I. J. Engineering and Manufacturing, 2026, 5, 312-329
Published Online on October 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijem.2026.05.17

Spatially Adaptive Infrared–visible Fusion with EfficientDet for Night-time Multi-class Intrusion Detection

Rithick S.
School of Computer Science and Tech., Karunya Institute of Technology and Sciences, India
E-mail: rithick.s@karunya.edu
ORCID iD: https://orcid.org/0009-0008-7997-126X

Jenefa A.*
School of Computer Science and Tech., Karunya Institute of Technology and Sciences, India
E-mail: jenefaa@karunya.edu
ORCID iD: https://orcid.org/0000-0002-6697-1788
*Corresponding Author

Abirami M. K.
School of Computer Science and Tech., Karunya Institute of Technology and Sciences, India
E-mail: abirami.ms@karunya.edu
ORCID iD: https://orcid.org/0009-0006-1691-3041

G. Sheeba Merlin
School of Computer Science and Tech., Karunya Institute of Technology and Sciences, India
E-mail: sheebamerlin@karunya.edu
ORCID iD: https://orcid.org/0000-0001-6728-8182

Antony Taurshia
School of Computer Science and Tech., Karunya Institute of Technology and Sciences, India
E-mail: antonytaurshia@karunya.edu
ORCID iD: https://orcid.org/0000-0001-9129-1859

Lincy A.
Computer Science & Engineering, National Engineering College, Kovilpatti, India
E-mail: lincy-cse@nec.edu.in
ORCID iD: https://orcid.org/0000-0002-4104-6529

Received: May 21, 2026; Revised: June 15, 2026; Accepted: July 22, 2026; Published: October 8, 2026

Abstract: Night-time perimeter monitoring in low illumination remains challenging because cluttered terrain, partial occlusion, and thermal crossover conditions simultaneously distort object boundaries and visual cues. Visible spectrum sensing loses discriminative texture at night, while infrared sensing preserves thermal salience but lacks structural context. Conventional pipelines that rely on single-modality detection or simple blending commonly produce unstable recall or avoidable false alarms when the background changes. The present paper presents an EfficientDet Fusion Intrusion Detector (EFID) which combines registered infrared and visible frames using an attention-guided weighting module and a spatially adaptive activity-weighted pixel fusion step. The feature-level attention weights estimate the reliability of each modality and guide the activity-weighted pixel fusion stage. The fused RGB image is resized from 1024 × 768 to 896 × 896 before being processed by EfficientDet-D3. The resultant fusion retains visible structural edges, while infrared target evidence is simultaneously enhanced. The resulting three-channel representation is processed by an EfficientDet D3 detector augmented with a bidirectional feature pyramid network (BiFPN) for multi-scale localisation under night conditions and clutter. Experiments use the public Multi-scenario Multi-modality Fusion and Detection (M³FD) benchmark containing 4,200 aligned infrared and visible pairs at 1024×768 resolution, annotated with 33,603 bounding boxes across six classes. The proposed EFID achieves 91.7% mAP@0.5 and 67.3% mAP@0.5:0.95, with 93.4% precision and 89.8% recall at 34.2 FPS on an NVIDIA RTX 3080. Systematic ablation

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

confirms the independent contribution of the attention module, the activity-weighted pixel fusion term, and each BiFPN iteration. Cross-dataset experiments demonstrate that zero-shot transfer retains 84.1% mAP@0.5, rising to 86.0% with 10% target fine-tuning. The results indicate practical night robustness and real-time feasibility, while extreme occlusion and calibration drift remain limiting factors that motivate alignment-aware training and lightweight deployment optimisation as future work.

Index Terms: Infrared and visible image fusion, Night-time intrusion detection, EfficientDet D3, Bidirectional feature pyramid network, Multimodal surveillance, Real-time object detection, Attention-guided fusion

1. Introduction

Surveillance missions under low-lighting or nocturnal settings require sensor modalities that are complementary since no one sensor modality can provide adequate evidence to facilitate effective object recognition under all possible operating environments. Visible spectrum cameras can record fine detail of colour and texture, but performance is sharply impaired when the light is dim, when shadows are cast, or where the camouflage of the targets is antagonistic. In contrast, infrared sensors obtain thermal differences in emission, which are quite independent of ambient light and colour of the surface. The resulting imagery is often not fine in terms of edge structure and can be lost in thermal crossover of the target and background where the target and background temperature converge. In this regard, the combination of these complementary data sources has been observed to be a popular area of research in improving on image quality and the final performance of detection processes. Even though considerable progress has been made in the methodology of fusion, there is still a substantive difference between fusion-oriented metrics of image quality and fusion-oriented metrics of detection performance. Many published fusion pipelines are tested based on perceptual measures such as entropy, spatial frequency and structural similarity, but with no attempt to determine the effect of the fused image on detector accuracy and recall. Separate optimisation of fusion and detection may produce visually sharp fused images without improving downstream object localisation, since perceptual fusion objectives may not align with the feature requirements of the detector. Computational feasibility is yet another challenge. A significant portion of multispectral detectors with high performance require two-stage architectures or computationally-intensive multi-branch networks that are not scalable to achieve real-time throughput in one GPU, thus limiting their use to deployed surveillance systems. Frame drops or buffering latency in constant perimeter monitoring means that at times of transient intrusion, a detection may be missed, making the throughput requirement an operational constraint and not an engineering footnote.

The third challenge is the scope of categories and situations. Numerous published assessments focus on pedestrian detection or one lighting scenario. Contrary to this, real outdoor surveillance requires the detection of people, diverse vehicle categories, and infrastructure such as luminaires, under different lighting conditions: day, grey, night, and extreme artificial light, as well as adverse weather such as fog and rain.

The proposed work addresses all three challenges through a unified pipeline called the EfficientDet Fusion Intrusion Detector (EFID). The key design choices are: (i) an attention-guided fusion module that produces spatially adaptive modality weights so that each region of the fused image reflects the more informative sensor at that location; (ii) a spatially adaptive activity-weighted pixel fusion that explicitly preserves visible structural edges in the fused pixel domain to support accurate boundary-level localisation; and (iii) an EfficientDet D3 detector with a bidirectional feature pyramid network (BiFPN) that propagates both top-down and bottom-up context for stable multi-scale detection.

