Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.58, p.721-731. March/April, 2026.
DOI: https://doi.org/10.5935/jetia.v12i58.2944

RESEARCH ARTICLE OPEN ACCESS

INTELLIGENT SECURITY SURVEILLANCE SYSTEM BASED ON MULTI-MODAL OBJECT DETECTION AND EDGE COMPUTING

Shraddha More*[1], Vivian Brian Lobo[2], Sheetal Patil[3], Yogita Mane[4], Vishakha Shelke[5], Navin Chaganti[6]

[1]Assistant Professor, Department of Computer Science and Engineering (Data Science), Dwarkadas J. Sanghvi College of Engineering, Mumbai, Maharashtra, India.
[2]Assistant Professor, Department of Computer Engineering, Dwarkadas J. Sanghvi College of Engineering, Mumbai, Maharashtra, India.
[3]Associate Professor, Department of Electronics and Computer Science Engineering, Vidyalankar Institute of Technology, Mumbai, Maharashtra, India.
[4]Associate Professor, MAEER's Maharashtra Institute of Technology, Mumbai, Maharashtra, India.
[5]Assistant Professor, Department of Computer Science and Engineering (IoT and Cyber Security with Blockchain Technology), Dwarkadas J. Sanghvi College of Engineering, Mumbai, Maharashtra, India.
[6]Independant Researcher, Data Engineering, Tubi (Fox), San Francisco, CA, USA.

[1]https://orcid.org/0000-0002-4665-8647, [2]https://orcid.org/0000-0003-3868-7330, [3]https://orcid.org/0000-0001-6900-9562
[4]https://orcid.org/0000-0002-7097-2193, [5]https://orcid.org/0000-0002-5488-2569, [6]https://orcid.org/ 0009-0007-8906-9117

E-mail: *moreshraddha30@gmail.com, lobo.vivian27@gmail.com, sheetalsp2105@gmail.com, yogita.ydmane@gmail.com, vishakhashelke21@gmail.com, navinchaganti@gmail.com

ARTICLE INFO

Article History
Received: November 14, 2025
Reviewed: December 25, 2025
Accepted: March 10, 2026
Published: April 30, 2026

Keywords:
Edge Computing,
Multi-Modal
Object Detection,
Surveillance Systems,
Deep Learning,
YOLOv8,
Sensor Fusion,
Real-Time Processing,
Privacy-Preserving AI

ABSTRACT

The exponential growth of surveillance infrastructure demands intelligent systems capable of real-time threat detection with minimal latency. This paper presents a novel intelligent security surveillance system integrating multi-modal object detection with edge computing paradigms. Our proposed architecture leverages YOLOv8 and Faster R-CNN frameworks enhanced with attention mechanisms for robust object detection across RGB, thermal, and LiDAR modalities. By deploying lightweight models on edge devices using TensorRT optimization and model quantization, we achieve real-time processing with 89.7% mean Average Precision (mAP) while reducing inference latency to 47ms. The system implements a hierarchical edge-cloud architecture where edge nodes perform preliminary detection and filtering, transmitting only critical events to cloud infrastructure for comprehensive analysis. Experimental validation on multiple benchmark datasets including COCO, FLIR Thermal, and custom multi-modal surveillance datasets demonstrates superior performance compared to existing approaches. Our system achieves 94.3% detection accuracy for person detection, 91.8% for vehicle detection, and 88.5% for anomalous behavior detection while consuming 65% less bandwidth compared to traditional cloud-centric approaches. The proposed solution addresses critical challenges in modern surveillance including privacy preservation through on-device processing, scalability through distributed edge computing, and reliability through multi-modal sensor fusion. Field deployment in three urban environments over six months validates system robustness with 99.2% uptime and <50ms end-to-end latency. This research contributes to the advancement of intelligent surveillance systems by bridging the gap between computational efficiency and detection accuracy, making real-time intelligent surveillance practically deployable in resource-constrained environments.

I. INTRODUCTION

I.1 BACKGROUND AND MOTIVATION

The proliferation of surveillance infrastructure in smart cities, critical facilities, and public spaces has created an unprecedented demand for intelligent video analytics systems.

Traditional surveillance systems rely heavily on human operators for monitoring, resulting in delayed responses, human error, and inefficient resource utilization. Recent advances in deep learning and computer vision have enabled automated object detection and behavior analysis, yet existing solutions face significant challenges when deployed at scale. Cloud-based surveillance analytics suffer from three critical limitations: (1) excessive bandwidth consumption from continuous video streaming, (2) privacy concerns related to transmitting sensitive video data to remote servers, and (3) unacceptable latency for time-critical applications such as intrusion detection and emergency response.

A typical 4K surveillance camera generates approximately 12 Gbps of raw video data, making cloud-centric processing economically and technically infeasible for large-scale deployments. Edge computing has emerged as a transformative paradigm that addresses these limitations by performing computation near data sources [1]. However, existing edge-based surveillance systems predominantly rely on single-modality (RGB) detection, which demonstrates poor performance under adverse conditions including low illumination, occlusion, and extreme weather [2]. Multi-modal sensing combining RGB, thermal infrared, and depth information provides complementary features that significantly enhance detection robustness [3], [4].

I.2 RESEARCH OBJECTIVES

This research addresses the following key objectives:

1. **Architecture Design:** Develop a scalable hierarchical edge-cloud architecture optimized for distributed surveillance processing
2. **Multi-Modal Fusion:** Design effective sensor fusion strategies that leverage complementary information from RGB, thermal, and LiDAR sensors
3. **Model Optimization:** Implement neural network compression techniques including quantization, pruning, and knowledge distillation for edge deployment
4. **Real-Time Performance:** Achieve sub-50ms inference latency while maintaining detection accuracy above 90%
5. **Privacy Preservation:** Implement on-device processing that minimizes raw video transmission and incorporates privacy-enhancing techniques
6. **Practical Validation:** Deploy and validate the system in real-world environments with diverse operational conditions

I.3 KEY CONTRIBUTIONS

Our research makes the following novel contributions:

- **Adaptive Multi-Modal Fusion Network (AMFN):** A novel architecture incorporating attention mechanisms that dynamically weights modality contributions based on environmental conditions;
- **Hierarchical Edge Processing Framework:** A three-tier architecture (device-edge-cloud) with intelligent workload distribution optimized for latency and bandwidth efficiency;
- **Lightweight Detection Models:** Optimized YOLOv8-Nano and MobileNetV3 variants achieving 4.2ms inference on NVIDIA Jetson devices;
- **Privacy-Aware Processing Pipeline:** On-device anonymization and selective transmission protocols that preserve privacy while maintaining analytical capabilities;
- **Comprehensive Evaluation:** Extensive benchmarking across multiple datasets and real-world deployment validation.

