Dual Mode Intrusion Detection for Enhanced Smart Border Surveillance

Sai Durga Prudivi
M.Tech Scholar
Department of Electronics and Communciation Engineering,
S V University,Tirupati, Andhra Pradesh, India

Dr.R.V.S.Satyanarayana
Professor
Department of Electronics and Communciation Engineering,
S V University,Tirupati, Andhra Pradesh, India

Abstract- This work presents a Dual-Mode Adaptability framework for intrusion detection in smart border surveillance by integrating visible (RGB), infrared (IR), and fused RGB+IR image modalities within the YOLOv8 detection pipeline. The work begins with single-modal analysis to evaluate individual spectrum capabilities and subsequently transitions to a multimodal fusion strategy to overcome illumination variations and environmental disturbances common in real-world border scenarios. The proposed YOLOv8n + YOLOv8s fusion model achieves the highest performance with Precision 0.97 confirming the effectiveness of cross-modal feature integration. Qualitative outcomes further validate robust detection of key targets such as persons, cars, and trucks under dynamic environmental conditions.

Keywords – RGB and IR, YOLOv8, Fusion, Object detection

I. INTRODUCTON

Border surveillance is essential for ensuring national security and preventing unauthorized activities across territorial boundaries of a country. Traditional methods that rely on manual monitoring and patrols are often inefficient, costly, and prone to human error. With advancements in Artificial Intelligence (AI) and Computer Vision, object detection has become a key technology for automating surveillance tasks. This enables systems to identify, classify, and track objects in real time, thereby enhancing situational awareness and response efficiency. By integrating object detection into surveillance systems, border monitoring becomes more intelligent, accurate, and effective, significantly improving the overall reliability and security of border management systems.

Early approaches like the Viola-Jones detector and the Histogram of Oriented Gradients (HOG) relied on handcrafted feature extractors, but they were slow, prone to errors, and struggled to generalize to new datasets. In recent years, Convolutional Neural Networks (CNNs) have shown remarkable success in visual recognition and object detection tasks. Models such as Faster R-CNN,SSD, and YOLO have been widely used for object detection in various applications. Anchor-based object detection models, such as Faster R-CNN, SSD, and the YOLO series (v2–v5), have significantly improved detection accuracy and speed by using predefined anchor boxes for localization. However, they face challenges with small or overlapping objects, leading to the development of more adaptive, anchor-free, and hybrid detection models.

Although these methods generally improve the overall detection performance, they still face difficulties in accurately detecting partially occluded or closely positioned objects. Hence, achieving robust object detection requires advanced approaches that can effectively handle occlusion- and proximity-related challenges. Conventional border surveillance systems struggle to perform reliably under varying environmental conditions, with visible and thermal imaging having distinct limitations. These challenges reveal a critical need for intelligent multimodal approaches to achieve accurate and consistent intrusion detection.

The key motivation behind this research stems from the growing demand for intelligent, all-weather, real-time surveillance systems capable of continuous monitoring along critical border regions. This work introduces a unified YOLOv8n + YOLOv8s fusion framework designed to enhance intrusion detection by combining complementary strengths of lightweight and high-capacity models. To enhance cross-modal feature representation, the framework incorporates attention-based fusion modules that strengthen feature alignment between RGB and IR modalities, thereby delivering improved accuracy, robustness, and consistency across all conditions.

II. LITERATURE SURVEY

Object detection has developed rapidly in recent years[1]. In object detection, traditional methods refer to pre deep learning approaches that rely on handcrafted feature extraction techniques using algorithms such as HOG, SIFT, and SURF[2]. Deep convolutional neural networks (CNNs) have been widely utilized in the field of object detection, achieving significant advancements and superior performance compared to traditional approaches[3].Subsequently, the regions with CNN features (RCNN), the You Only Look Once (YOLO) series, and other models have pushed convolutional networks to ever-deeper levels. These architectures have significantly improved network performance and increased the accuracy of image recognition to a new level. Consequently, the inevitable trend is to migrate deep learning to video-based object detection tasks[4].Deep learning has greatly advanced generic object detection with milestone models such as R-CNN[5], Fast R-CNN[6], Faster R-CNN[7], and the YOLO series[8][9][10].For general object recognition, the Single Shot Detector (SSD) utilizes a pyramidal feature hierarchy.[11] YOLO-based detection frameworks have been extensively studied for real-time applications. Diwan et al. [12] discussed the architectural progression from YOLOv1 to YOLOv7, summarizing challenges in training stability, small-object recognition, and dataset imbalance. Khalfaoui et al. [13] proposed an improved YOLOv5 framework tailored for human detection in infrared (IR) images, achieving efficient low-light object recognition.