The main contributions of this paper are as follows.

- The main contribution is the coordinated use of attention-guided modality weighting, activity-based pixel fusion, and EfficientDet-based multi-scale detection within a unified night-time surveillance framework.
- A detection-oriented end-to-end pipeline that couples the proposed fusion module with EfficientDet D3 and a three-iteration BiFPN head for scale-consistent localisation in cluttered night scenes, achieving 34.2 FPS at 896×896 input resolution.
- A systematic ablation study isolates and quantitatively measures the contribution of the individual parts, which include fusion, attention weighting, activity-weighted pixel fusion, and BiFPN depth of iteration, which gives clear and quantitative advice to the practitioner.
- Cross-dataset transfer and scenario-wise analyses demonstrate the model's generalisation to unseen sensor and illumination conditions without test-time domain adaptation.

The remaining sections of this paper will be organized in the following manner. Section 2 reviews the earlier research on infrared visible fusion and multispectral object detection. Section 3 describes the methodology of the system and provides the full theoretical derivation. Section 4 describes the experimental setup, data set and benchmarking. All quantitative results are presented in section 5. Section 6 mentions findings, their practical

applicability, and limitations of the study. Lastly, Section 7 provides a summary of the work and advances future research directions.

2. Related Work

Liu et al. [1] (2022) constructed the M³FD benchmark of 4,200 aligned infrared-visible pairs across six categories and four illumination settings, establishing a reliable evaluation standard for fusion-driven detection. Tan et al. [2] (2020) proposed EfficientDet, which compound-scales backbone depth and resolution through a bidirectional feature pyramid network to achieve strong accuracy-efficiency trade-offs. Tang et al. [3] (2022) introduced PIAFusion, a progressive fusion network that assigns illumination-conditioned weights to infrared and visible modalities under night-time and mixed-lighting scenarios. Roy et al. [4] (2023) developed a multimodal fusion transformer for remote sensing classification that learns cross-modal spectral attention and generalises across diverse sensor combinations. Ma et al. [5] (2025) studied scene-adaptive unsupervised crowd counting and confirmed that domain shift severely impairs models trained on fixed illumination distributions. Ni et al. [6] (2024) proposed cross-dimensional multispectral edge fusion for traffic scene segmentation, showing that preserving modality-specific edges is critical under changing ambient light. Santhiya et al. [7] (2024) applied Vision Transformer models to traffic sign detection in autonomous driving, demonstrating that attention-based architectures capture fine-grained sign features more reliably than standard convolutional methods. Yang et al. [8] (2023) designed an instance-aware attention mechanism for multi-person pose estimation that improves localisation accuracy for occluded targets. Yang et al. [9] (2024) proposed a multi-scale information integration framework for infrared-visible fusion that aligns feature hierarchies before blending to improve textured-region representation. Taurshia et al. [10] (2025) applied EfficientDet-based monitoring to pothole detection, demonstrating that scalable detection architectures transfer effectively to infrastructure surveillance under variable lighting. Li and Wu [11] (2018) introduced DenseFuse, a compact densely connected encoder-decoder that projects infrared and visible modalities into a shared latent space and became a widely adopted fusion baseline. Li et al. [12] (2021) proposed RFN-Nest, an end-to-end residual fusion network that propagates fine-grained texture across both modalities to reduce deep compression losses. Ma et al. [13] (2020) generalised adversarial fusion through DDcGAN, a dual-discriminator conditional model that enforces texture and edge coherence across multiple output resolutions. Xu et al. [14] (2020) unified several fusion objectives into FusionDN, a single densely connected network that maintains competitive perceptual performance with minimal architectural overhead. Santhiya et al. [15] (2024) investigated lane detection from a segmentation perspective, reinforcing that segmentation-driven detection holds up under poor visibility conditions relevant to night-time perception. Xiang et al. [16] (2025) showed that joint optimisation of infrared-visible fusion and detection consistently outperforms decoupled pipelines regardless of the fusion backbone used. Gallagher and Oughton [17] (2025) surveyed multispectral YOLO adaptations and found high perceptual fusion quality but reduced detection recall in adverse settings, identifying this gap as a key unsolved problem. Wang et al. [18] (2025) proposed DICFusion, which integrates deep semantic features to suppress background clutter and sharpen object edges in night-time outputs. Madhasu and Pande [19] (2025) paired nocturnal infrared surveillance with deep learning for wildlife protection, confirming that thermal evidence is essential for detecting low-contrast objects in dark environments. Joshua Samuel et al. [20] (2023) developed a deep learning-based license plate recognition system for counter-terrorism that achieves reliable detection across variable lighting and cluttered backgrounds. Tu et al. [21] (2025) demonstrated that multidomain sensor fusion captures human fatigue patterns from non-invasive pressure data, supporting the value of multimodal sensing in safety-critical monitoring. Vasantharaj et al. [22] (2025) validated that scaled YOLOv11 detectors sustain real-time throughput for wild animal detection in resource-constrained deployments. Wang et al. [23] (2025) proposed a boundary dynamic perception network for low-light railway environments that enhances edge response under photon-limited conditions. Rocha et al. [24] (2025) surveyed electrical grid asset inspection and identified detection robustness under lighting variability as a persistent open challenge in aerial surveillance. Aman et al. [25] (2025) introduced an adaptive hybrid detector that adjusts its strategy based on contextual cues, achieving stable performance across diverse scene complexities. Ande et al. [26] (2023) established that deep convolutional networks significantly reduce missed detections in real-time obstacle localisation on cluttered rail tracks. Rampriya et al. [27] (2024) validated that multimodal deep fusion achieves superior boundary-level accuracy for fault detection and semantic segmentation on railway infrastructure. Santhiya et al. [28] (2024) proposed BiPoolNet, a pooling-enhanced UNet architecture that improves lane detection reliability for autonomous vehicles under degraded visibility. Isaac et al. [29] (2025) combined attention-enhanced detection with real-time voice feedback for road sign recognition, maintaining high accuracy under variable illumination. RS et al. [30] (2024) published a UAV-acquired multimodal dataset for railroad segmentation, emphasising that registration-quality infrared-visible alignment is a prerequisite for effective small-object localisation.