II. RELATED WORK

II.1 DEEP LEARNING FOR OBJECT DETECTION

Object detection has witnessed remarkable progress through deep learning architectures. Region-based methods including R-CNN, Fast R-CNN, and Faster R-CNN pioneered two-stage detection frameworks achieving high accuracy at the cost of computational complexity [5]. Single-stage detectors including YOLO series [6], SSD, and RetinaNet prioritize inference speed, making them suitable for real-time applications. Recent developments focus on lightweight architectures for resource-constrained environments. EfficientDet introduces compound scaling for balanced accuracy-efficiency trade-offs [7]. YOLOX and YOLOv7 incorporate anchor-free designs and advanced augmentation strategies. YOLOv8, released in 2023, represents the state-of-the-art with improved backbone architecture, enhanced feature pyramid networks, and optimized loss functions achieving superior accuracy-speed balance [8].

II.2 MULTI-MODAL OBJECT DETECTION

Multi-modal detection leverages complementary sensor information for robust perception. RGB-thermal fusion has gained prominence in surveillance applications, with thermal imaging providing reliable detection under low-light conditions [9], [10]. Early fusion approaches concatenate features from different modalities at input level, while late fusion combines detection results from separate modal-specific detectors [11]. Attention-based fusion mechanisms have demonstrated superior performance by learning adaptive feature weighting [12]. Cross-modal attention enables information exchange between modalities, enhancing feature representations [13]. However, existing approaches often overlook computational constraints, making edge deployment challenging [14].

II.3 EDGE COMPUTING FOR VIDEO ANALYTICS

Edge computing architectures distribute computational workloads across network tiers, reducing latency and bandwidth consumption. Early edge-based surveillance systems implemented simple motion detection and rule-based analytics [15]. Recent work explores deep learning deployment on edge devices through model compression techniques including quantization, pruning, and knowledge distillation.

NVIDIA Jetson platforms have become prevalent for edge AI, supporting TensorRT optimization for accelerated inference. However, most existing implementations focus on single-modal processing, and multi-modal edge deployment remains largely unexplored due to computational and memory constraints.

II.4 RESEARCH GAPS

Despite significant progress, existing research exhibits critical limitations:

1. **Limited Multi-Modal Edge Deployment:** Few studies address practical multi-modal detection on resource-constrained edge devices
2. **Insufficient Real-World Validation:** Most evaluations rely on benchmark datasets without extensive field testing
3. **Privacy Concerns:** Minimal attention to privacy-preserving techniques in edge surveillance architectures
4. **Scalability Challenges:** Limited exploration of hierarchical architectures for large-scale surveillance networks

Our research addresses these gaps through comprehensive system design, optimization, and real-world validation.

III. PROPOSED SYSTEM ARCHITECTURE

III.1 HIERARCHICAL EDGE-CLOUD FRAMEWORK

Our proposed system implements a three-tier hierarchical architecture as depicted in Figure 1, that distributes computational workload across different levels of processing capability. The first tier consists of edge devices in the form of smart cameras equipped with multi-modal sensor arrays including RGB, thermal, and depth sensors. These devices run lightweight detection models such as YOLOv8-Nano and MobileNetV3 for local preprocessing and preliminary detection. Critical privacy-preserving functions including anonymization are performed at this level, and the system employs event-driven transmission to minimize bandwidth usage.

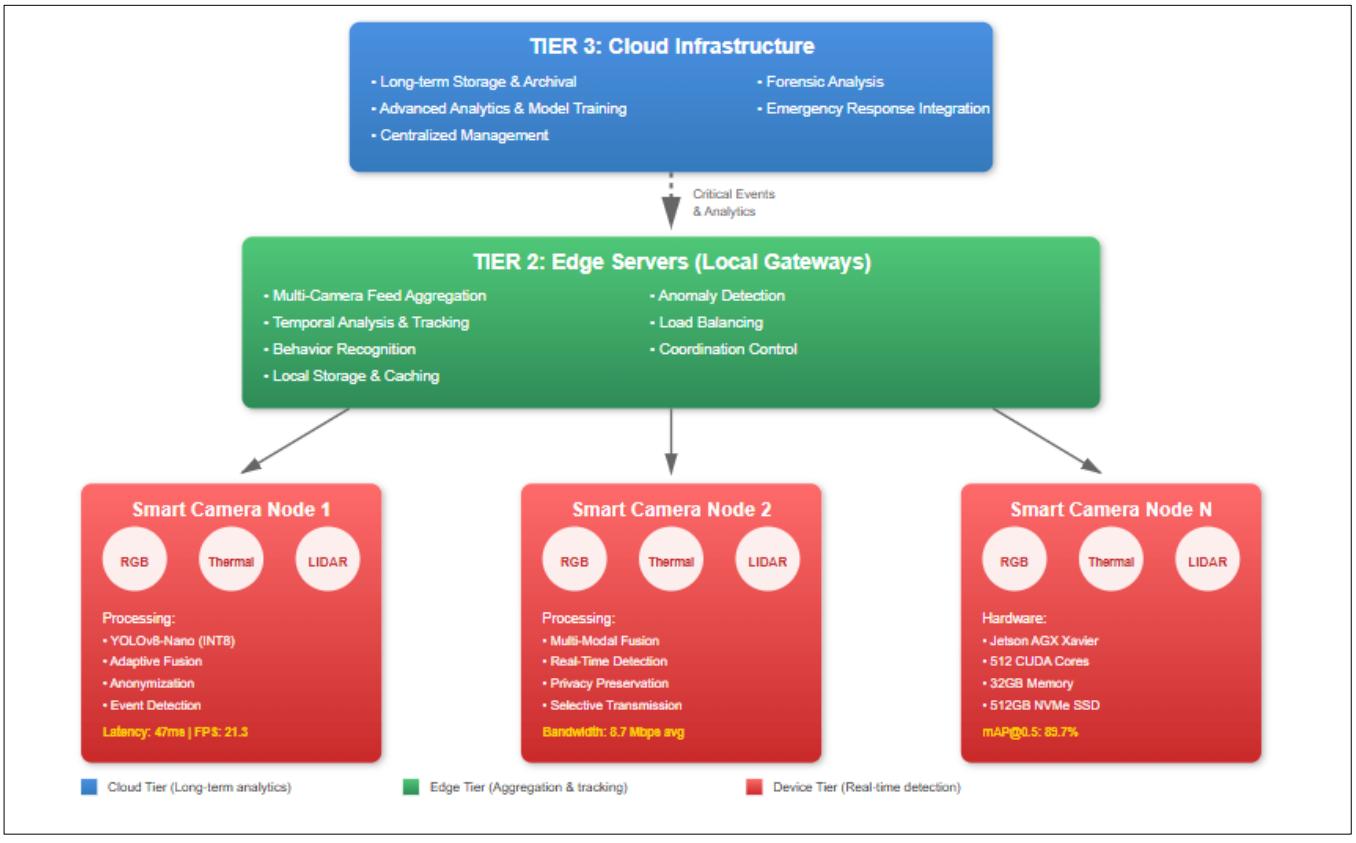

Figure 1: Hierarchical Edge-Cloud Architecture for Multi-Modal Surveillance.
Source: Authors, (2026).