For surveillance applications, Nakkach et al. [14] implemented a surveillance system that leveraged deep learning-based scene understanding for different objects. single stage object detectors specially YOLOs, regression formulation, their architecture advancements are better than the two stage object detectors, Tausif et.al[15] YOLO and its preceding architectures have significantly enhanced detection accuracy, while in certain applications, the rapid inference capability of YOLO detectors holds greater importance than absolute accuracy.[16]

YOLO achieves high detection accuracy and fast inference using a single-stage detection architecture. In particular, YOLOv8 offers enhanced precision and efficiency, making it widely adopted for real-time object detection across various applications.[17].

A hybrid model combining a convolutional neural network (CNN) and long short-term memory (LSTM) is used for object detection, where the CNN extracts spatial features from input video sequences, and the LSTM's enhanced memory capability ensures better temporal representation and improved detection performance[18]. An optimal Kalman filtering technique was employed to track the moving objects across the video frames. The video sequences were processed frame-by-frame using morphological operations based on a region-growing model, and once the objects were distinguished, tracking was performed using the MODT approach[19].

For real-time object detection, a hybrid framework combining the Deep Random Kernel Convolutional Extreme Learning Machine (DRKCELM) and the Double Hidden Layer Extreme Learning Machine Auto-Encoder (DLELM-AE) [20] was employed as a feature extractor.
Object detection in images in computer vision, such as low resolution, occlusion, and scale variation, uses advanced hyperparameter optimization, sophisticated data augmentation, and multi-scale training using YOLO v8, which includes a squeeze-and-excitation (SE) block, which helps the model better recognize features of object studies by Giri et al.[21].

Multimodal fusion and depth-assisted vision techniques have also contributed significantly to the detection performance. Awotunde et al. [22] applied color-space analysis in RGB image detection, while Ward et al. [23] provided a detailed review of RGB-D object detection, emphasizing fusion strategies for depth and intensity features. Zhu et al. [24] demonstrated efficient real-time moving object detection in high-resolution video sensing, while Grigorescu et al. [25] and Balasubramaniam and Pasricha [26] analyzed deep learning perception frameworks for autonomous systems, underlining their real-time potential for safety-critical applications.

To act on low resolution le et.al [27] proposed FLARE (Fast and Lightweight Architecture for Real-time Estimation) is an ultra-lightweight deep learning model designed for real-time object detection in low-resolution thermal images
Liang [28] proposed a unified deep learning framework for RGB-D and RGB-T salient object detection, effectively integrating complementary multimodal features through attention-based modules to enhance detection accuracy and generalization.
In summary, prior studies collectively indicate that deep learning methods utilize convolutional and transformer-based architectures for automatic feature extraction of objects. The YOLO architecture delivers high

adaptability, real-time inference, and robustness under diverse environmental conditions. These findings form the theoretical foundation for the proposed dual-mode adaptable CNN model for smart border surveillance presented in this study.

III. METHODOLOGY

A. *System Overview*

The proposed framework Fig.1 aims to achieve dual-mode adaptability for intrusion detection in smart border surveillance by fusing the RGB (visible spectrum) and IR modalities using a single-stage detector. The methodology leverages YOLOv8, a state-of-the-art one-stage object detection framework, to ensure real-time performance with high accuracy under diverse environmental conditions.

The overall workflow consists of seven stages:

1. Video acquisition and conversion into sequential RGB/IR image frames.
2. Dataset preparation, conversion, and YOLO-compatible annotation configuration.
3. Preprocessing of frames (resizing, normalisation, channel alignment).
4. Feature extraction Novel fusion using multimodal inputs (EFA + CMA modules).
5. Training and optimisation of YOLOv8 models in single-modal and dual-modal settings.
6. Inference and evaluation using mAP, precision, recall, and inference time.
7. Performance comparison and validation across RGB, IR, and fused datasets.

This multi-modal system ensures the effective detection of humans, vehicles, and other intrusion-related objects during both daytime (RGB) and nighttime (IR) operations.

