Multispectral Pedestrian Detection in Low-Light Conditions: Infrared, Visible, and Fusion-Based Approaches for CCTV Applications

Ylli Rexhaj[1] ,Redon Rexhepi[2] , Arxhend Jetullahu[3] ,Ideal Rafuna[4]

UBT College, Prishtina, Rep. of Kosovo

ylli.rexhaj@ubt-uni.net
redon.rexhepi@ubt-uni.net
arxhend.jetullahu@ubt-uni.net
ideal.rafuna@ubt-uni.net

Abstract. Reliable pedestrian detection under low-light conditions remains a big challenge for intelligent surveillance and autonomous monitoring systems. Visible (VIS) or RGB sensor fail in darkness due to limited illuminations, while in the other hand infrared (IR) camera lack fine visual texture. This paper presents an adaptive multispectral pedestrian detection framework to address these limitations that combines the strengths of VIS and IR modalities through Weighted Boxes Fusion (WBF) applied to separately trained YOLOv8 models. Three variants of YOLOv8 the nano, small and medium models were trained independently on the LLVIP dataset. During inference, their outputs were merged using WBF with adaptive modality weighting determined by the mean brightness of each image. The models and fusions were evaluated using Precision, Recall, mAP@0.50 and mAP@0.50:0.95 as performance metrics.

Results demonstrate that the adaptive fusion outperforms both single-modality detectors achieving up to 0.97 precision and 0.914 recall. Also, the correlation between image brightness and detection confidence is negligibly negative ($-0.03 \leq r \leq -0.06$), confirming that the adaptive weighting successfully neutralizes the influence of illumination on detection reliability. Providing a lightweight and scalable solution for real-time CCTV and smart-city surveillance application these findings validate the effectiveness of brightness-guided late fusion in achieving illumination-invariant multispectral pedestrian detection.

Keywords: Multispectral pedestrian detection, Adaptive fusion, YOLOv8, Weighted Boxes Fusion, Low-light surveillance.

1 Introduction

The ability to accurately detect human presence across diverse environmental conditions enables reliable decision-making and enhances situational awareness in modern smart-city infrastructures, which is a critical component in intelligent surveillance, autonomous driving, and public safety systems. Most conventional vision-based pedestrian detection systems often rely solely on visible (VIS) spectrum imagery, which is highly sensitive to illumination. Under low-light or nighttime, RGB cameras suffer from loss of texture, motion blur, and noise, which results in a significant drop in detection accuracy. To mitigate these limitations, infrared (IR) sensors have been introduced as complementary sources of information. Infrared cameras capture thermal radiation emitted by objects, allowing pedestrian detection even in total darkness or adverse weather conditions. Nonetheless, IR imagery alone lacks fine visual detail and is sensitive to temperature variations and background heat sources. That's when these complementary properties suggest that multispectral (VIS+IR) fusion can improve detection robustness across variable lightning environments. Fusions can be categorized in 3 categories which the first one is the early fusion, which combines VIS and IR data at pixel level before feature extraction, while mid-level fusion merges features within a shared neural network architecture. Although these methods integrate information tightly, they require careful alignment, large annotated datasets, and significant computational resources. In the other hand, late fusion approaches combine the final predictions (bounding boxes) from independent detector, offering a lightweight, flexible, and hardware-independent solution. This study uses YOLOv8 (You Only Look Once) as the base detector for both VIS and IR modalities which has achieved state-of-the-art performance in real-time object detection. Also, in this study introduces a brightness-guided adaptive late fusion framework that dynamically adjusts the contribution of each modality according to scene illumination.

Firstly we train YOLOv8 models the nano, small and medium in visible and infrared subsets of the LLVIP dataset, which provides paired multispectral pedestrian images. During inference prediction from VIS and IR models then are merged using Weighted Boxes Fusion (WBF), which is a method that refined final bounding boxes by averaging overlapping detections weighted by their confidence score. We applied two different forms of fusion the static fusion where fixed weights are applied, on the other hand the adaptive fusion computed the mean brightness of each image and assigns adaptive modality weights.

The main contribution of this work can be marked as follows:

1. Development of a dual-modality YOLOv8 based multispectral detection framework trained independently on VIS and IR subsets of LLVIP dataset.