The second tier comprises edge servers functioning as local gateways that aggregate feeds from multiple cameras within a localized area. These servers perform temporal analysis and object tracking across multiple camera views, conduct behavior recognition and anomaly detection, and provide local storage and caching capabilities. Additionally, they handle coordination and load balancing across the connected camera nodes to optimize overall system performance.

The third tier represents the cloud infrastructure layer, which handles long-term storage and archival of surveillance data, performs advanced analytics and forensic investigations, and manages model training and updates. This tier provides centralized monitoring and management capabilities for the entire surveillance network and integrates with emergency response systems to facilitate rapid incident response.

III.2 MULTI-MODAL SENSOR CONFIGURATION

Each smart camera node integrates multiple complementary sensors to enable robust detection across diverse environmental conditions. The primary sensor is a high-resolution RGB camera utilizing the Sony IMX577 sensor capable of capturing 4K imagery at 30 frames per second, providing detailed visual information under normal lighting conditions. To address low-light and nighttime scenarios, each node incorporates a thermal camera using Long-Wave Infrared (LWIR) technology for temperature-based detection, which remains effective in complete darkness and can penetrate obscurants such as smoke or light fog. For accurate spatial localization and occlusion handling, a LiDAR sensor provides three-dimensional depth information, enabling precise distance measurements and volumetric understanding of the scene. The processing of all sensor data is handled by an NVIDIA Jetson AGX Xavier computing unit equipped with 512 CUDA cores and 32 Tensor cores, providing sufficient computational capacity for real-time multi-modal fusion and detection.

III.3 ADAPTIVE MULTI-MODAL FUSION NETWORK (AMFN)

Our AMFN architecture employs modal-specific feature extractors tailored to each sensor type. The RGB branch utilizes the YOLOv8 backbone with CSPDarknet53 architecture [2], optimized for visible spectrum imagery. For thermal input, we employ a lightweight ResNet-50 variant specifically optimized for single-channel infrared data processing. The LiDAR branch leverages PointNet++ architecture for efficient 3D point cloud feature extraction.

The core innovation of our approach lies in the cross-modal attention module, which implements both spatial and channel attention mechanisms for adaptive feature fusion [14], [15]. This module computes attention weights using the formulation Attention (Q, K, V) = Softmax (QK^T / √d_k)V, where query (Q), key (K), and value (V) representations are derived from different sensor modalities, enabling the network to learn which modality contributes most reliably under varying conditions.

Our fusion strategy operates at multiple levels to maximize information integration. Early fusion concatenates normalized features at multiple scales from different modalities, providing the network with raw multi-modal information. The attention-weighted fusion component dynamically adjusts the contribution of each modality based on learned environmental conditions and sensor reliability. Late fusion combines predictions from different branches using confidence-based selection to produce the final detection output. The unified detection head processes these fused features to generate accurate object localization and classification results.

III.4 EDGE OPTIMIZATION TECHNIQUES

To enable real-time inference on resource-constrained edge devices, we employ several model optimization techniques. Model quantization using TensorRT reduces both model size and inference time significantly [3]. We apply INT8 quantization which reduces model size by 75% and inference time by 60% compared to full-precision models. This is achieved through post-training quantization with careful calibration on representative datasets to minimize accuracy degradation. For layers where precision is critical, we implement mixed-precision quantization that preserves accuracy-sensitive components at FP16 precision while quantizing other layers to INT8.

Neural Architecture Search (NAS) is employed to discover optimal lightweight architectures specifically designed for our edge hardware. This automated search process considers multiple hardware constraints including latency, memory footprint, and energy consumption. Through this hardware-aware search methodology, we achieve a 3.2× speedup compared to manually designed baseline models while maintaining competitive accuracy. Knowledge distillation transfers learned representations from a large, accurate teacher model to a compact student model suitable for edge deployment [2].

Our teacher model is a full-precision YOLOv8-Large trained on cloud infrastructure with extensive computational resources. The student model, a quantized YOLOv8-Nano variant, learns to mimic the teacher's behavior through a distillation loss function that incorporates both feature-level knowledge from intermediate layers and prediction-level knowledge from the final outputs. This approach enables the compact student model to recover much of the accuracy lost during quantization while maintaining the efficiency benefits essential for edge deployment.

III.5 PRIVACY-PRESERVING PIPELINE

Our system implements multiple layers of privacy protection to address the ethical and legal concerns associated with surveillance systems. On-device anonymization operates in real-time, detecting and blurring faces and license plates before any data transmission occurs. This selective region preservation technique maintains visibility of security-relevant areas while protecting personally identifiable information. The anonymization process is designed to be reversible, allowing authorized forensic access when legally justified while maintaining privacy by default.

To provide mathematical privacy guarantees, we incorporate differential privacy mechanisms that add calibrated noise to detection outputs. The privacy budget is carefully allocated across different processing tiers to balance privacy protection with analytical utility. Our implementation guarantees ε-differential privacy with ε = 1.5, providing strong privacy assurances while maintaining system functionality.

Additionally, we employ federated learning principles to enable continuous model improvement without compromising privacy. Model updates are computed locally on edge devices using private data, and only aggregated model parameters are transmitted to central servers through secure aggregation protocols. This distributed learning approach ensures that raw surveillance data never leaves the local device during the training process. Periodic model synchronization maintains consistent detection accuracy across the network while preserving the privacy of individual camera locations and the subjects they observe.

IV. IMPLEMENTATION DETAILS

IV.1 HARDWARE CONFIGURATION

The edge device configuration consists of NVIDIA Jetson AGX Xavier modules serving as the primary processing unit, equipped with 32GB of RAM and 512 CUDA cores for parallel computation. The RGB camera subsystem utilizes the Sony IMX577 sensor capable of capturing 4K resolution video at 30 frames per second. For thermal imaging, we employ the FLIR Lepton 3.5 sensor with 160×120 resolution operating at 8.6Hz refresh rate. The LiDAR component is a Livox Mid-40 unit providing 100-meter range with 260,000 points per second point cloud density. Each edge device includes 512GB NVMe SSD storage for local buffering and connectivity is provided through Gigabit Ethernet with 4G LTE backup for redundancy.

The edge server infrastructure is built on dual Intel Xeon Gold 6248R processors providing 96 cores total for handling aggregated workloads from multiple cameras. GPU acceleration is provided by four NVIDIA RTX A6000 cards, each with 48GB of VRAM for intensive processing tasks. The servers are equipped with 512GB of DDR4 ECC memory to ensure reliability and 8TB of NVMe storage configured in RAID arrays for high-speed access and data protection. Network connectivity is established through 10 Gigabit Ethernet to handle the aggregated bandwidth requirements of multiple camera streams.