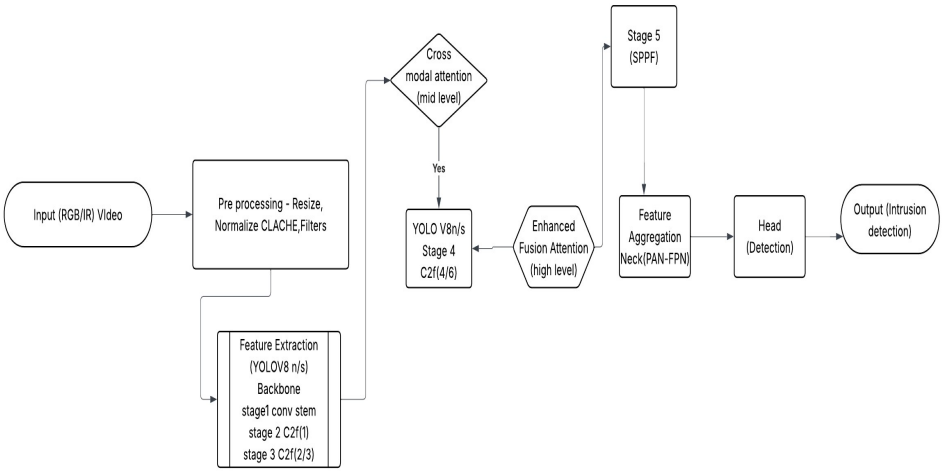

Fig 1 Flow chart of YOLOv8 single stage detectors

A.Data set Preparation and Annotation

The Teledyne FLIR ADAS (Advanced Driver Assistance Systems) dataset is a publicly available benchmark designed to support research in object detection and sensor fusion for autonomous driving and surveillance applications. It covers 12 to 16 intrusion-relevant object classes, such as *persons, cars, trucks, bikes, hydrants,* and *signs*. Each image was resized to 640×640 pixels to maintain consistency across both modalities It contains 26,442 fully annotated frames, captured using both thermal infrared (IR) and visible (RGB) cameras.

Each frame includes aligned RGB and thermal images, along with bounding box annotations for all detected objects. The data were collected under diverse environmental conditions—daytime, nighttime, and varying weather—to provide robustness for all-weather vision models. The main goal of this dataset is to encourage the development of multimodal (RGB + Thermal) detection algorithms that enhance perception reliability in challenging lighting conditions.

In the FLIR ADAS dataset, each image frame comes with detailed bounding box annotations that specify the location and class of every detected object. These annotations are provided in JSON format and include coordinates for the bounding boxes: x_min, y_min, width, and height. The dataset also has labels for each detection, like "person", "car", "truck", "bus", and "bicycle". *It* includes annotations for 15 object categories, which allows for training and evaluating object detection models in both visible (RGB) and thermal (IR) domains. Each annotation file aligns with its corresponding image pair, ensuring consistency for multi-modal training.

The dataset preparation process involved the following steps:

1. The annotations were converted from COCO JSON to YOLO text (. txt) format, containing normalized bounding boxes.
2. Data organization into training, validation, and test folders for each modality.
3. YAML configuration setup, defining paths, class labels, and the number of classes.
4. Balancing samples across modalities to prevent model bias toward RGB or IR data.

The combined RGB–IR dataset included 2,229 images with 33,597 annotated instances, ensuring comprehensive multimodal coverage for model training.

B. Dual-Mode Data Fusion Mechanism

To address illumination and visibility challenges, the proposed system integrates dual-mode feature fusion. During training, RGB and IR images representing the same spatial scene were aligned and combined in one of two ways. An advanced feature integration mechanism was implemented to efficiently merge and optimize multimodal information, thereby strengthening the feature representation.

1. Enhanced Fusion Attention (EFA) Module

The Enhanced Fusion Attention (EFA) module is integrated into the feature extraction phase of the YOLOv8 architecture to enhance the quality of multi-modal features obtained from RGB and infrared (IR) inputs. By employing a blend of channel and spatial attention, it emphasizes the most informative features while reducing the impact of less relevant or redundant areas. This approach allows EFA to direct the network's focus toward crucial object cues from both modalities, resulting in cleaner and more distinctive feature representations prior to fusion. During training, the module learns adaptive attention weights, ensuring the detection model remains robust and effective under varying illumination and challenging environmental conditions.