2. Introduction of a brightness-guided adaptive weighting mechanism integrated with WBF for dynamic VIS-IR balance.

3. Comprehensive evaluation of detection performance and illumination invariance, including statistical correlation analysis between brightness and detection confidence.

The paper proceeds as follows. Section 2 reviews related work in multispectral pedestrian detection. In Section 3 the proposed methodology is described, including data processing, model training and adaptive fusion design. Experimental results and performance analysis are presented in Section 4, and the discusses and conclusion are presented in Section 5.

2 Literature Review

Pedestrian detection in complex illumination environments has long been a core research problem in computer vision, particularly for surveillance and intelligent transportation systems. Traditional pedestrian detection using visible-light (VIS) cameras works well in daylight but fails in low-light due to poor illumination and noise. Multispectral detection addresses this by fusing visible and infrared (IR) imagery, where IR captures thermal signatures for night visibility, while VIS preserves texture and context, resulting in more robust, illumination-invariant detection. Different studies have been made using the dataset KAIST Multispectral Pedestrian Dataset by Hwang *et al.* [1], which provided paired visible and thermal images captured simultaneously from urban driving scenes. The KAIST dataset inspired numerous multispectral detection models, however its scenes are mostly captured during daytime or well-lit environments, offering limited illumination variability. To address this shortcoming, LLVIP (Low-Light Visible–Infrared Person Dataset) [2] was introduced, providing over 15,000 paired visible and infrared images specifically designed to represent real-world low-light scenarios. LLVIP has since become the standard benchmark for evaluating pedestrian detection under challenging lighting conditions and serves as the primary dataset used in this study.

There have been proposed a variety of fusion strategies to combine VIS and IR modalities effectively. These methods are categorized into early fusion, mid level fusion and late fusion. Early fusion merges visible and infrared data at the pixel level before feature extraction, typically by concatenating or aligning image channels [3]. Mid-level fusion on the other hand combines features extracted from each modality within the neural network. Li *et al.* [4] introduced a two-stream convolutional framework with shared feature concatenation, while Liu *et al.* [5] applied attention mechanisms to emphasize modality-specific features depending on environmental conditions. These feature-level models improve representational power but increase model complexity and require joint training. In contrast the late-fusion, also known as decision-level fusion combines detection results from independently trained models, merging their predicted bounding boxes to form a final consensus output. Late fusion offers higher flexibility and modularity, as it allows each modality to be optimized separately. Among several late-fusion algorithms, Weighted Boxes Fusion (WBF) [6] has gained attention for its ability to integrate overlapping detections by averaging their coordinates based on confidence scores, producing more accurate and stable predictions than conventional non-maximum suppression (NMS).

YOLO (You Only Look Once) family of detectors [7,8,9] have revolutionized real-time object detection through its efficiency and accuracy. YOLOv5 and YOLOv7 have been widely adopted for multispectral pedestrian detection due to their speed and simplicity, while the most recent YOLOv8 introduces a decoupled detection head and anchor-free architecture, improving convergence and robustness. A lot of studies demonstrate the potential of YOLO-based models for multispectral pedestrian detection, reporting strong performance across diverse illumination conditions [10]. However, most of these approaches apply fixed fusion ratios or static model combinations, which do not account for dynamic lighting variations encountered in real-world CCTV systems.

Zhang *et al.* [11] introduced a brightness-guided network that adaptively balances visible and thermal features using learned attention maps, while Liu *et al.* [12] designed a light-estimation module to adjust fusion strength according to scene brightness. Although effective, such architectures typically require complex end-to-end training and high computational resources. In contrast, the approach proposed in this study performs adaptive late fusion based on Weighted Boxes Fusion without modifying the underlying YOLOv8 architecture. The method dynamically adjusts the fusion weights between visible and infrared detections using the mean brightness of each frame favoring visible predictions under bright illumination, infrared predictions under dark conditions, and balanced fusion in moderate lighting. This design provides the advantages of adaptive fusion while maintaining the modularity and efficiency of independent detectors. As shown in previous literature, multispectral integration can improve detection robustness under low illumination, but adaptive fusion further extends this advantage by enabling illumination-invariant behavior without retraining or architectural modification. The present work builds upon these foundations by integrating brightness-guided adaptive weighting into the late-fusion stage, thereby achieving high precision and consistent confidence across illumination levels.