3. System Methodology

This section presents the theoretical formulation of the proposed EFID pipeline, covering the problem definition, attention-guided fusion module, activity-weighted pixel fusion step, bidirectional feature pyramid aggregation, prediction heads, optimisation objective, and evaluation metrics.

3.1. Problem Definition

Let $\mathbf{I}_{ir} \in \mathbb{R}^{H\times W}$ denote a registered infrared frame and $\mathbf{I}_{vis} \in \mathbb{R}^{H\times W\times 3}$ the aligned visible spectrum frame captured at the same time instant by a co-located sensor pair. The registration step that produces spatial alignment is treated as pre-processing and is not modelled here. The fusion goal is to construct a fused representation $\mathbf{I}_f \in \mathbb{R}^{H\times W\times 3}$ according to

$$\mathbf{I}_f = \mathcal{F}\big(\mathbf{I}_{ir}, \mathbf{I}_{vis}; \Theta_f\big) \tag{1}$$

where Θ_f denotes learnable fusion parameters. The detector then maps $\mathbf{I}_f$ to a set of predicted detections

$$\widehat{\mathcal{B}} = \mathcal{D}\big(\mathbf{I}_f; \Theta_d\big) \tag{2}$$

where $\widehat{\mathcal{B}}$ is the set of predicted bounding boxes with class labels and confidence scores and Θ_d denotes detection parameters. The joint optimisation minimises the detection loss with respect to both Θ_f and Θ_d, so that fusion parameters adapt to what the detector needs rather than to a standalone perceptual objective.

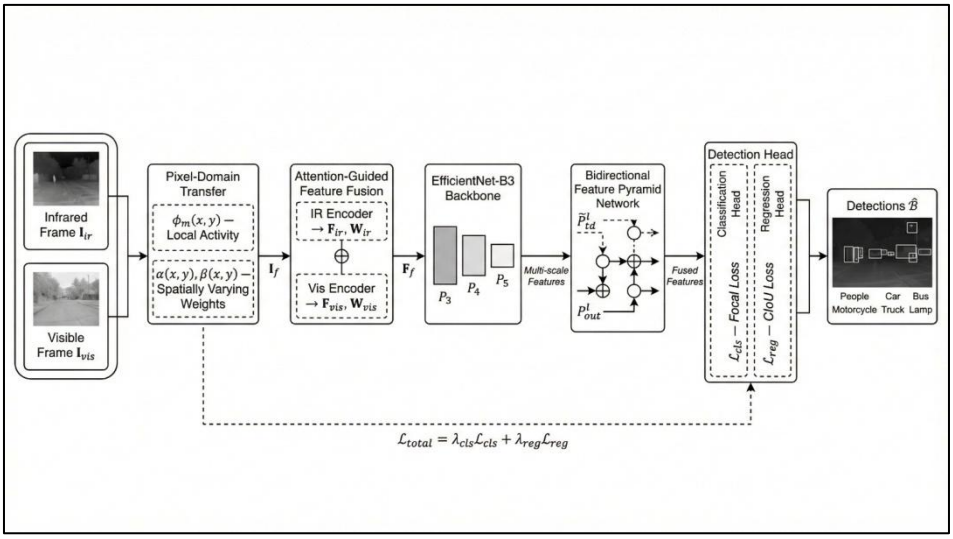


Fig. 1. Overall architecture of the proposed EfficientDet Fusion Intrusion Detector (EFID).

Figure 1 illustrates the complete EFID pipeline. The two sensor streams are first encoded by a shared-weight feature extractor to produce modality-specific feature maps. An attention block assigns spatially varying importance weights to each map; the feature-level attention module estimates the relative reliability of the infrared and visible modalities. The registered source images are then processed through spatially adaptive pixel-domain fusion to preserve thermal salience and visible structural information. The resulting three-channel image is fed into an EfficientNet-B3 backbone, and its P_3–P_7 feature levels are aggregated by the BiFPN before entering the class and box prediction heads.

3.2. Attention-Guided Feature Fusion

Both modality frames are projected into a common feature space through a shared lightweight encoder consisting of three convolutional layers with batch normalisation and ReLU activation. Let $\mathbf{F}_{ir} \in \mathbb{R}^{H\prime\times W\prime\times C}$ and $\mathbf{F}_{vis} \in \mathbb{R}^{H\prime\times W\prime\times C}$ denote the resulting feature maps. The attention function $\mathcal{A}(\cdot)$ is a channel-squeeze-and-excitation module [31] that first performs global average pooling over spatial dimensions, applies two fully connected layers with a bottleneck ratio of 16, and outputs a channel-wise scalar vector. The per-modality attention weights are

$$\mathbf{W}_{ir} = \sigma\big(\mathcal{A}(\mathbf{F}_{ir})\big), \quad \mathbf{W}_{vis} = \sigma\big(\mathcal{A}(\mathbf{F}_{vis})\big) \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid activation. The fused feature map is formed by the normalised weighted combination

$$\mathbf{F}_f = \frac{\mathbf{W}_{ir} \odot \mathbf{F}_{ir} + \mathbf{W}_{vis} \odot \mathbf{F}_{vis}}{\mathbf{W}_{ir} + \mathbf{W}_{vis} + \epsilon} \tag{4}$$

where $\odot$ denotes element-wise multiplication and $\epsilon = 10^{-4}$ prevents division by zero. The denominator normalisation ensures that the output magnitude is bounded regardless of the absolute weight values, which stabilises training when one modality is temporarily uninformative (for example, during thermal crossover). The feature-level attention stage and the pixel-domain fusion stage perform complementary functions. Feature attention captures the semantic importance of each modality, while pixel-domain weighting preserves local intensity variations and object boundaries at the original image resolution. The registered infrared and visible images are consequently used in the next stage to construct the fused image supplied to the detector. The feature-level attention stage estimates the relative reliability of the infrared and visible modalities by assigning adaptive importance weights to their encoded representations. These modality-specific cues are then used to guide the activity-weighted pixel fusion process described in Section 3.3. The resulting fused RGB image integrates thermal salience and visible structural information before being forwarded to the EfficientDet-based detection network. The feature-level attention stage estimates the relative reliability of the infrared and visible modalities. These modality-specific cues guide the activity-weighted pixel fusion stage described in Section 3.3.