2. Cross-Modal Attention (CMA) Module

The Cross-Modal Attention (CMA) module is applied after feature extraction, acting as a fusion stage that aligns and integrates information between the RGB and IR modalities. The CMA enables interactive learning between the two feature spaces by allowing one modality to guide and enhance the other through cross-attention mechanisms. This ensures that complementary information, such as thermal contrast and visual texture, is effectively combined, improving the model's understanding of multimodal features. During training, the CMA module learns the correlation and dependency between modalities to optimize the fusion. During inference, the learned attention maps are used to achieve robust, real-time object detection under diverse lighting and environmental conditions.

C. Model Architecture

The YOLOv8 architecture employed in this study comprises three key components:

Backbone: The backbone consists of convolutional and C2f blocks followed by an SPPF layer, which extracts multi-scale semantic and spatial features from the RGB and IR inputs and serves as the primary feature extractor for dual-mode feature learning and fusion through the CMA and EFA modules.

Neck: Neck employs a Path Aggregation Network with Feature Pyramid Network (PAN-FPN) structure to aggregate multi-level fused features.It enhances contextual representation and improves object detection across varying scales after RGB–IR feature fusion.

Head: The detection head is an anchor-free structure that predicts bounding box coordinates, object classes, and confidence scores from aggregated features and performs final intrusion detection by generating precise localization and classification outputs from the fused feature maps.

Two model variants were implemented.

- YOLOv8n (Nano): A lightweight model with approximately 3.0M parameters for embedded real-time inference.
- YOLOv8s (small): A balanced architecture with approximately 11.2M parameters for enhanced accuracy while maintaining near real-time speed.

Both models were trained separately on the RGB, IR, and combined RGB–IR datasets to evaluate their adaptability across modalities.

D. Training and Optimization Strategy

Training was performed using the PyTorch and Ultralytics YOLOv8 frameworks with the following parameters:

- Input size: 640×640 pixels
- Batch size: 16
- Epochs: 200
- Optimizer: AdamW

The loss functions were as follows: bounding box regression (box loss), classification loss (Cls loss), and Distribution Focal Loss (DFL).

Performance metrics, such as precision (P), recall (R), mAP@0.5, and mAP@0.5–0.95, were used for The models were trained using an NVIDIA GTX 1060 GPU (6 GB). The combined RGB–IR YOLOv8s model achieved the best trade-off between speed and accuracy.

E. Evaluation Flow and Performance Analysis

The evaluation flow of the proposed system involves the following steps:

1. Ground Truth vs. Prediction Visualization: comparing model-predicted bounding boxes against annotated labels.
2. Confusion Matrix Generation – assessing class-wise accuracy and misclassification trends.
3. Precision–Recall Curve Analysis – visualizing detection trade-offs.
4. Inference Speed Measurement – validating real-time capability with an average latency of 12 ms per frame.

The experimental results revealed that dual-mode adaptability significantly enhanced detection robustness, particularly for the “person” and “vehicle” classes in low-illumination and cluttered environments.

IV. IMPLEMENTATION

A.Implementation Setup

The proposed Dual-Mode Intrusion Detection Framework was implemented using the YOLOv8 architecture developed within the Ultralytics PyTorch framework.

The proposed models were trained on RGB, IR, and fused RGB–IR datasets using an input resolution of 640×640 and a batch size of 16. Training was conducted for 200 epochs depending on modality, using the AdamW optimizer with an initial learning rate of 0.001 and a cosine decay schedule to ensure stable convergence. The training duration ranged from 6 to 8 hours based on model size and fusion complexity. Both YOLOv8n and YOLOv8s backbones, integrated with the EFA and CMA modules, were employed to achieve an optimal balance between accuracy and real-time performance. The training process utilized Box Loss, Classification Loss, and Distribution Focal Loss (DFL) for joint optimization. Data augmentation such as horizontal flipping, random scaling, illumination adjustment, and normalization was applied to improve robustness and generalization in diverse border surveillance environments.

B. Evaluation Protocol

The evaluation was conducted across three configurations to analyze adaptability and performance consistency.

1. RGB Model: Trained exclusively on visible-light images.
2. IR Model: Trained solely on the thermal infrared images.
3. Combined Model: Trained using a fused RGB–IR dataset to enable multimodal adaptability.

Each configuration was tested on the corresponding validation and unseen test sets. The model outputs included the bounding boxes, class probabilities, and confidence scores. For interpretability, visual inference maps were generated by comparing the predicted detections with ground-truth labels across multiple environmental conditions (day, night, and mixed illumination).