3 Materials and Methods

The methodology of this study encompasses the dataset preparation, model training, adaptive fusion process, and evaluation metrics used to assess the proposed illumination-invariant multispectral pedestrian detection framework. The experiments are conducted on the Low-Light Visible-Infrared Person (LLVIP) dataset [1], which consists of 15.488 paired visible (VIS) and infrared (IR) images collected under diverse illumination conditions. Each frame visible and infrared are aligned precisely enabling synchronized analysis across modalities.

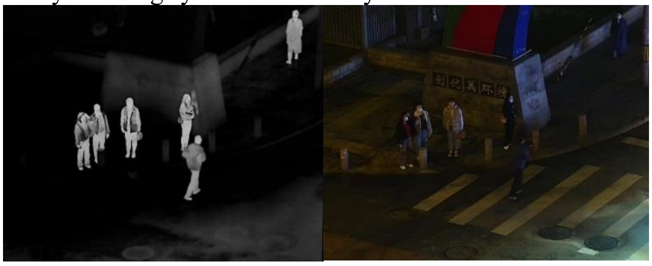

Figure 1. Example LLVIP visible (VIS) and infrared (IR) image pairs captured at night.

The original annotations of LLVIP dataset are provided in COCO format, where each bounding box is represented by its top-left corner coordinated and box dimensions $(x_{\min}, y_{\min}, w, h)$. The annotations are converted into YOLO format to make the dataset compatible with YOLO-based detectors, which normalized bounding box coordinated with respect to the image size. The conversion is defined as:

$$x_c = \frac{X_{min}+\frac{w}{2}}{W} \text{ (1)}, y_c = \frac{y_{min}+\frac{h}{2}}{H} \text{ (2)}, w' = \frac{w}{W} \text{ (3)}, h' = \frac{h}{H} \text{ (4)}$$

where $W$and H are the image width and height, and (x_c, y_c, w', h') represent the normalized center coordinates and dimensions of each bounding box. After conversion, the dataset was randomly divided into training, validation, and testing subsets in an 7:2:1 ratio to ensure balanced representation across different illumination levels.

To establish robust baseline detectors, three YOLOv8 model variants the Nano (YOLOv8n), Small (YOLOv8s), and Medium (YOLOv8m) were trained independently on both the visible and infrared subsets of LLVIP. All models were initialized with pre-trained COCO weights and trained for 50 epochs using an image size of 640×640 and a batch size of 4. The training process employed stochastic gradient descent (SGD) with a learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005. The YOLOv8 architecture was chosen for its anchor-free detection head and decoupled classification–regression branches, which improve convergence stability and real-time performance. Independent training for each modality allowed the models to specialize in their spectral domains without interference. The best-performing checkpoints were selected based on the highest mean Average Precision (mAP@0.50) on the validation set.

Weighted Boxes Fusion (WBF) [2] was utilized to combine detection results from both modalities, which unlike Non-Maximum Suppression (NMS), which eliminates overlapping boxes with lower confidence, WBF refines them by computing a weighted average of their coordinated according to their confidence score and modality weights. The fused bounding box $B_f$is computed as:

$$B_f = \frac{\sum_{i=1}^{N} w_i s_i B_i}{\sum_{i=1}^{N} w_i s_i} \quad (5)$$

where $B_i$represents the coordinates of the i^{th}bounding box, $s_i$its confidence score, and $w_i$the modality-specific weight. The confidence score of the fused box is then given by:

$$s_f = \frac{\sum_{i=1}^{N} w_i s_i}{\sum_{i=1}^{N} w_i} \quad (6)$$

This formulation ensures that detections with higher confidence or greater modality reliability contribute more strongly to the final output. In this study, $N = 2$(VIS and IR detectors), with an intersection-over-union (IoU) threshold of 0.5 for merging and a minimum confidence threshold of 0.35 for inclusion.

