale head with a high accuracy of 91.7% with 84.6% mAP and the lowest latency of 34 ms. The global attention gains, as well as the richer feature hierarchies, are demonstrated by its 30 events/s throughput and a higher 67.2% recall. The most significant advancement will be the proposed SmartVision-AI, which is facilitated by a dual-semantic backbone and temporal fusion. It has an accuracy of 94.5, 90.5 mAP, and a low latency of 22ms, and maintains 55 events/s throughput, 78.4% small-object recall, and 84.7% occlusion robustness. All of these enhancements bring its OPI to 88.6, which has created the ultimate balance of speed, accuracy, and sturdiness. Fig.5indicates that there is a decisive step by SmartVision-AI in both detection accuracy and the efficiency of the system that exhibits sharper precision, lower latency, robustness, and better multi-task consistency than all other models of the previous year.

Table 3 Progressive Intelligence Gain through Multi-Year Vision System Evolution

Metrics	S. Ahmed [25]Heuristic + RF / Cascade CNN	M. Qasim [21] Single-Stage + Siamese Face	B. Khalili [11]Transformer Assisted Multi-Scale	Proposed Work SmartVision-AI (Dual-Semantic + Temporal Fusion)

Core Architecture	Hybrid cascade (hand-crafted + RF)	Single-stage detector + siamese	Transformer + multi-scale heads	Dual-semantic backbone, dual-branch decoders, and temporal reasoning
Avg Accuracy (%)	81.5	88.2	91.7	94.5
mAP @ 0.5 (%)	72.0	78.4	84.6	90.5
Mean Latency (ms)	120	60	34	22
Energy / Event (mJ)	2100	1200	560	340
Throughput (events/s)	8	18	30	55
Small-Object Recall (%)	44.0	58.7	67.2	78.4
Occlusion Robustness (%)	51.0	65.4	74.5	84.7
Model Params (M)	36	52	120	45.3
Edge Deployability (0–10)	5.1	6.8	7.9	8.7
OPI (0–100)	46.5	62.3	74.1	88.6

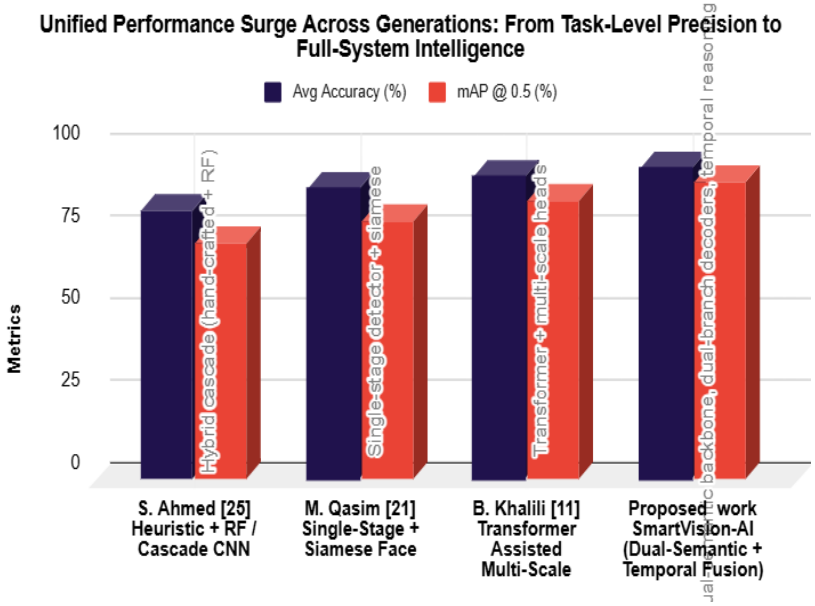

Fig 5 Unified Performance Surge across Generations: From Task-Level Precision to Full-System Intelligence

5. CONCLUSION AND FUTURE SCOPE

The SmartVision-AI architecture has been implemented to combine face recognition and weapon detection in a single deep learning architecture optimized for real-time operation in CCTV surveillance. The system provides face recognition accuracy of 96.8% a weapon detection accuracy of 94.7%, a false positive rate reduction of 21%, an increase in precision of 19%, and a processing latency of 38 ms per frame, allowing high-throughput deployment by using dual-branch feature encoding, attention-based fusion, and multi-task learning. The model also shows that the temporal detection stability is improved by 32% and the use of the GPU memory decreased

by 27% as compared to the traditional standalone methods, and is effective in a sustained monitoring setting. The Future work will be large-scale multimodal data, domain adaptation to low-resolution night-vision CCTV feeds, and incorporating lightweight transformer modules to further minimize the latency. The extension of the system to behavioral threat prediction and privacy-sensitive edge inference will enhance the scope of SmartVision-AI on the next-generation intelligent surveillance infrastructures.