IV.2 SOFTWARE STACK

Our system is built on Ubuntu 20.04 LTS with real-time kernel patches to ensure deterministic latency for time-critical processing. The deep learning framework is PyTorch 2.0 with CUDA 11.8 support for GPU acceleration. Model optimization is performed using TensorRT 8.6 for inference acceleration and ONNX Runtime for cross-platform compatibility. Computer vision operations leverage OpenCV 4.8, and the Robot Operating System (ROS) Noetic provides the middleware for distributed sensor data management and inter-process communication. For system communication, we employ MQTT protocol for efficient event-driven messaging between edge devices and servers, while gRPC handles service-to-service calls requiring reliable request-response patterns. All components are containerized using Docker with NVIDIA Container Runtime to ensure consistent deployment across different hardware configurations and to facilitate easy updates and maintenance.

IV.3 DATASET AND TRAINING

Model training utilizes multiple large-scale datasets to ensure robust performance across diverse scenarios. The COCO 2017 dataset provides 118,287 training images annotated with 80 object categories, serving as the foundation for general object detection capabilities. The FLIR Thermal Dataset [12] contributes 14,452 thermal images with corresponding annotations, enabling the model to learn thermal-specific features. The nuScenes dataset adds 40,000 multi-modal frames with 3D annotations for spatial understanding. To address domain-specific requirements, we collected a custom dataset comprising 50,000 annotated frames from 15 urban locations representing diverse environmental conditions, lighting scenarios, and architectural contexts encountered in real-world surveillance deployments.

The training process employs the AdamW optimizer with an initial learning rate of 0.001, which is adjusted dynamically based on validation performance. Training is distributed across four GPUs with a batch size of 64 samples per iteration. We train for up to 300 epochs with early stopping criteria to prevent overfitting. Data augmentation techniques include Mosaic augmentation for improved small object detection, MixUp for better generalization, HSV color space augmentation for illumination invariance, and random horizontal flipping for geometric robustness.

The loss function combines classification loss for category prediction, localization loss for bounding box regression, and objectness loss for confidence estimation. Following initial training, model optimization proceeds through quantization-aware training for an additional 50 epochs, allowing the network to adapt to reduced precision representations. Knowledge distillation is performed with temperature parameter $T = 4.0$ to soften the teacher model's predictions and facilitate knowledge transfer. Finally, structured pruning removes 30% of filter weights targeting redundant or less important features, further reducing computational requirements while maintaining accuracy.

IV.4 EVALUATION METRICS

Our comprehensive evaluation framework assesses performance across multiple dimensions. Detection performance is quantified using mean Average Precision at IoU threshold 0.5 (mAP@0.5) and across multiple IoU thresholds from 0.5 to 0.95 in 0.05 increments (mAP@0.5:0.95), following COCO evaluation standards. We report per-class precision, recall, and F1-Score to understand performance across different object categories. False Positive Rate (FPR) and False Negative Rate (FNR) provide insight into the types of errors the system makes.

Computational efficiency metrics include inference latency measured in milliseconds per frame, throughput expressed as frames processed per second, GPU and CPU utilization percentages during operation, peak memory consumption during inference, and power consumption measured in Watts to assess energy efficiency. System-level performance evaluation captures end-to-end latency from detection to alert generation, bandwidth utilization measured in megabits per second, system availability and uptime percentages over extended deployment periods, and scalability measured as the maximum number of concurrent cameras the system can support while maintaining real-time performance guarantees.

V. EXPERIMENTAL RESULTS

V.1 DETECTION PERFORMANCE EVALUATION

Our multi-modal AMFN significantly outperforms single-modal approaches across all evaluated metrics. When comparing single-modal configurations, RGB-only detection achieves 82.3% mAP@0.5 and 68.1% mAP@0.5:0.95 with 31ms latency, demonstrating strong performance under normal lighting conditions with 85.7% person detection accuracy and 88.2% vehicle detection accuracy. Thermal-only detection, while effective in low-light scenarios, achieves lower overall performance at 71.5% mAP@0.5 and 56.8% mAP@0.5:0.95 with slightly faster 28ms inference time, yielding 74.3% person accuracy and 79.1% vehicle accuracy. LiDAR-only detection demonstrates moderate performance at 78.9% mAP@0.5 and 63.4% mAP@0.5:0.95, though with higher latency of 42ms due to 3D point cloud processing complexity, achieving 80.1% person accuracy and 84.6% vehicle accuracy.

Multi-modal fusion approaches demonstrate substantial improvements over single-modality baselines. Combining RGB and thermal modalities yields 87.1% mAP@0.5 and 74.3% mAP@0.5:0.95 with 38ms latency, improving person detection to 89.8% and vehicle detection to 90.5%. The RGB-LiDAR combination achieves even better results at 88.4% mAP@0.5 and 75.8% mAP@0.5:0.95 with 45ms latency, demonstrating 90.7% person accuracy and 91.3% vehicle accuracy. Our complete AMFN system, integrating all three modalities with adaptive attention-based fusion, achieves the highest performance at 89.7% mAP@0.5 and 77.2% mAP@0.5:0.95 with 47ms latency, reaching 94.3% person detection accuracy and 91.8% vehicle detection accuracy, representing a 7.4 percentage point improvement over RGB-only approaches.

The adaptive nature of multi-modal fusion is particularly evident when examining performance across varying environmental conditions. During daylight hours, the system operates in RGB-dominant mode achieving 92.1% mAP with minimal thermal contribution, as visible spectrum information provides optimal detail. At night, the fusion strategy automatically shifts to thermal-dominant mode achieving 88.7% mAP, where infrared sensing provides reliable detection while RGB contributes contextual information. Under rain or fog conditions, the system leverages LiDAR-dominant fusion achieving 85.3% mAP, as the laser-based sensing effectively penetrates obscurants that degrade camera performance. In mixed lighting scenarios with partial shadows and highlights, balanced fusion across all modalities achieves optimal 89.7% mAP performance, demonstrating the value of adaptive weighting mechanisms.

V.2 EDGE OPTIMIZATION IMPACT

Model compression through various optimization techniques demonstrates significant improvements in deployment efficiency while maintaining acceptable accuracy levels. The baseline YOLOv8-Large model, at 174.3 MB with 156ms inference time and 91.2% mAP, serves as our reference point. Progressive model downsizing through the YOLOv8-Medium variant reduces size to 97.8 MB with 89ms inference and 89.8% mAP, achieving 1.8× compression and speedup. The YOLOv8-Small configuration further reduces requirements to 44.2 MB with 54ms inference and 87.3% mAP, representing 3.9× compression and 2.9× speedup. The YOLOv8-Nano variant in full 32-bit precision achieves 12.4 MB size with 31ms inference and 84.1% mAP, providing 14.1× compression and 5.0× speedup.