3.3. Spatially Adaptive Activity-Weighted Pixel Fusion

To preserve visible structural edges that support boundary-level localisation, the fused pixel intensity $\mathbf{I}_f$ is constructed with spatially varying blending coefficients. The luminance channel of the visible frame is first extracted as $\mathbf{I}_{vis}^{lum} = 0.299\,R + 0.587\,G + 0.114\,B$. To complement the semantic information obtained from feature-level attention, pixel-domain fusion is applied directly to the registered source images. This stage retains local thermal contrast from the infrared image and structural details from the visible image. The fused pixel intensity is constructed using spatially varying activity weights calculated from the local variations of both modalities. The fused intensity is then

$$\mathbf{I}_f(x,y) = \alpha(x,y)\,\mathbf{I}_{ir}(x,y) + \beta(x,y)\,\mathbf{I}_{vis}^{lum}(x,y) \tag{5}$$

where the weight terms are derived from the local activity of each modality. Specifically,

$$\alpha(x,y) = \frac{\phi_{ir}(x,y)}{\phi_{ir}(x,y) + \phi_{vis}(x,y)}, \beta(x,y) = 1 - \alpha(x,y) \tag{6}$$

and the activity measure ϕ_m at position (x,y) is the local standard deviation computed over an $N \times N$ neighbourhood $\mathcal{N}$:

$$\phi_m(x,y) = \sqrt{\frac{1}{|\mathcal{N}|}\sum_{(i,j)\in\mathcal{N}} \left(\mathbf{I}_m(i,j) - \bar{\mathbf{I}}_m(x,y)\right)^2} \tag{7}$$

where $\bar{\mathbf{I}}_m(x,y)$ is the local mean and $N = 5$ in all experiments. A high ϕ_{ir} value, corresponding to a thermally active region, increases the infrared contribution, while a high ϕ_{vis} value, corresponding to a visually textured edge, shifts weight toward the visible channel. This behaviour explicitly counters thermal crossover by allowing the visible stream to compensate whenever the infrared contrast is low. Figure 2 illustrates the complementary role of each modality and motivates the spatially adaptive weighting in (6).

The three output channels of $\mathbf{I}_f$ are assembled by applying the fused intensity to the luminance channel and copying the a^* and b^* channels from the visible frame in CIELAB colour space, after which the result is converted back to RGB. This preserves colour information that assists the detector in discriminating vehicle types and background regions while keeping the luminance channel thermal-aware. The fused RGB representation forms the direct input to the EfficientNet-B3 backbone and subsequent BiFPN aggregation layers. During end-to-end training, the detection loss is propagated through the differentiable fusion operations, allowing the fusion module and detection network to be jointly optimised. This unified training process aligns the fused representation with the requirements of downstream object localisation and classification. The resulting fused RGB image is directly supplied to the EfficientNet-B3 backbone. During end-to-end training, the detection loss is propagated through the differentiable fusion and detection components, allowing their parameters to be jointly optimised.

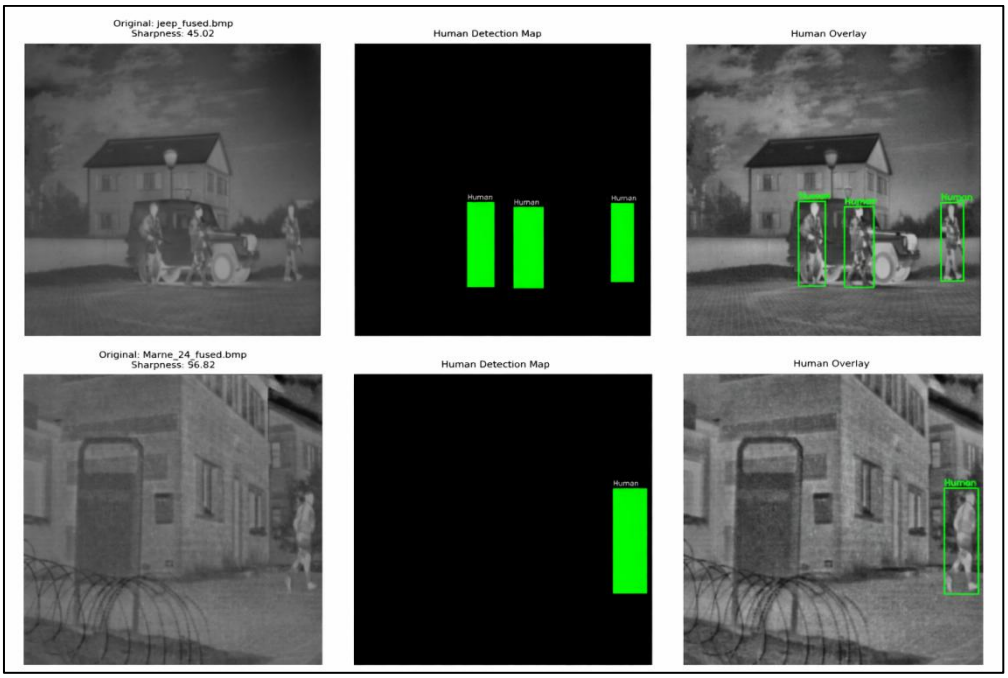


Fig. 2. Visible, infrared, and fused target highlights across three night scenes.

3.4. Bidirectional Feature Pyramid Aggregation

The EfficientNet-B3 backbone extracts feature levels $\{P_3, P_4, P_5, P_6, P_7\}$ with strides {8,16,32,64,128} relative to the input. Multi-scale aggregation is performed through $K = 3$ iterations of the BiFPN, which applies fast normalised weighted fusion.