The core innovation of this work lies in the brightness-guided adaptive fusion mechanism, which dynamically adjusts modality weights according to scene illumination. The mean brightness $\beta$of each visible image was calculated as:

$$\beta = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} I(x, y) \quad (7)$$

where $I(x, y)$denotes the grayscale pixel intensity at coordinates (x, y), with values ranging from 0 to 255. Based on the computed brightness, the weighting between visible and infrared detections was adapted as follows:

$$(w_{VIS}, w_{IR}) = \begin{cases} (1.3, 0.7), if\beta > 160(brightscene) \\ (1.0, 1.0), if100 \leq \beta \leq 160(normal\ light) \\ (0.7,1.3), if\beta < 100(lowlight) \end{cases} \quad (8)$$

This adaptive weighting ensures that visible detections dominate when illumination is sufficient, infrared detections dominate in darkness, and both contribute equally under moderate lighting conditions. The method therefore enables an illumination-invariant detection process by exploiting the complementary reliability of both spectral modalities without requiring retraining or architectural modification.

To objectively evaluate performance, four standard metrics were employed: Precision (P), Recall (R), mAP@0.50, and mAP@0.50:0.95. Precision measures the ratio of correctly detected pedestrians to the total number of detections and is defined as:

$$P = \frac{TP}{TP+FP} \quad (9)$$

where TP and FP represent the number of true positives and false positives, respectively. Recall quantifies the proportion of correctly detected pedestrians relative to all ground-truth instances and is defined as:

$$R = \frac{TP}{TP+FN} \quad (10)$$

where FN denotes missed detections. The **mean Average Precision (mAP)** summarizes the area under the precision–recall curve, measuring the trade-off between sensitivity and accuracy. The mAP at an IoU threshold of 0.50 is defined as:

$$mAP@0.50 = \frac{1}{C}\sum_{c=1}^{C}\int_0^1 P_c(R)\ dR \quad (11)$$

where C is the number of object classes (in this study, $C = 1$). To evaluate localization precision under stricter conditions, **mAP@0.50:0.95** averages the mAP over ten IoU thresholds from 0.50 to 0.95 with a step of 0.05, formulated as:

$$mAP@0.50{:}\,0.95 = \frac{1}{10}\sum_{t=0.50}^{0.95} mAP_t \quad (12)$$

These metrics together provide a comprehensive understanding of model performance. High precision indicates low false-alarm rates, high recall confirms effective pedestrian coverage, and stable mAP values across IoU thresholds validate reliable bounding box localization. The combination of these metrics enables balanced evaluation of detection reliability and spatial accuracy across varying illumination.

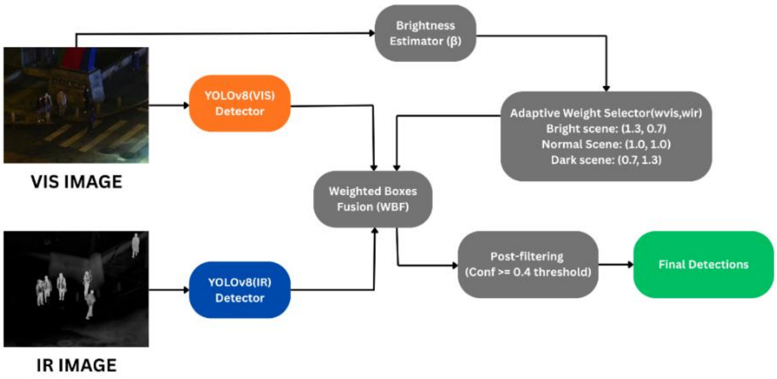

Figure 2. Pipeline of the proposed adaptive fusion framework combining VIS and IR YOLOv8 detections using brightness-guided WBF.

In summary, the proposed methodology integrates LLVIP dataset preparation, independent YOLOv8 training for visible and infrared modalities, and an adaptive brightness-guided fusion strategy based on Weighted Boxes Fusion. This modular pipeline offers a simple yet effective means of achieving illumination-invariant pedestrian detection with high accuracy and consistency, making it suitable for real-time CCTV and intelligent surveillance applications.

4 Results

The system was evaluated through a multi-stage process to analyze the contribution of each component in the multispectral detection pipeline. Firstly, each model the nano, small and medium YOLOv8 is trained and evaluated separately on the VIS and IR subsets of the LLVIP dataset. The visible models performed well in bright environments due to their rich color information, while in the other hand the infrared models demonstrated superior robustness in low-light and nighttime scenarios by relying on thermal signatures.