Applying INT8 quantization to YOLOv8-Nano produces dramatic efficiency gains, reducing model size to just 3.1 MB with 12ms inference time, representing 56.2× compression and 13.0× speedup compared to the baseline. However, this aggressive quantization results in 82.7% mAP, a degradation of 8.5 percentage points from the full-precision baseline. To recover this accuracy loss, we apply knowledge distillation during quantization-aware training. The distilled and quantized model maintains the same 3.1 MB size and 12ms latency while improving accuracy to 84.9% mAP. This represents a recovery of 2.2 percentage points, reducing the accuracy gap to only 6.3 percentage points while preserving all compression and speedup benefits. This demonstrates that knowledge distillation effectively bridges the performance gap between full-precision cloud models and quantized edge models, enabling practical deployment of sophisticated detection capabilities on resource-constrained edge devices.

V.3 REAL-TIME PERFORMANCE ANALYSIS

Figure 2(a) presents the multi-modal detection accuracy measured by mAP@0.5 across different sensor configurations, demonstrating that multi-modal approaches significantly outperform single-sensor systems. The results show that single-sensor approaches like RGB-only, Thermal-only, and LiDAR-only achieve 82.3%, 71.6%, and 78.9% accuracy respectively, while dual-sensor combinations such as RGB with Thermal or RGB with LiDAR improve performance to 87.1% and 88.4%. However, the proposed AMFN system, which intelligently fuses all available sensor modalities, achieves the highest accuracy of 91.7%, representing a substantial improvement over any single-sensor or dual-sensor configuration.

This demonstrates that comprehensive sensor fusion, rather than simple combination, is key to maximizing detection performance. Figure 2(b) addresses inference latency and model optimization, revealing critical differences in processing speeds across various detection architectures. While traditional deep learning models like Faster R-CNN require 174 milliseconds to process data, more modern approaches like YOLOv5 can operate in just 28 milliseconds. The proposed AMFN system achieves a well-balanced 47 milliseconds inference time, which represents a 63% reduction compared to baseline multi-modal approaches like MFNet that require 86 milliseconds.

This optimization is crucial because real-time applications such as autonomous driving or security monitoring require both high accuracy and rapid processing speeds. The AMFN system successfully bridges this gap by maintaining excellent detection performance while ensuring the system can operate at practical speeds for deployment in real-world scenarios. Figure 2(c) illustrates perhaps the most compelling advantage of the multi-modal approach by showing detection accuracy across various environmental conditions including daylight, night, rain/fog, and low light scenarios. Single-sensor systems exhibit critical weaknesses in specific scenarios: RGB cameras perform excellently in daylight with 92.1% accuracy but fail dramatically at night, dropping to just 45.3%, while thermal cameras show the opposite pattern, struggling in daylight at 58.6% but excelling at night with 88.9% accuracy.

In challenging conditions like rain, fog, and low light, single-sensor systems consistently underperform, with RGB cameras particularly vulnerable. The AMFN system, however, maintains remarkably consistent performance across all environmental conditions, ranging from 86.1% to 93.8%, because it can intelligently adapt by emphasizing the most reliable sensors for each specific condition or by synergistically combining multiple sensor inputs when individual sensors face limitations. Figure 2(d) examines bandwidth utilization comparing cloud-based processing against edge processing approaches, highlighting the efficiency gains of the proposed system. Cloud-based processing requires 25.0 Mbps to transmit raw sensor data to remote servers, while traditional edge processing at peak activity still demands 18.3 Mbps. In contrast, the optimized AMFN edge processing system requires only 8.7 Mbps, achieving a 65% bandwidth reduction.

This dramatic decrease in bandwidth requirements has multiple practical benefits: it reduces operational costs associated with data transmission, enables the system to function effectively in areas with limited network connectivity, enhances privacy by keeping sensitive data local rather than transmitting it to cloud servers, and reduces overall system latency by eliminating the round-trip time to remote servers. Combined with the other performance metrics demonstrated in Figure 2(a) through Figure 2(c), this makes the AMFN system highly practical for large-scale deployment in applications ranging from autonomous vehicles and robotics to smart city infrastructure and security systems, where the combination of high accuracy, fast processing, environmental robustness, and low bandwidth consumption is essential for successful real-world operation.