The top-down pass at level l is

$$\tilde{P}_{td}^{l} = \frac{w_1^l P_{in}^l + w_2^l \text{Resize}\left(P_{in}^{l+1}\right)}{w_1^l + w_2^l + \epsilon} \tag{8}$$

and the bottom-up pass is

$$P_{out}^{l} = \frac{w_3^l P_{in}^l + w_4^l \tilde{P}_{td}^l + w_5^l \text{Resize}(P_{out}^{l-1})}{w_3^l + w_4^l + w_5^l + \epsilon} \tag{9}$$

where all weights w_k^l are non-negative and learnable after passing through a ReLU. The Resize operator performs nearest-neighbour upsampling or max-pooling depending on direction. The skip connection from P_{in}^l to the bottom-up node (term $w_3^l P_{in}^l$ in equation (9)) preserves input-level information that would otherwise be lost after repeated top-down resampling. The BiFPN fusion weights are initialised to 1 and constrained to remain non-negative through ReLU activation. They are jointly updated with the detector parameters using the detector learning rate of 5×10^{-5}.

3.5. Prediction Heads

Following the BiFPN, class and box prediction heads apply a sequence of four separable convolutional layers with weight sharing across levels. The class head predicts the probability of each of $|\mathcal{C}| = 6$ categories for $A = 9$ anchor configurations per level. The box head predicts four regression offsets relative to the anchor centre and dimensions. Non-maximum suppression with an IoU threshold of 0.50 and a confidence threshold of 0.45 is applied at inference time.

Optimisation Objective

Training uses a combined loss that balances classification and localisation contributions:

$$\mathcal{L}_{total} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{reg}\mathcal{L}_{reg} \tag{10}$$

with $\lambda_{cls} = 1.0$ and $\lambda_{reg} = 1.0$ throughout.

The classification term is a focal loss [32] that down-weights the contribution of easy negative examples:

$$\mathcal{L}_{cls} = -\frac{1}{N_{pos}} \sum_{i} \sum_{j} \left[(1 - p_{ij})^{\gamma} y_{ij} \log(p_{ij}) + p_{ij}^{\gamma} (1 - y_{ij}) \log(1 - p_{ij}) \right] \tag{11}$$

where $\gamma = 2.0$, $y_{ij} \in \{0,1\}$ is the binary class label, p_{ij} is the predicted class probability, and N_{pos} normalises by the number of positive anchors.

The localisation term is the Complete-IoU (CIoU) loss [33], which simultaneously optimises overlap, centre distance, and aspect-ratio consistency:

$$\mathcal{L}_{reg} = 1 - \mathrm{IoU} + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2} + \alpha_v v \tag{12}$$

where $\rho(\cdot,\cdot)$ is the Euclidean distance between predicted and ground-truth centres, c is the diagonal length of the smallest enclosing box, $v = (2/\pi)^2 \left(\arctan(w^{gt}/h^{gt}) - \arctan(w/h) \right)^2$ measures aspect-ratio consistency, and $\alpha_v = v/\left((1 - \mathrm{IoU}) + v \right)$.

3.6. Evaluation Metrics

Precision, recall, and F1-score are computed as

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{13}$$

$$F_1 = 2 \cdot \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{14}$$

Mean Average Precision at IoU threshold τ is defined as

$$\mathrm{mAP@}\tau = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \int_0^1 p_c(r)\, dr \tag{15}$$

where $p_c(r)$ is the interpolated precision of class c at recall r. The aggregate metric mAP@0.5:0.95 averages over $\tau \in \{0.50, 0.55, \ldots, 0.95\}$ at step 0.05. Frames per second (FPS) is measured as the mean throughput over the entire test set excluding data loading, using batch size 1 to reflect single-stream deployment conditions.

4. Experimental Setup

The section describes the dataset, division of the dataset, settings under which the implementation is done, and the baseline mechanisms applied to assess the proposed EFID. Each experiment was conducted under the same conditions in order to make a fair and reproducible comparison of methods.

4.1. Dataset

The public Multi-scenario Multi-modality Fusion and Detection (M³FD) benchmark [1] is used for all experiments, and its profile is summarised in Table 1. The benchmark provides 4,200 aligned infrared and visible image pairs at 1024×768 resolution, covering four real-world scenarios: Daytime, Overcast, Night, and Challenge. The infrared frames are recorded in the longwave thermal band (8–14 μm) and are warped to match the resolution of the visible image before they can be fused. It contains annotations on six object types: People, Car, Bus, Motorcycle, Truck, and Lamp and has 33,603 per-instance bounding boxes, and therefore, it is one of the largest publicly available space-version object detector benchmarks based on multimodal fusion. The variety of scenarios and categories makes M³FD especially appropriate to test the robustness of night-time detection and generalisation to other scenarios.

Table 1. Dataset profile of the M³FD benchmark.

Attribute	Value
Dataset name	Multi-scenario Multi-modality Fusion and Detection (M³FD)
Modalities	Registered visible spectrum and infrared (paired)
No. of pairs	4,200 aligned pairs
Visible resolution	1024×768
Infrared resolution	640×512 (warped to 1024×768 for alignment)
Infrared band	8–14 μm (longwave thermal infrared)
Scenarios	Daytime, Overcast, Night, Challenge
No. of classes	6 (People, Car, Bus, Motorcycle, Truck, Lamp)
Total annotations	33,603 bounding boxes
Annotation type	Per-instance axis-aligned bounding boxes

4.2. Data Split and Instance Distribution

The 4200 pairs of images were divided into training, validation and test sets in a 70/15/15 proportion. Scenario stratification ensures that each condition of illumination exists in proportion in each split. Table 2 presents the instances of the three subsets by the classes. There are 2,940 pairs in the training set and 30,490 annotations, which are sufficient to train the model to learn fusion in all six categories. Each of the validation and test sets has 630 pairs each and their class distribution is very similar to that of the training set to facilitate effective generalisation testing. People and Car are the most common with 17,900 and 14,850 annotations respectively, and Lamp is the least common with 1,702 annotations, which is realistic in outdoor surveillance. The issue of class imbalance is met through focal loss, a training method that does not involve oversampling or class-weighted sampling, thus the evaluation of the trained model represents the real-world deployment scenario.

Table 2. Train, validation, and test split with class-wise instance totals (total = 33,603 annotations).