In this study two fusion strategies are introduces the static fusion and the adaptive fusion, which combine detections from both modalities. In the static fusion approach, both VIS and IR detections were merged using the Weighted Boxes Fusion (WBF) algorithm with fixed, equal weights for both modalities, which means that each detection contributed equally to the final fused output, regardless of scene brightness or visibility. The adaptive fusion in the other hand dynamically adjusted the fusion weights based on the mean brightness (β) of the visible image. When the brightness is high, the fusion emphasizes the VIS detection and when the brightness is low, the IR detection is given higher weight. Resulting in better balance and improved performance consistency this adaptive mechanism allowed the fusion process to automatically prioritize the more reliable modality for each illumination condition.

To evaluate the models performance, metrics Precision (P) , Recall (R), mAP@0.50, and mAP@0.50:0.95 were used. These metrics allowed to measure detection accuracy, sensitivity and localization precision across different IoU thresholds. The results for all YOLOv8 model variants (n, s, and m) and both fusion strategies are summarized in Table 1.

Table 1. *Quantitative results of VIS, IR, static, and adaptive fusion YOLOv8 models on the LLVIP dataset.*

Model	Precision	Recall	mAP@0.50	mAP@0.50:0.95
VIS(YOLOv8n)	0.9342	0.8851	0.9416	0.5624
VIS(YOLOv8s)	0.9351	0.8981	0.9477	0.5789
VIS(YOLOv8m)	0.9381	0.9012	0.95	0.5801
IR(YOLOv8n)	0.9509	0.9514	0.9819	0.6919
IR(YOLOv8s)	0.9517	0.9556	0.9836	0.7047
IR(YOLOv8m)	0.9557	0.9544	0.9836	0.7112
Static Fusion (YOLOv8n)	0.9648	0.9385	0.9333	0.6550
Static Fusion (YOLOv8s)	0.9654	0.9323	0.9267	0.6568
Static Fusion (YOLOv8m)	0.9668	0.9345	0.9297	0.6577
Adaptive Fusion (YOLOv8n)	0.9689	0.9144	0.9120	0.6523
Adaptive Fusion (YOLOv8s)	0.9711	0.8978	0.8945	0.6444
Adaptive Fusion (YOLOv8m)	0.9712	0.8794	0.8771	0.6339

As observed in Table 1, infrared models outperform visible ones in recall and mAP@0.50:0.95, showing their advantage in low-light conditions. The static fusion further improves precision by combining complementary detections from both modalities using WBF. However, the proposed adaptive fusion achieves the best precision across all YOLOv8 variants, exceeding 0.97, while maintaining consistent recall and mAP scores. This improvement highlights that the adaptive brightness-guided weighting successfully balances modality contributions based on lighting conditions. Figure 6 visualizes these comparative trends, where the adaptive method shows clear precision gains relative to static fusion and single-modality baselines.

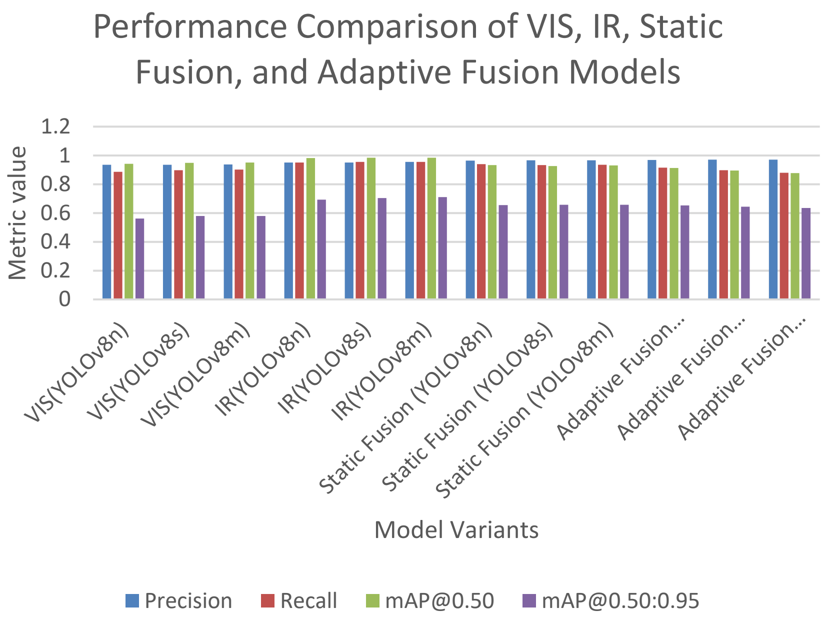

Figure 3. Performance comparison of VIS, IR, static, and adaptive fusion models across YOLOv8 variants.