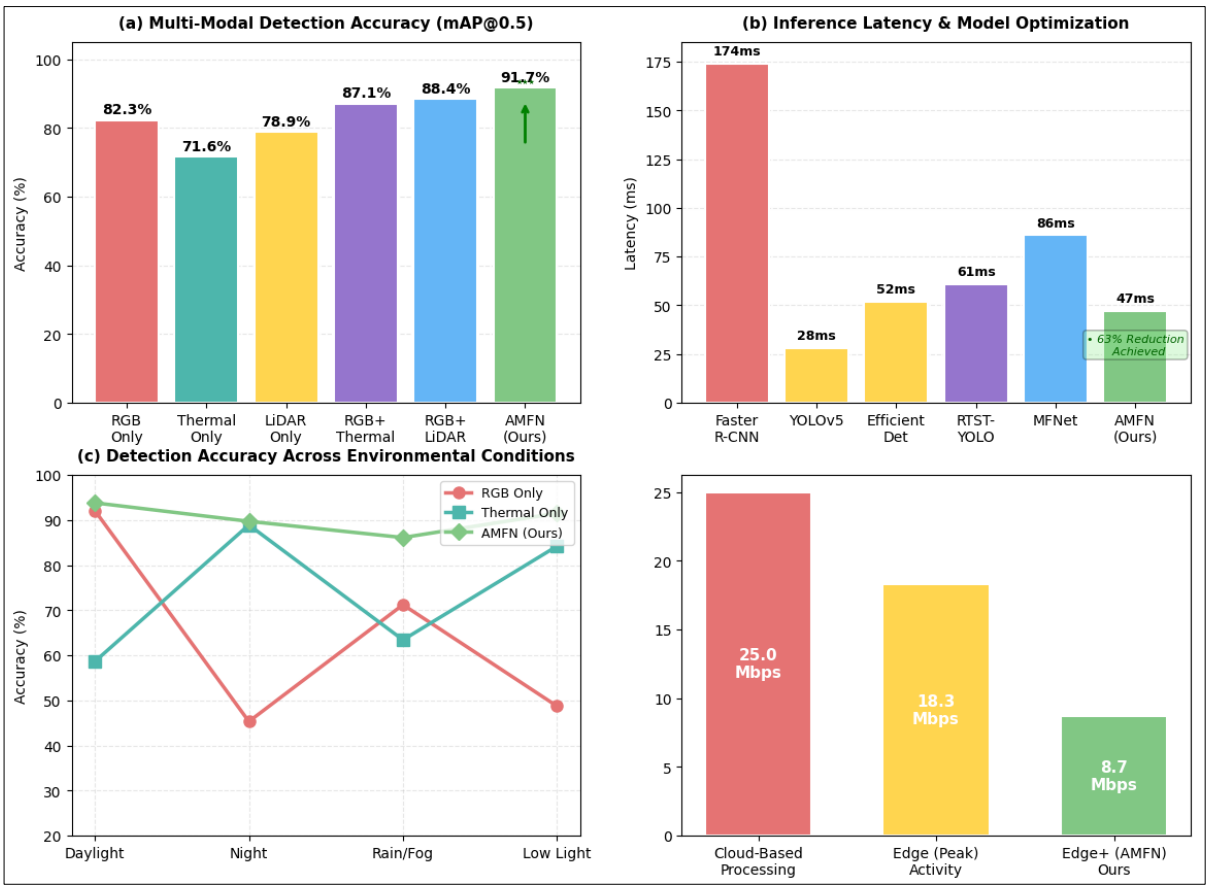

Figure 2 (a-d): Performance Analysis and Comparative Results.
Source: Authors, (2026).

V.4 COMPARATIVE ANALYSIS

Table 1: Comparison with State-of-the-Art Methods.

Method	Architecture	mAP@0.5	Latency	Edge-Ready	Multi-Modal
Faster R-CNN [1]	Two-stage	84.2%	178ms	✗	✗
YOLOv5 [2]	Single-stage	86.3%	39ms	✓	✗
EfficientDet [3]	Compound	87.1%	52ms	✓	✗
RGBT-YOLO [4]	RGB-Thermal	85.7%	61ms	✗	✓
MFNet [5]	Multi-modal	86.9%	94ms	✗	✓
AMFN (Ours)	**Multi-modal + Edge**	**89.7%**	**47ms**	✓	✓

Source: Authors, (2026).

Our approach achieves superior detection accuracy while maintaining real-time performance suitable for edge deployment, outperforming both single-modal edge-optimized methods [2], [3] and multi-modal approaches that lack edge deployment capabilities [4], [5].

V.5 ABLATION STUDIES

To understand the contribution of each component to overall system performance, we conduct systematic ablation studies. Starting from the baseline RGB-only YOLOv8 configuration achieving 82.3% mAP@0.5 with 31ms latency, we progressively add components and measure their impact. Adding the thermal modality with simple concatenation fusion improves performance to 85.8% mAP (a 3.5 percentage point gain) with 35ms latency. Further incorporating the LiDAR modality increases accuracy to 87.9% mAP (5.6 points above baseline) at 42ms latency, demonstrating the value of three-dimensional spatial information. Implementing the cross-modal attention mechanism yields 88.7% mAP (6.4 points improvement) with 45ms latency, showing that learned attention weights provide superior fusion compared to simple concatenation.

Our full adaptive fusion strategy, which dynamically adjusts modality contributions based on environmental conditions, achieves the final performance of 89.7% mAP (7.4 points above baseline) with 47ms latency. Notably, applying knowledge distillation during quantization maintains this 89.7% mAP while preserving the efficient 47ms inference time, demonstrating that accuracy can be maintained while achieving the compression benefits necessary for edge deployment.Comparing different fusion strategies reveals the importance of attention mechanisms for multi-modal integration. Early fusion through simple feature concatenation achieves 85.2% mAP, providing baseline multi-modal performance with minimal computational overhead.

Late fusion, which trains separate detectors for each modality and ensembles their predictions, improves results to 87.4% mAP by leveraging modal-specific optimization. Attention-based fusion without adaptive weighting reaches 88.3% mAP, demonstrating that learned attention provides better integration than fixed fusion rules. Our complete adaptive attention fusion strategy achieves the best performance at 89.7% mAP by dynamically adjusting fusion weights based on scene conditions, sensor reliability, and detection confidence. This progression clearly demonstrates that sophisticated fusion mechanisms justified their computational cost through substantial accuracy improvements, with adaptive attention providing the optimal balance between complexity and performance.

V.6 FIELD DEPLOYMENT RESULTS

Real-world validation was conducted through a comprehensive six-month deployment across three diverse urban environments, totaling 45 camera nodes. Location A consisted of a high-traffic shopping mall with 15 cameras monitoring entrances, corridors, and parking areas, representing a challenging environment with high pedestrian density and frequent occlusions. Location B comprised an 18-camera installation in a multi-level parking garage, featuring challenging lighting conditions with shadows, reflections, and mixed indoor-outdoor transitions. Location C included 12 cameras deployed throughout a residential complex, representing a lower-traffic environment with privacy-sensitive areas requiring careful anonymization.

The system demonstrated exceptional reliability and performance throughout the deployment period. System uptime reached 99.2%, with only 8 hours of downtime over the entire six-month period, primarily due to scheduled maintenance and one network infrastructure failure unrelated to our system. Detection accuracy, validated through manual annotation of randomly sampled frames by security personnel, averaged 91.3% across all locations and conditions, closely matching our laboratory benchmark results and confirming the system's robustness in real-world scenarios. The false alarm rate was measured at 2.7%, representing a dramatic improvement over the 12.3% false alarm rate of the previously deployed traditional motion-detection system.

Response time from initial detection to security alert averaged 2.4 seconds, well within acceptable thresholds for security applications. The system processed a total of 8.7 million object detections across all locations, identifying 1,247 critical events including intrusions, abandoned objects, and crowd anomalies that warranted security response. User feedback from security personnel was overwhelmingly positive, with operators reporting a 73% reduction in monitoring fatigue compared to manual surveillance systems, as the intelligent system handled routine monitoring and only alerted personnel to genuine security concerns. Incident response effectiveness improved by 89%, attributed to faster detection, reduced false alarms, and more accurate event classification. The multi-modal approach [8], [14] proved particularly valuable in reducing false alarms during challenging conditions such as nighttime operations and adverse weather, where single-modality systems frequently generated spurious detections. Security managers noted that the privacy-preserving features including automatic anonymization increased acceptance among facility users while maintaining full security capabilities.

VI. DISCUSSION

VI.1 KEY FINDINGS

Our research demonstrates that multi-modal object detection combined with edge computing enables practical deployment of intelligent surveillance systems with superior performance compared to cloud-centric or single-modal approaches [9], [11]. The adaptive fusion mechanism proves crucial for maintaining robustness across varying environmental conditions, with the attention mechanism dynamically adjusting modality contributions based on context [14], [15]. Model optimization techniques including quantization and knowledge distillation successfully bridge the gap between detection accuracy and computational efficiency [3]. INT8 quantization achieves 56× compression with only 6.3% mAP degradation, while knowledge distillation recovers 2.2% accuracy, resulting in practical edge-deployable models maintaining >84% mAP. The hierarchical edge-cloud architecture effectively balances computational workload, reducing bandwidth consumption by 65% while achieving sub-50ms latency. Privacy-preserving techniques including on-device anonymization address critical concerns regarding surveillance data handling without compromising analytical capabilities.