Split	Pairs	People	Car	Bus	Moto	Truck	Lamp
Train (70%)	2,940	12,550	10,420	1,420	2,610	2,310	1,180
Val (15%)	630	2,680	2,230	300	560	510	260
Test (15%)	630	2,670	2,200	310	580	520	262
Total	4,200	17,900	14,850	2,030	3,750	3,340	1,702

4.3. Implementation Details

All the models are run on PyTorch 2.1 and trained using an NVIDIA RTX 3080 with 10GB VRAM. Each method had its input resolution set to 896×896, and the EfficientNet-B3 backbone was initialized using ImageNet-pretrained weights. Our Adam optimiser takes learning rates in the fusion module and the detector head individually. Learning rate is reduced by 0.9 every ten epochs, and a cosine anneal is applied over the last twenty epochs to facilitate gradual convergence. Our data-augmentation pipeline includes horizontal flips, affine transformations, translations, paired mosaic, and colour jitter which are only applied to the visible channel to preserve the radiometric consistency of the infrared channel. In inference, we use a confidence threshold of 0.45 and non-maximum suppression IoU threshold of 0.50 on all methods. Table 3 below defines all implementation settings; the settings are the same across all comparison methods to provide a fair and reproducible assessment.

Table 3. Implementation settings applied uniformly across all methods.

Setting	Value
Framework	PyTorch 2.1
Input resolution	896×896
Batch size	8
Epochs	120
Optimiser	Adam (β1=0.9, β2=0.999)
Initial LR (fusion)	1×10−4
Initial LR (detector)	5×10−5
LR schedule	Decay by 0.9 every 10 epochs; cosine annealing in final 20 epochs
Weight decay	5×10−4
Augmentation	Horizontal flip (p=0.5), affine (±15°), translation (up to 10%), paired mosaic, colour jitter (visible only)
Confidence threshold	0.45
NMS IoU threshold	0.50
Backbone pre-training	ImageNet (EfficientNet-B3)
Hardware (training)	NVIDIA RTX 3080 (10 GB VRAM)
Hardware (FPS)	NVIDIA RTX 3080 (batch size 1)

4.4. Baseline Methods

Five baseline methods are selected to cover a broad range of detection strategies, from single-modality detection to fusion-based approaches, as summarised in Table 4. YOLOv8-L operating on infrared-only and visible-only inputs serves as the single-modality reference for each sensor stream. Faster R-CNN combined with DenseFuse represents a two-stage detection approach with a perceptual fusion front-end. EfficientDet D3 on visible input provides a direct compound-scaled baseline without any fusion benefit. The illumination-aware fusion combined with YOLOv8-L represents the strongest prior fuse-then-detect approach. All baselines are retrained or fine-tuned from their publicly released weights under the settings in Table 3 to ensure a fair and controlled comparison with the proposed EFID.

Table 4. Baseline methods compared under identical experimental conditions.

Method	Input	Key idea
YOLOv8-L (IR only)	Infrared	Single-stage YOLO on thermal input
YOLOv8-L (visible only)	Visible	Single-stage YOLO on visible input
Faster R-CNN + DenseFuse	Fused	Two-stage detector on DenseFuse output
EfficientDet D3 (vis. only)	Visible	Compound-scaled baseline without fusion
Illum.-aware fusion + YOLOv8	Fused	PIAFusion front-end with YOLOv8-L
Proposed EFID	Fused	Attention-guided fusion + EfficientDet D3 with BiFPN

Table 5. Overall detection performance on the test split (values in %). Bold denotes best; FPS is measured at batch size 1 on RTX 3080.

Method	mAP@0.5	mAP@0.5:0.95	Prec.	Rec.	F1	FPS
YOLOv8-L (IR only)	78.6	52.7	80.1	75.8	77.9	61.5
YOLOv8-L (vis. only)	72.4	47.9	74.3	69.2	71.7	62.8
Faster R-CNN + DenseFuse	82.1	56.3	83.4	80.2	81.8	18.7
EfficientDet D3 (vis.)	77.9	53.1	79.0	76.1	77.5	39.8
Illum.-aware + YOLOv8	88.3	62.4	90.2	86.0	88.0	33.5
Proposed EFID	**91.7**	**67.3**	**93.4**	**89.8**	**91.5**	34.2

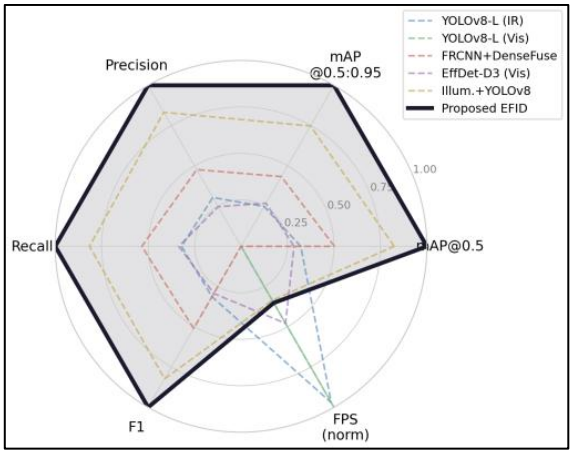


Fig. 3. Multi-metric radar comparison of all methods across six normalised evaluation dimensions.

4.5. Overall Detection Performance

Table 5 reports overall detection results on the test split. The proposed EFID achieves 91.7% mAP@0.5 and 67.3% mAP@0.5:0.95, with 93.4% precision and 89.8% recall at 34.2 FPS as shown in Figure 3. The proposed method (as compared to the best baseline (Illumination-aware fusion + YOLOv8)) increases mAP@0.5 by 3.4 and mAP@0.5:0.95 by 4.9 percentage points while retaining almost the same FPS (34.2 vs. 33.5). This gain indicates that attention-guided fusion combined with the BiFPN head produces feature representations better suited for detection than PIAFusion does. Faster R-CNN + DenseFuse has the lowest FPS (18.7), which makes it inappropriate for continuous streaming even though its mAP at 0.5 is competitive. Single-modality visible-only detectors are the most affected in terms of accuracy, which validates the requirement of infrared data during multimodal night surveillance.