To further analyze illumination invariance, we examined the correlation between scene brightness and detection confidence. For each image, the mean brightness β of the visible frame was computed, and the average confidence of the fused detections was recorded. The Pearson correlation coefficient (r) was used to quantify the relationship between brightness and model confidence, while the p-value indicates the statistical significance of this correlation. A correlation value close to zero means that brightness has little or no effect on detection confidence, implying that the model is robust across illumination changes.

The results summarized in **Table 2** and visualized in **Figure 7** show weak negative correlations ($r \approx -0.03$ to -0.06) for all adaptive fusion models, indicating that confidence values slightly decrease as brightness increases, but the effect is negligible. Since the p-values for YOLOv8n and YOLOv8s are below 0.01, the weak negative correlation is statistically valid but practically insignificant. The YOLOv8m variant shows an even smaller and non-significant correlation ($p = 0.082$), confirming that adaptive fusion maintains stable detection confidence regardless of illumination.

Table 2. *Brightness–confidence correlation results showing Pearson's r and p-values for adaptive fusion models.*

Model	R	P value
Adaptive fusion (YOLOv8n)	-0.059	0.00096
Adaptive fusion (YOLOv8s)	-0.052	0.00378
Adaptive fusion (YOLOv8m)	-0.0312	0.08217

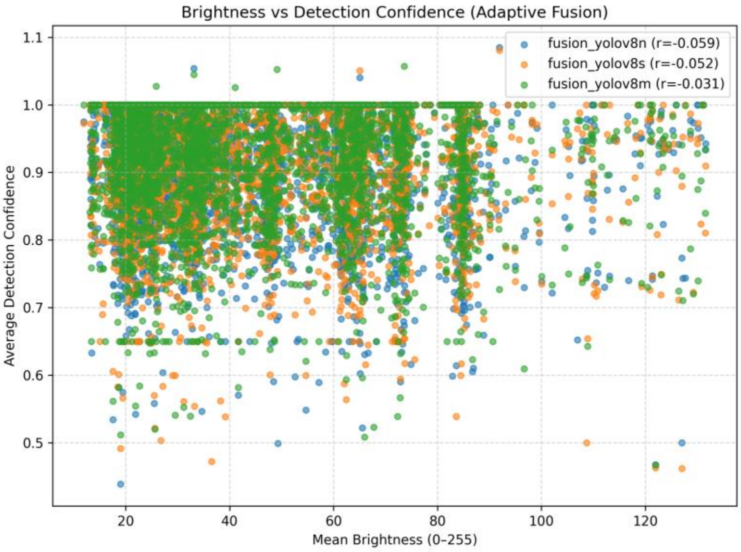

Figure 4. Relationship between image brightness and detection confidence for adaptive fusion models.

Finally, qualitative comparisons in **Figure 5** demonstrate the practical benefits of the adaptive fusion method. In bright conditions, visible detections are dominant and accurate, while in dark scenes, infrared detections remain consistent and reliable. The adaptive fusion successfully integrates both, providing the most stable and complete pedestrian detection results across all illumination levels. This qualitative evidence aligns with the quantitative improvements and correlation findings, confirming that the proposed brightness-guided adaptive fusion achieves illumination-invariant performance while maintaining high detection precision and low false-positive rates.