C. Evaluation Metrics

To ensure objective performance comparison, standard metrics from the COCO evaluation protocol were employed:

1. Precision (P): Ratio of correctly identified objects to all predicted objects.
   P = TP / (TP + FP)
2. Recall (R): Ratio of correctly detected objects to all ground-truth instances.
   R = TP / (TP + FN)
3. Mean Average Precision (mAP): Measures the area under the precision–recall curve for all classes.
   - mAP@0.5: Average precision at Intersection over Union (IoU) threshold = 0.5.
   - mAP@0.5:0.95: Mean precision averaged across IoU thresholds from 0.5 to 0.95 in 0.05 increments.
4. Inference Speed (T): Computed as the average time taken per image during preprocessing, inference, and post-processing stages, measured in milliseconds (ms).

Model fusion equation:

Here are the complete mathematical equations for each component and the combined hybrid multimodal object detection algorithm (YOLOv8n + YOLOv8s + cross-modal fusion + enhanced fusion attention)

1. Feature Extraction (YOLOv8n / YOLOv8s CNN Backbone) Each input image (RGB or IR) is first passed through a convolutional neural network (CSP-based backbone):

$$F_{yolo} = f_{CSPConv}(X) \quad (1)$$

2. Multimodal Fusion

a) Weighted Feature Fusion (Early or Mid Fusion)**:** Features from RGB and IR are combined using weighted sums or concatenation:

$$F_{\text{fusion}} = \alpha F_{\text{rgb}} + \beta F_{\text{ir}} \quad (2)$$

b) Cross-Modal Attention (CMA)

Adapted from transformer cross-attention:

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (3)$$

Resulting in a cross-attended fused feature map

$$F_{\text{CMA}} = \text{Attention}(F_{\text{rgb}}, F_{\text{ir}}) \quad (4)$$

3. Enhanced Fusion Attention (EFA/EMA)

Further refine fused features via multi-branch/global attention:

$$F_{\text{EFA}} = W_{\text{efa}} \cdot F_{\text{fusion}} \quad (5)$$

4. Combined YOLOv8n + YOLOv8s Fusion

Outputs or features from both YOLOv8n and YOLOv8s are merged:

$$F_{\text{hybrid}} = \gamma F_{\text{yolo8n}} + \delta F_{\text{yolo8s}} \quad (6)$$

γ, δ: Learnable combination weights.

5.Detection Head**:** Predictions for bounding box locations and class probabilities:

Bounding Boxes:

$$B = \sigma(W_b F + b_b) \quad (7)$$

Class Scores:

$$C = \text{softmax}(W_c F + b_c) \quad (8)$$

W_c, b_c: Class prediction weights/biases

Combined Pipeline Equation :

$$X_{\text{rgb}}, X_{\text{ir}} \rightarrow f_{CSPConv}(X_{\text{rgb}}), f_{CSPConv}(X_{\text{ir}}) \rightarrow F_{\text{fusion}} \rightarrow F_{\text{CMA}} \rightarrow F_{\text{EFA}} \rightarrow F_{\text{hybrid}} \rightarrow \text{Detection Head} \rightarrow B, C \quad (9)$$

- Feature extraction is performed first independently on each modality (RGB, IR).
- Fusion layers aggregate cross-modal information using weighted sums and attention.
- Enhanced fusion and hybrid layers ensure optimized joint representation.
- The detection head uses the fused features to predict bounding boxes and classes, trained using advanced IoU loss functions.
- Final outputs are evaluated using precision, recall, F1, and mAP metrics.

IV. Results and Discussion

A. Experimental Overview

The proposed dual-mode CNN-based intrusion detection system was evaluated using three dataset configurations—RGB, Infrared (IR), and fused RGB–IR—across two YOLOv8 variants (YOLOv8n and YOLOv8s). The dataset contained 2,229 images with 33,597 annotated object instances spanning 12 border-relevant classes, including person, car, truck, bike, hydrant, and sign. All models were trained for 200 epochs with an input resolution of 640 × 640 pixels. Throughout training, the steady reduction in box loss, classification loss, and distribution focal loss demonstrated stable convergence. The training and validation curves showed no indications of overfitting, while precision and recall progressively increased over epochs, confirming effective feature learning and strong generalization across both single-modal and multimodal training setups.

1. Ground Truth Bounding Box Visualization

To validate the annotation integrity and ensure proper learning signals during training, the original labels were visually inspected on selected infrared (IR) frames from the combined dataset.