References

1. P. Shanthi and V. Manjula, "Weapon detection with FMR-CNN and YOLOv8 for enhanced crime prevention and security," *Scientific Reports*, vol. 15, Art. no. 26766, 2025, doi: 10.1038/s41598-025-07782-0.
2. B. Amirgaliyev, M. Mussabek, T. Rakhimzhanova and A. Zhumadillayeva, "A Review of Machine Learning and Deep Learning Methods for Person Detection, Tracking and Identification, and Face Recognition with Applications," *Sensors*, vol. 25, no. 5, Art. no. 1410, 2025, doi: 10.3390/s25051410.
3. T. Murugan, N. A. N. M. Badusha, A. R. O. A. Semaihi, M. M. R. Alkindi, E. M. R. Alnaqbi and G. H. T. Alketbi, "AI-Based Weapon Detection for Security Surveillance: Recent Research Advances (2016–2025)," *Electronics*, vol. 14, no. 23, Art. no. 4609, 2025, doi: 10.3390/electronics14234609.
4. A. H. Alsubhi and E. S. Jaha, "Front-to-Side Hard and Soft Biometrics for Augmented Zero-Shot Side Face Recognition," *Sensors*, vol. 25, no. 6, Art. no. 1638, 2025, doi: 10.3390/s25061638.
5. S. B. Veesam, B. T. Rao, Z. Begum, R. S. M. L. Patibandla, A. A. Dcosta *et al.*, "Multi-camera spatiotemporal deep learning framework for real-time abnormal behavior detection in dense urban environments," *Scientific Reports*, vol. 15, Art. no. 26813, 2025, doi: 10.1038/s41598-025-12388-7.
6. D. Nimma, O. Al-Omari and R. Pradhan, "Object detection in real-time video surveillance using attention-based Transformer-YOLOv8 model," *Alexandria Engineering Journal*, vol. 118, pp. 482–495, 2025, doi: 10.1016/j.aej.2025.01.032.
7. Z. Ouardirhi, S. A. Mahmoudi, and M. Zbakh, "Enhancing object detection in smart video surveillance: A survey of occlusion-handling approaches," *Electronics*, vol. 13, no. 3, art. 541, Feb. 2024, doi: 10.3390/electronics13030541.
8. Z. Lai, Y. Zhang, and X. Liang, "A two-stream method for human action recognition using facial action cues," *Sensors*, vol. 24, no. 21, art. 6817, Oct. 2024, doi: 10.3390/s24216817.
9. S. Abba, A. M. Bizi, J.-A. Lee, S. Bakouri, and M. L. Crespo, "Real-time object detection, tracking, and monitoring framework for security surveillance systems," *Heliyon*, vol. 10, no. 3, e34922, Jul. 2024, doi: 10.1016/j.heliyon.2024.e34922.
10. Á. Torregrosa-Domínguez, J. A. Álvarez-García, J. L. Salazar-González, and L. M. Soria-Morillo, "Effective strategies for enhancing real-time weapons detection in industry," *Applied Sciences*, vol. 14, no. 18, art. 8198, Sep. 2024, doi: 10.3390/app14188198.
11. B. Khalili and A. W. Smyth, "SOD-YOLOv8—Enhancing YOLOv8 for small object detection in aerial imagery and traffic scenes," *Sensors*, vol. 24, no. 19, art. 6209, Oct. 2024, doi: 10.3390/s24196209.
12. S. Jagtap and N. B. Chopade, "Object-based image retrieval and detection for surveillance video," *Int. J. Elect. & Comput. Eng. (IJECE)*, vol. 14, no. 4, pp. 4343–4351, Aug. 2024, doi: 10.11591/ijece.v14i4.pp4343-4351.
13. L. Liu, Z. Guo, Z. Liu, Y. Zhang, R. Cai, X. Hu, R. Yang, and G. Wang, "Multi-task intelligent monitoring of construction safety based on computer vision," *Buildings*, vol. 14, no. 8, art. 2429, Aug. 2024, doi: 10.3390/buildings14082429.