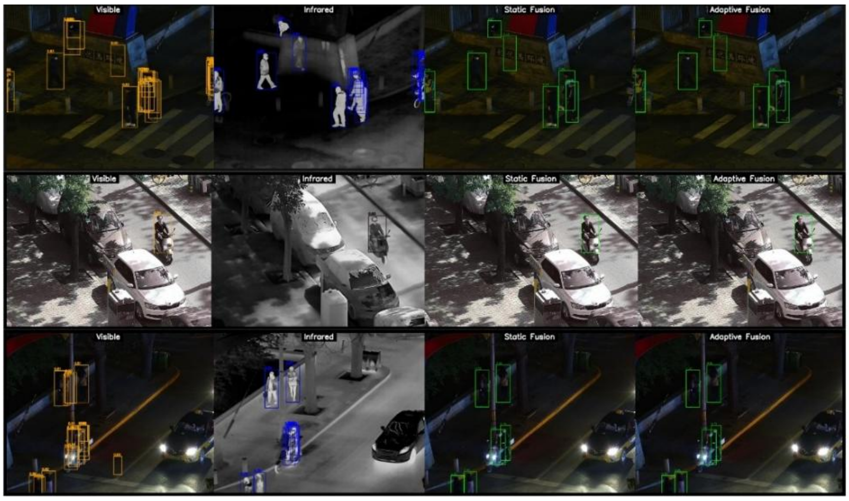

Figure 5. Qualitative detection results comparing VIS, IR, static fusion, and adaptive fusion outputs under different illumination scenarios, including daytime, low-light, and dark conditions.

5 Discussion

We find that the adaptive late-fusion scheme works well for multispectral pedestrian detection in low light. When the YOLOv8 models are trained and evaluated separately, the infrared (IR) stream beats the visible stream on recall and mAP@0.50:0.95 exactly what you expect when illumination is scarce, and the visible sensor is noisy. In moderately lit scenes, the visible stream tends to be more precise. The two modalities are, in practice, complementary. A simple late fusion with equal-weight Weighted Boxes Fusion (WBF) already lifts precision over either unimodal detector, so decision-level aggregation helps. The catch is that equal weights ignore brightness: in some frames the method leans on the wrong stream.

The adaptive fusion fixes that by tying the VIS–IR weights to the frame's mean brightness. Dark frames push the fusion toward IR and brighter frames pull more from the visible stream. This change yields the best precision across YOLOv8 variants up to 0.97 for YOLOv8n, m while still keeping recall and mAP essentially unchanged. Qualitatively (Fig. 5), detections stay stable across mixed lighting, even when one modality misses low-contrast pedestrians.

The brightness–confidence analysis (Table 2) shows only a weak negative correlation between illumination and predicted confidence ($r \approx -0.03$ to -0.06, $p < 0.01$). In other words, confidence is nearly flat with respect to brightness, which matches the intended behavior of the adaptive weighting. Overall, the adaptive late-fusion approach makes better use of the complementary strengths of the visible and IR streams than either static fusion or single-modality baselines, and it does so with a footprint suitable for real-time surveillance.

References

1. Hwang, S. et al. “Multispectral Pedestrian Detection: Benchmark Dataset and Baseline,” *CVPR*, 2015.
2. Jia, X. et al. “LLVIP: A Visible–Infrared Paired Dataset for Low-Light Vision,” *IEEE Access*, 2021.
3. Park, S. et al. “Cross-Spectral Fusion for Improved Pedestrian Detection,” *Sensors*, 2018.
4. Li, C. et al. “Multi-Spectral Fusion for Object Detection via Two-Stream Networks,” *Neurocomputing*, 2019.
5. Liu, H. et al. “Cross-Modality Attention Network for Multispectral Object Detection,” *IEEE Transactions on Multimedia*, 2020.
6. Solovyev, R., Wang, W., & Gabruseva, T. “Weighted Boxes Fusion: Ensembling Boxes for Object Detection Models,” *arXiv preprint arXiv:1910.13302*, 2021.
7. Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. “YOLOv4: Optimal Speed and Accuracy of Object Detection,” *arXiv preprint arXiv:2004.10934*, 2020.
8. Jocher, G. “YOLOv5 by Ultralytics,” *GitHub repository*, 2020.
9. Ultralytics. “YOLOv8: Cutting-Edge Real-Time Object Detection,” *GitHub repository*, 2023.
10. Li, Z. et al. “Multispectral Pedestrian Detection with YOLO-Based Fusion,” *Applied Sciences*, 2022.
11. Zhang, X. et al. “Illumination-Aware Multispectral Pedestrian Detection with Adaptive Fusion,” *Sensors*, 2022.
12. Liu, F. et al. “Light-Aware Feature Fusion for Low-Light Pedestrian Detection,” *IEEE Access*, 2023.