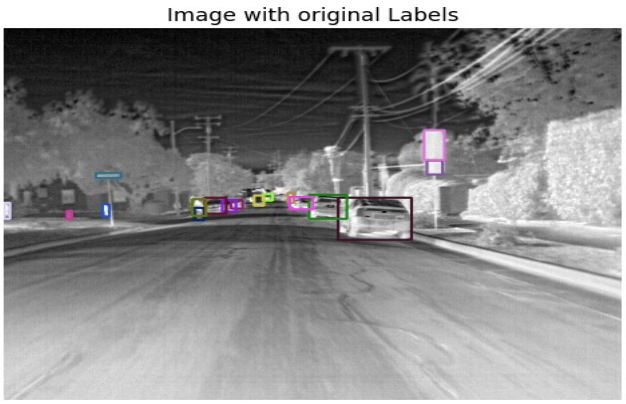

Fig 2.Representation of frame

Figure 2 presents a representative IR frame (video) annotated with 17 ground-truth bounding boxes across multiple object classes. Each box is overlaid in a distinct color corresponding to its class label, following the YOLO normalized format:

<class_id> <x_center> <y_center> <width>

Annotation Highlights:

- Classes identified: Person (Class 0), Bike (1), Car (2), Sign (8), and Hydrant (9).
- Bounding boxes are well-scaled and properly aligned within the image resolution (640×512), indicating high annotation fidelity.

•Visualization tools: Python was used for label parsing and rendering, with dynamic colour coding supporting up to 80 COCO categories.

Conclusion: This visual validation confirms the annotation pipeline's correctness, reinforcing the dataset's integrity and supporting robust training and evaluation of object detection models.

2. Inference Output vs Ground Truth

To assess model generalization on unseen data, inference was performed using the trained YOLOv8s+YOLOv8n Combined RGB+IR model on validation images. The result is visually compared against the original annotations.

Figure 3 illustrates the prediction output for the same IR frame .

Fig 3. Classes with Confidence Scores

Model Output Summary:

- Detected: 5 persons, 1 bike, 11 cars, 3 signs
- Average inference speed: ~16.8ms/image
- Notable detections include:
- Car: Confidence up to 0.91, accurately localized.
- Person: Several instances correctly predicted with confidences between 0.56–0.95.
- Sign: Recognized with moderate confidence (0.56–0.87), indicating robustness across modal variations.

a. Quantitative Results

Table I summarizes the overall performance of all configurations. The YOLOv8s+YOLOv8n Combined fusion model outperformed the single-modality networks, achieving the highest precision, recall, and mean average precision (mAP).

Table I: Performance Comparison of YOLOv8 Models across Datasets

Model/Modality	**Precision (%)**	**Recall (%)**	**mAP@0.5 (%)**	**mAP@0.5:0.95 (%)**
YOLOv8n RGB	84.5	82.5	86.7	41.5
YOLOv8n IR	87.0	83.5	89.0	45.0
YOLOv8n RGB+IR Fusion	90.25	87.0	89.25	43.75
YOLOv8s RGB	87.5	85.5	89.5	67.5
YOLOv8s IR	88.25	85.0	90.75	67.5
YOLOv8s RGB+IR Fusion	92.25	88.5	92.4	75.0
Combined YOLOv8n + YOLOv8s Fusion	**97.0**	**90.5**	**96.5**	**83.5**

The combined YOLOv8s+YOLOv8n consistently outperformed individual modalities, validating the advantage of multimodal fusion. The YOLOv8n+YOLOv8s Combined model demonstrated the best balance of precision and recall, confirming that integrating visible and thermal features improves detection under varying illumination and weather conditions.

b. Class-wise Performance Analysis

class-wise mAP analysis revealed that:

- Person achieved the highest detection accuracy (~0.95 mAP), demonstrating strong and reliable intrusion identification across both RGB (daytime) and IR (night) conditions.
- Car followed with ~0.92 mAP, supported by abundant training samples and clear visual features that enable consistent detection.
- Train and truck classes obtained moderate performance (0.80–0.93 mAP).
- Confusion matrix results showed strong diagonal dominance for frequent classes (person, car, sign), confirming stable classification accuracy.