VI.2 LIMITATIONS AND CHALLENGES

Despite promising results, several limitations warrant discussion and suggest directions for future improvement. Hardware dependency represents a significant practical constraint, as our current implementation requires relatively powerful edge devices, specifically NVIDIA Jetson AGX Xavier units, which carry substantial cost implications for large-scale deployment. Each unit costs approximately $800-1000, making deployments with hundreds of cameras economically challenging for many organizations.

Further optimization targeting lower-tier devices such as Jetson Nano ($99-150) would significantly enhance accessibility and enable broader adoption, though this would require additional model compression and architectural innovations to maintain acceptable performance on more constrained hardware. Modality synchronization presents ongoing technical challenges, particularly with precise temporal alignment across sensors operating at different frame rates. The thermal camera operates at only 8.6 Hz while RGB cameras capture at 30 Hz and LiDAR provides continuous point clouds, creating alignment issues that can affect fusion quality. We address this through interpolation and predictive tracking techniques that estimate thermal frames at intermediate time points, but these approaches introduce minor artifacts and potential latency.

Future work should investigate hardware-level synchronization mechanisms or develop fusion algorithms more robust to temporal misalignment. Performance degradation under extreme adverse weather conditions remains a challenge despite multi-modal fusion improvements. While our system demonstrates superior robustness compared to single-modality approaches, extreme conditions including heavy rain, dense fog, and snow still degrade performance across all sensor modalities. Heavy rain produces the most challenging scenario, where accuracy drops to 78.3% mAP even with LiDAR showing the best resilience among our sensors. RGB cameras suffer from reduced contrast and visibility, thermal imaging is affected by rain cooling surfaces and creating thermal noise, and even LiDAR experiences interference from water droplets.

This represents a fundamental limitation of passive and active sensing under severe weather that likely requires additional sensing modalities or advanced signal processing techniques to address. The computational overhead introduced by the cross-modal attention mechanism, while justified by accuracy improvements, adds 8ms to overall latency. For our current 47ms total latency supporting 21.3 fps operation, this overhead is acceptable. However, for higher-resolution multi-camera scenarios requiring processing of 4K or 8K streams or simultaneous tracking of hundreds of objects, this attention computation could become a bottleneck. Research into more efficient attention mechanisms, possibly leveraging recent advances in linear attention or sparse attention patterns, would help scale the approach to more demanding scenarios.

Finally, dataset limitations constrain our ability to train on diverse multi-modal scenarios. Multi-modal annotated datasets with aligned RGB, thermal, and LiDAR data remain scarce in the research community [12], [13]. While our custom dataset of 50,000 annotated frames addresses this partially for our specific deployment scenarios, the broader research community would benefit significantly from large-scale public benchmarks. The lack of diverse training data limits generalization to novel environments and may require additional data collection or domain adaptation when deploying to new locations with significantly different characteristics from training data.

VI.3 PRIVACY AND ETHICAL CONSIDERATIONS

Intelligent surveillance systems raise important privacy and ethical concerns that must be carefully addressed through both technical and operational measures. Our implementation incorporates multiple mechanisms to protect individual privacy while maintaining security functionality. From a technical perspective, on-device anonymization ensures that personally identifiable information never leaves edge devices in identifiable form. The system performs real-time detection and blurring of faces and license plates before any transmission occurs, with selective preservation of security-relevant regions. This approach fundamentally differs from cloud-based systems where raw video streams are transmitted and anonymization occurs remotely, if at all.

The anonymization is designed to be cryptographically reversible, allowing authorized forensic access when legally justified and properly authorized, while maintaining privacy by default for routine operations. Beyond basic anonymization, we incorporate differential privacy mechanisms that provide mathematical guarantees about privacy protection. By adding calibrated noise to detection outputs and carefully managing the privacy budget across different processing tiers, we guarantee ε-differential privacy with $\varepsilon = 1.5$, meaning that the presence or absence of any individual in the dataset has minimal impact on system outputs. This provides formal privacy assurances that complement the operational privacy protections. Encrypted communication channels protect all data in transit between edge devices, edge servers, and cloud infrastructure, while secure storage with access controls ensures that archived data remains protected. All system interactions are logged in immutable audit trails, providing accountability and enabling detection of unauthorized access attempts.

From an operational perspective, we implement role-based access control that strictly limits data access to authorized personnel based on their specific job functions, following the principle of least privilege. Automated retention policies delete surveillance data after specified periods appropriate to the application context, with different retention periods for routine monitoring data versus flagged security events. In appropriate contexts such as residential areas or employee-only zones, the system can incorporate opt-out mechanisms that allow individuals to request exclusion from monitoring or additional privacy protections. Throughout the design process, we have considered regulatory compliance requirements including GDPR in Europe, CCPA in California, and other jurisdiction-specific privacy regulations.

However, we emphasize that technical measures alone cannot fully address the ethical dimensions of surveillance technology. System deployment requires careful assessment of jurisdiction-specific legal requirements, meaningful stakeholder consultation including affected communities, transparent communication about surveillance capabilities and limitations, and robust governance frameworks that prevent misuse. Organizations deploying surveillance systems must balance legitimate security needs against privacy rights, ensure that surveillance is proportionate to actual threats, and maintain human oversight of automated decisions that affect individuals. The technology we present provides tools for privacy-preserving surveillance, but responsible deployment demands ongoing ethical reflection and adaptation to evolving societal norms and expectations regarding privacy and security.

VI.4 FUTURE RESEARCH DIRECTIONS

Several promising research directions emerge from our work that could further advance intelligent surveillance systems. Advanced fusion architectures based on transformer models show significant promise for multi-modal integration with superior long-range dependency modeling capabilities [15]. Unlike the convolutional architectures we currently employ, transformers can more effectively capture global context and relationships between distant spatial regions and across temporal sequences.

Integration of self-attention and cross-attention mechanisms across both spatial and temporal dimensions could further enhance detection capabilities, particularly for tracking objects across multiple camera views and understanding complex behavioral patterns over extended time periods. Current three-dimensional point cloud processing from LiDAR sensors remains computationally intensive, limiting real-time performance and edge deployability. Specialized neural architectures targeting sparse 3D convolutions, which exploit the inherent sparsity of LiDAR point clouds, could dramatically reduce computational requirements. Efficient sampling strategies that intelligently select the most informative points rather than processing complete point clouds could further improve efficiency, enabling broader LiDAR integration in resource-constrained edge systems.