4.6. Per-Class Average Precision

Table 6 reports per-class AP@0.5 on the test split. The Lamp category has the lowest AP of all methods, as can be expected, since its instances are small and often appear near image boundaries, where less BiFPN context is available. Fusion is most advantageous to People and Car due to their high thermal contrast to the background asphalt and vegetation. Motorcycle shows the largest relative improvement over single-IR detection (90.1% vs. 74.2%), suggesting

that small two-wheeled vehicles benefit substantially from the visible structural edges introduced by the activity-weighted pixel fusion step as shown in Figure 4. Figure 5 provides qualitative examples confirming that the detector produces tight bounding boxes under low sharpness and in cluttered backgrounds, consistent with the quantitative recall gains for the People category.

Table 6. Per-class average precision at IoU 0.5 (%) on the test split.

Method	People	Car	Bus	Moto	Truck	Lamp
YOLOv8-L (IR only)	79.8	83.1	77.6	74.2	76.9	80.0
YOLOv8-L (vis. only)	73.9	76.8	70.4	69.3	71.7	72.3
Faster R-CNN + DenseFuse	84.6	86.9	81.0	78.4	80.7	80.8
EfficientDet D3 (vis.)	78.1	82.3	75.8	72.6	76.5	82.0
Illum.-aware + YOLOv8	89.6	92.1	87.3	85.7	86.8	88.3
Proposed EFID	**93.2**	**95.7**	**91.6**	**90.1**	**91.9**	**87.7**

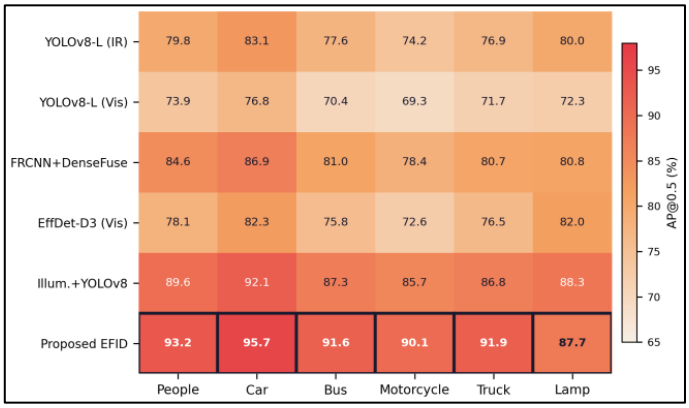


Fig. 4. Per-class AP@0.5 heatmap; bold border highlights the proposed EFID row.

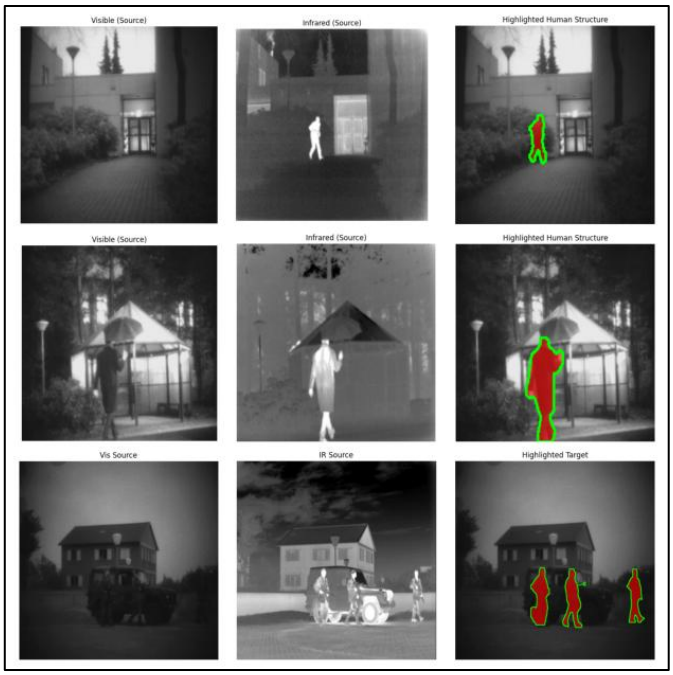


Fig. 5. Qualitative detection results: fused image, detection map, and bounding-box overlay.

4.7. Confidence–Recall Trade-off Analysis

Table 7 reports precision and recall at three confidence thresholds for the proposed method, illustrating the operational flexibility of EFID.

Table 7. Proposed EFID precision and recall at varying confidence thresholds (mAP@0.5 computed over the full threshold sweep; all values in %).

Confidence Threshold	Precision	Recall	F1
0.30	88.7	93.5	91.0
0.45	93.4	89.8	91.5
0.60	95.8	83.1	89.0

The confidence threshold can also be reduced to 0.30 (accepting a decrease in precision to 88.7) by operators that need higher recall (e.g., maximum-sensitivity perimeter alarm). An operator requiring minimal false alarms (e.g., automated access control) can increase it to 0.60, obtaining 95.8 percent precision at the cost of recall.

4.8 Scenario-Wise Robustness

Table 8 reports mAP@0.5 broken down across the four M³FD scenarios. The largest margin over the best competing method is seen in the Daytime scenario (+3.8 pp), with a much smaller margin in the Challenge scenario (+2.6 pp). The drop of 3.2 pp of the proposed method compared to the drop of 8.8 pp of the visible-only YOLOv8-L (75.6 to 66.8) in the Night scenario confirms that it is specifically the infrared evidence used in the fusion that reduces the illumination-induced accuracy gap as shown in Figure 6. The Challenge scenario involves foggy and mixed artificial light; the activity-weighted pixel fusion term is useful in this regard since it is based on local contrast, but not the absolute levels of intensity.

Table 8. Scenario-wise mAP@0.5 (%) on the M³FD test split.