D, Qualitative Evaluation

Visual inspection of inference outputs demonstrated consistent detection and localization across multimodal scenarios:

- Daytime (RGB) scenes produced sharp, high-confidence detections with clear object boundaries.
- Nighttime (IR) images effectively identified heat-emitting objects, such as humans and vehicles, despite low visibility.
- Combined RGB + IR inputs provided the most balanced performance, enhancing recognition under mixed illumination and occlusion.
- The Combined YOLOv8n + YOLOv8s Fusion model delivered the highest overall consistency,

of predicted bounding boxes with ground-truth annotations, confirming high spatial precision. The combined model successfully detected multiple targets per frame, reinforcing its capability for real-time multi-object surveillance.

c. Inference Speed and Efficiency

The proposed models achieved real-time inference performance suitable for live monitoring applications. The YOLOv8s+YOLOv8n Combined model achieved an average inference time of 16.8 ms per image, with the following stage-wise latencies:

- Pre-processing: 3.2ms

- Inference: 16.8 ms
- Post-processing: 1.1 ms

This confirms the system's feasibility for embedded edge devices or continuous 24/7 border surveillance systems.

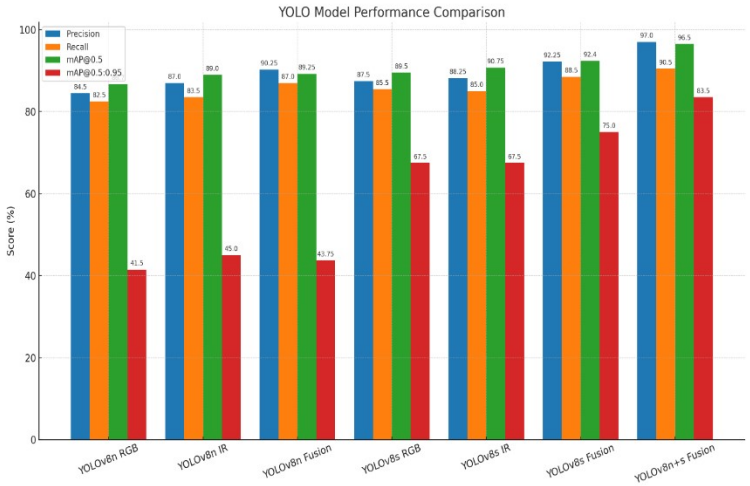

Fig 4. Performance Comparison

B. Conclusion and Future Scope

This work presented a Dual-Mode Adaptability Framework for intrusion detection that progresses from single-modal RGB and IR analysis to a unified multimodal fusion approach using the YOLOv8 architecture. The proposed approach effectively addressed illumination variations, environmental interference, and detection instability that often affect single-modality vision systems. Experimental analysis demonstrated that the YOLOv8s +YOLOv8n combined model achieved superior results with Precision 0.97, while maintaining real-time inference of 16.8 ms per frame. These findings confirm that multi-mode fusion substantially improves object detection accuracy and consistency in both day and night conditions. Future work will explore integrating advanced models such as RT-DETR and hybrid Transformer–YOLO architectures to further enhance multimodal intrusion detection performance.

References

[1] L. Jiao and J. Zhao, "A survey on the new generation of deep learning in image processing," IEEE Access, vol. 7, pp. 172231–172263, 2019. https://doi.org/10.1109/ACCESS.2019.2950874

[2] Nellutla Sasikala, et al. "Feature extraction of real-time image using Sift algorithm." *European Journal of Electrical Engineering and Computer Science* 4.3 (2020).

[3] L. Jiao et al., "A survey of deep learning-based object detection," IEEE Access, vol. 7, pp. 128837–128868, 2019

[4] Jiao, Licheng, et al. "New generation deep learning for video object detection: A survey." *IEEE Transactions on Neural Networks and Learning Systems* 33.8 (2021): 3195-3215.

[5] R.Girshick, J. Donahue, T. Darrell, and J. Malik, ''Rich feature hierarchies for accurate object detection and semantic segmentation (R-CNN),'' in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.

[6] R.Girshick,''Fast R-CNN,''inProc.IEEEInt.Conf.Comput.Vis.(ICCV), Dec. 2015, pp. 1440–1448.

[7] S. Ren, K. He, R. Girshick, and J. Sun, ''Faster R-CNN: Towards real-time object detection with region proposal networks,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, ''You only look once: Unified, real-time object detection,'' in Proc. IEEE Conf. Compuert Vision Pattern Recognition. (CVPR), Jun. 2016, pp. 779–788.