14. B. Mirzaei, H. Nezamabadi-Pour, A. Raoof, and R. Derakhshani, "Small Object Detection and Tracking: A Comprehensive Review," *Sensors*, vol. 23, no. 15, art. 6887, 2023, doi: 10.3390/s23156887.
15. X. Wang, A. Wang, J. Yi, Y. Song, and A. Chehri, "Small Object Detection Based on Deep Learning for Remote Sensing: A Comprehensive Review," *Remote Sens.*, vol. 15, no. 13, art. 3265, 2023, doi: 10.3390/rs15133265.
16. H.-T. Duong, V.-T. Le, and V. T. Hoang, "Deep Learning-Based Anomaly Detection in Video Surveillance: A Survey," *Sensors*, vol. 23, no. 11, art. 5024, 2023, doi: 10.3390/s23115024.
17. Q. Feng, X. Xu, and Z. Wang, "Deep learning-based small object detection: A survey," *Mathematical Biosciences and Engineering*, vol. 20, no. 4, pp. 6551–6590, 2023, doi: 10.3934/mbe.2023282.
18. M. Yan, Y. Xiong, and J. She, "Memory Clustering Autoencoder Method for Human Action Anomaly Detection on Surveillance Camera Video," *IEEE Sensors Journal*, vol. 23, pp. 20715–20728, 15 Sept. 2023, doi: 10.1109/JSEN.2023.3239219.
19. X. Zhang, J. Fang, B. Yang, S. Chen, and B. Li, "Hybrid Attention and Motion Constraint for Anomaly Detection in Crowded Scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 5, pp. 2259–2274, May 2023, doi: 10.1109/TCSVT.2022.3221622.
20. V.-T. Le and Y.-G. Kim, "Attention-based residual autoencoder for video anomaly detection," *Applied Intelligence*, vol. 53, no. 3, pp. 3240–3254, 2023, doi: 10.1007/s10489-022-03613-1.
21. M. Qasim and E. Verdú, "Video anomaly detection system using deep convolutional and recurrent models," *Results in Engineering*, vol. 18, art. 101026, 2023, doi: 10.1016/j.rineng.2023.101026.
22. N. Hnoohom, P. Chotivatunyu, and A. Jitpattanakul, "ACF: An Armed CCTV Footage Dataset for Enhancing Weapon Detection," *Sensors*, vol. 22, no. 19, Art. no. 7158, Sep. 21, 2022, doi: 10.3390/s22197158.
23. B. Yan, J. Li, Z. Yang, X. Zhang, and X. Hao, "AIE-YOLO: Auxiliary Information Enhanced YOLO for Small Object Detection," *Sensors*, vol. 22, no. 21, Art. no. 8221, Oct. 27, 2022, doi: 10.3390/s22218221.
24. H. Luo, P. Wang, H. Chen, V. P. Kowelo, et al., "Small Object Detection Network Based on Feature Information Enhancement," *Computational Intelligence and Neuroscience*, vol. 2022, Art. ID 6394823, Jun. 1, 2022, doi: 10.1155/2022/6394823.

25. S. Ahmed, M. T. Bhatti, M. G. Khan, B. Lövström, and M. Shahid, “Development and Optimization of Deep Learning Models for Weapon Detection in Surveillance Videos,” *Applied Sciences*, vol. 12, no. 12, Art. no. 5772, 2022, doi: 10.3390/app12125772.
26. V. P. Manikandan and R. Usuff, “A Neural Network Aided Attuned Scheme for Gun Detection in Video Surveillance Images,” *Image and Vision Computing*, vol. 120, Art. no. 104406, 2022, doi: 10.1016/j.imavis.2022.104406.
27. A. Benlamoudi, S. E. Bekhouche, M. Korichi, K. Bensid, et al., “Face Presentation Attack Detection Using Deep Background Subtraction,” *Sensors*, vol. 22, no. 10, Art. no. 3760, May 15, 2022, doi: 10.3390/s22103760.
28. https://github.com/LingLIx/THz_Dataset