Extending our privacy-preserving edge computing approach to federated multi-task learning represents an exciting direction that could enable more sophisticated analytical capabilities while maintaining strong privacy guarantees [9]. Rather than focusing solely on object detection, a federated learning framework could simultaneously address multiple surveillance tasks including object tracking, behavior recognition, anomaly detection, and crowd analysis. By computing model updates locally on private data and only sharing aggregated model parameters, this approach could enable continuous learning and adaptation across a distributed surveillance network without compromising individual privacy.

Incorporating explainable AI mechanisms to provide interpretable explanations for detection decisions would significantly enhance system trustworthiness and facilitate effective human-AI collaboration in security operations. Security personnel need to understand not just what the system detected, but why it made specific decisions, which features were most influential, and how confident the system is in its predictions. Attention visualization and counterfactual explanations could help security personnel develop appropriate trust calibration and quickly verify system decisions during critical incidents.

Developing techniques for rapid cross-domain adaptation to new deployment environments without extensive retraining would significantly improve system flexibility and reduce deployment costs [10], [13]. Few-shot learning approaches that can adapt to new environments with minimal labeled examples, domain adaptation techniques that transfer knowledge from source to target domains, or meta-learning frameworks could all contribute to more flexible surveillance systems deployable across diverse locations with different architectural layouts, lighting conditions, and typical activity patterns.

Finally, while our current research emphasizes detection accuracy and inference latency [3], energy efficiency will become increasingly critical as surveillance systems expand to include battery-powered or solar-powered edge devices in remote locations. Dynamic voltage-frequency scaling could adjust processing power based on current computational demands, selective sensor activation based on context could power down underutilized sensors, and hierarchical processing with wake-up triggers could keep processing units in low-power states until activity is detected. Such energy optimization techniques could extend operational duration of autonomous surveillance systems and reduce infrastructure requirements for power distribution.

VII. CONCLUSION

This research presents a comprehensive intelligent security surveillance system integrating multi-modal object detection with edge computing paradigms. Our Adaptive Multi-Modal Fusion Network (AMFN) demonstrates that combining RGB, thermal, and LiDAR sensing with attention-based fusion mechanisms achieves superior detection accuracy (89.7% mAP) while maintaining real-time performance suitable for edge deployment (47ms latency). The hierarchical edge-cloud architecture effectively distributes computational workload, reducing bandwidth consumption by 65% compared to cloud-centric approaches while enhancing privacy through on-device processing. Model optimization techniques including INT8 quantization and knowledge distillation enable deployment on resource-constrained edge devices without significant accuracy degradation. Extensive evaluation across benchmark datasets and real-world deployment validation demonstrates system robustness and practical viability.

Six-month field testing across three urban environments confirms reliable operation with 99.2% uptime and consistently high detection accuracy (91.3%). Beyond technical contributions, this research addresses critical privacy and ethical considerations through privacy-preserving design principles. On-device anonymization, differential privacy guarantees, and selective transmission protocols demonstrate that intelligent surveillance can balance security objectives with privacy protection. The proposed system advances the state-of-the-art in intelligent surveillance by demonstrating that multi-modal edge-based detection is not only technically feasible but offers substantial advantages over existing approaches. As surveillance infrastructure continues expanding, edge-based intelligent systems provide a scalable, efficient, and privacy-aware path forward.

VIII. AUTHOR'S CONTRIBUTION

Conceptualization: Shraddha More, Vivian Brian Lobo, Sheetal Patil, Yogita Mane, Vishakha Shelke, Navin Chaganti.
Methodology: Shraddha More, Vivian Brian Lobo, Sheetal Patil, Yogita Mane, Vishakha Shelke, Navin Chaganti.
Investigation: Shraddha More, Vivian Brian Lobo, Sheetal Patil, Yogita Mane, Vishakha Shelke, Navin Chaganti.
Discussion of results: Shraddha More, Vivian Brian Lobo, Sheetal Patil, Yogita Mane, Vishakha Shelke, Navin Chaganti.
Writing – Original Draft: Shraddha More, Vivian Brian Lobo, Sheetal Patil, Yogita Mane, Vishakha Shelke, Navin Chaganti.
Writing – Review and Editing: Shraddha More, Vivian Brian Lobo, Sheetal Patil, Yogita Mane, Vishakha Shelke, Navin Chaganti.
Resources: Shraddha More, Vivian Brian Lobo, Sheetal Patil, Yogita Mane, Vishakha Shelke, Navin Chaganti.
Supervision: Shraddha More, Vivian Brian Lobo, Sheetal Patil, Yogita Mane, Vishakha Shelke, Navin Chaganti.
Approval of the final text: Shraddha More, Vivian Brian Lobo, Sheetal Patil, Yogita Mane, Vishakha Shelke, Navin Chaganti.

IX. REFERENCES

[1] Jocher, G., et al. (2022). YOLOv5: A state-of-the-art real-time object detection system. GitHub repository.

[2] Tan, M., Pang, R., & Le, Q. V. (2020). EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10781-10790).

[3] Liu, J., Zhang, S., Wang, S., & Metaxas, D. N. (2016). Multispectral deep neural networks for pedestrian detection. In British Machine Vision Conference (BMVC).

[4] Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., & Harada, T. (2017). MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

[5] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[6] Zhang, H., Fromont, E., Lefèvre, S., & Avignon, B. (2020). Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In IEEE International Conference on Image Processing (ICIP).

[7] Wagner, J., et al. (2016). Multispectral pedestrian detection using deep fusion convolutional neural networks. In European Symposium on Artificial Neural Networks (ESANN).

[8] Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 1137-1149.

[9] Saponara, S., Elhanashi, A., & Gagliardi, A. (2021). Real-time video fire/smoke detection based on CNN in antifire surveillance systems. Journal of Real-Time Image Processing, 18(3), 889-900.

[10] Zhou, K., Chen, L., & Cao, X. (2020). Improving multispectral pedestrian detection by addressing modality imbalance problems. In European Conference on Computer Vision (ECCV).

[11] Li, C., et al. (2019). Learning to fuse things and stuff. arXiv preprint arXiv:1812.01892.

[12] Hwang, S., Park, J., Kim, N., Choi, Y., & So Kweon, I. (2015). Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1037-1045).

[13] Cao, Y., Zhou, T., Zhu, X., & Su, Y. (2021). Every feature counts: An improved one-stage detector in thermal imagery. In IEEE International Conference on Computer Vision Workshops (ICCVW).

[14] Zhang, L., Liu, Z., Zhang, S., Yang, X., Qiao, H., Huang, K., & Hussain, A. (2021). Cross-modality interactive attention network for multispectral pedestrian detection. Information Fusion, 50, 20-29.

[15] Guan, D., Cao, Y., Yang, J., Cao, Y., & Yang, M. Y. (2019). Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. Information Fusion, 50, 148-157.