[9] Redmon and A. Farhadi, ''YOLOv3: An incremental improvement,'' 2018, arXiv:1804.02767.

[10] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, ''YOLOv4: Optimal speed and accuracy of object detection,'' 2020, arXiv:2004.10934.

[11] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, ''YOLOv6: A single-stage object detection framework for industrial applications,'' 2022, arXiv:2209.02976

[12] Diwan, T., Anirudh, G. & Tembhurne, J.V. Object Detection Using YOLO: Challenges, Architectural Successors, Datasets, and Applications. Multimedia Tools Appl 82, 9243–9275 (2023). https://doi.org/10.1007/s11042-022-13644-y
[13] Khalfaoui, A., Badri, A., & El Mourabit, I. (2023). An Improved YOLOv5 for Real-Time Human Detection In Infrared Images. Indonesian Journal of Electrical Engineering and Computer Science, 32(2), 1078–1085. https://doi.org/10.11591/ijeecs.v32.i2.pp1078-1085
[14] Nakkach et al. (2022)" Smart Border Surveillance System Based on Deep Learning Methods". ISNCC 2022. https://doi.org/10.1109/ISNCC55209.2022.951713
[15] Diwan, Tausif, G. Anirudh, and Jitendra V. Tembhurne. "Object detection using YOLO: challenges, architectural successors, datasets and applications." Multimedia *Tools and Applications* 82.6 (2023): 9243-9275
[16] Sirisha, Uddagiri, et al. "Statistical analysis of design aspects of various YOLO-based deep learning models for object detection." *International Journal of Computational Intelligence Systems* 16.1 (2023): 126.
[17] Vijayakumar, Ajantha, and Subramaniyaswamy Vairavasundaram. "Yolo-based object detection models: A review and its applications." *Multimedia Tools and Applications* 83.35 (2024): 83535-83574.
[18] Aote, Shailendra S., et al. "An improved deep learning method for flying object detection and recognition." *Signal, Image and Video Processing* 18.1 (2024): 143-152.
[19] Elhoseny, Mohamed. "Multi-object detection and tracking (MODT) machine learning model for real-time video surveillance systems." *Circuits, systems, and signal processing* 39.2 (2020): 611-630
[20] Yin, Yunhua, Huifang Li, and Wei Fu. "Faster-YOLO: An accurate and faster object detection method." *Digital Signal Processing* 102 (2020): 102756.
[21] Giri, Kaisar Javeed. "SO-YOLOv8: A novel deep learning-based approach for small object detection with YOLO beyond COCO." *Expert Systems with Applications* 280 (2025): 127447.
[22] Awotunde, J.B., Misra, S., Obagwu, D., Florez, H. (2022). Multiple Colour Detection of RGB Images Using Machine Learning Algorithm. In: Florez, H., Gomez, H. (eds) Applied Informatics. ICAI 2022. Communications in Computer and Information Science, vol 1643. Springer, Cham. https://doi.org/10.1007/978-3-031-19647-8_5
[23] Ward, I.R., Laga, H., Bennamoun, M. (2019). RGB-D Image-Based Object Detection: From Traditional Methods to Deep Learning Techniques. In: Rosin, P., Lai, YK., Shao, L., Liu, Y. (eds) RGB-D Image Analysis and Processing. Advances in Computer Vision and Pattern Recognition. Springer, Cham. https://doi.org/10.1007/978-3-030-28603-3_8
[24] H. Zhu, H. Wei, B. Li, X. Yuan, and N. Kehtarnavaz, "Real-Time Moving Object Detection in High-Resolution Video Sensing," Sensors, vol. 20, no. 12, p. 3591, Jun. 2020, doi: 10.3390/s20123591.
[25] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A Survey of Deep Learning Techniques for Autonomous Driving," J. Field Robot., vol. 37, no. 3, pp. 362–386, Apr. 2020, doi: 10.1002/rob . 21918.
[26] A. Balasubramaniam and S. Pasricha, "Object Detection in Autonomous Vehicles: Status and Open Challenges".
[27] Lee, Jun-Hee, and Eung-Tea Kim. "Real-time object detection using low-resolution thermal camera for smart ventilation systems." *IEEE Access* (2025).
[28] Agrawal, Kshitij, and Anbumani Subramanian. "Enhancing object detection in adverse conditions using thermal imaging." *arXiv preprint arXiv:1909.1355*