Method	Day	Overcast	Night	Challenge
YOLOv8-L (IR only)	80.6	78.9	79.1	74.0
YOLOv8-L (vis. only)	75.6	71.8	66.8	63.4
Faster R-CNN + DenseFuse	84.0	82.3	80.1	79.0
EfficientDet D3 (vis.)	80.1	78.2	74.6	72.7
Illum.-aware + YOLOv8	89.6	88.2	87.1	86.3
Proposed EFID	**93.4**	**92.1**	**90.2**	**88.9**

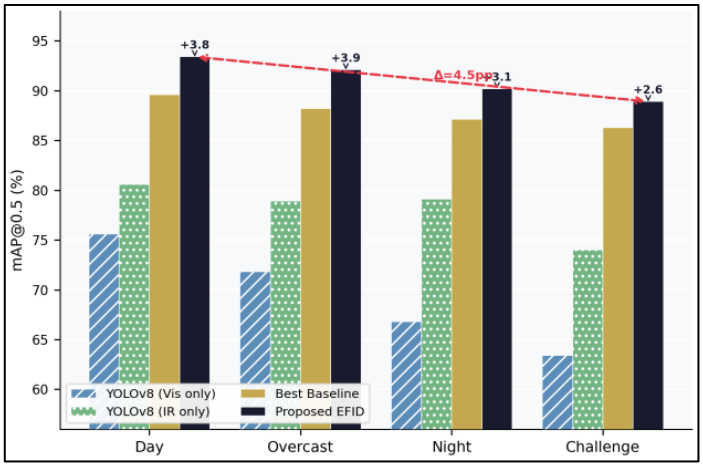


Fig. 6. Scenario-wise mAP@0.5 with per-scenario margin gains of EFID over the best baseline.

4.9. Ablation Study

Table 9 measures the independent contribution of every EFID component. Relative to a visible-only EfficientDet D3 baseline, adding pixel-domain fusion without attention increases mAP@0.5 by 6.1 pp. Adding the attention module increases mAP at 0.5 by 3.6 pp, at the price of a small 1.3X reduction in FPS. Gains of 2.2 and 1.9 pp respectively are obtained by increasing BiFPN operations to two and three respectively with a minimal cost in FPS at each step as shown in Figure 7. For a controlled comparison, only the evaluated component or BiFPN iteration count was varied, while the dataset split, backbone, input resolution, training settings, and evaluation protocol remained unchanged.

Table 9. Ablation study on the test split (all values in %). BiFPN column indicates number of aggregation iterations.

Configuration	BiFPN	mAP@0.5	F1	FPS
EfficientDet D3 (vis. only)	3	77.9	77.5	39.8
Pixel fusion (no attention)	3	84.0	83.5	36.9
Attention-guided fusion	1	87.6	87.2	35.6
Attention + BiFPN (2)	2	89.8	89.4	34.8
Full EFID (BiFPN 3)	3	**91.7**	**91.5**	34.2

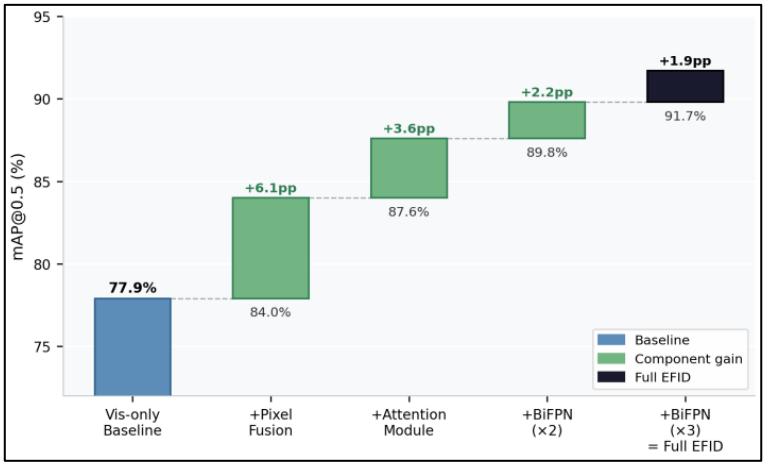


Fig. 7. Incremental mAP@0.5 gains from progressive addition of each EFID component.

4.10. Activity-Weighted Pixel Fusion Analysis:

To isolate the contribution of the activity-weighted pixel fusion term from the activity blending, Table 10 compares three pixel-domain fusion variants within the full pipeline (attention + BiFPN 3).

Table 10. Comparison of pixel-domain fusion variants within the complete detection pipeline. (attention enabled, BiFPN 3 iterations; values in %).

Pixel-domain fusion variant	mAP@0.5	Recall	F1
Simple average (α=β=0.5)	89.3	87.4	88.3
Activity-weighted blend (eq. 6–7)	90.8	88.9	89.8
Proposed activity-weighted pixel fusion	**91.7**	**89.8**	**91.5**

Each step in the pixel-domain fusion chain provides a measurable benefit. The activity-weighted blend adds 1.5 pp over the simple average, confirming that spatially adaptive weighting is important. The activity-weighted pixel fusion term adds a further 0.9 pp and 0.9 pp in recall, which corresponds to improved boundary-level recall on small and partially occluded targets.

4.11. Computational Profile

Table 11 summarises the computational footprint of EFID alongside the two strongest baselines. All models are measured at batch 1 on the same RTX 3080 without TensorRT optimisation.

Table 11. Computational comparison of the proposed EFID and strongest baselines.

Model	Params (M)	GFLOPs	FPS	VRAM (GB)
Faster R-CNN + DenseFuse	41.7	22.6	18.7	8.1
Illum.-aware + YOLOv8-L	43.6	19.3	33.5	7.4
Proposed EFID	28.4	12.4	34.2	6.2

EFID provides a favourable accuracy–efficiency balance, using 28.4 M parameters, 12.4 GFLOPs, and 6.2 GB of VRAM while achieving the highest mAP@0.5 among the compared methods. Relative to Faster R-CNN with DenseFuse, it reduces the parameter count by 32% and computational cost by 45%. Its throughput of 34.2 FPS is comparable to the YOLOv8-L-based fusion pipeline and supports real-time surveillance applications.

4.12. Cross-Dataset Generalisation

Table 12 reports transfer behaviour when EFID is trained on M³FD and evaluated on a separate target dataset with a different sensor configuration. Zero-shot transfer retains 84.1% mAP@0.5 compared with 91.7% on the source test set, corresponding to a 7.6 pp generalisation gap. Fine-tuning with only 10% of the target dataset (126 aligned pairs) recovers 2 pp, reaching 86.0%. This is noteworthy because YOLOv8-L trained on the *full* target set achieves 87.2%, indicating that EFID with minimal adaptation approaches fully supervised performance with a fraction of the target annotation effort. The result suggests that the attention-guided fusion module learns modality-invariant representations that transfer across sensor generations. Cross-dataset evaluation was performed on the specified target dataset without additional fine-tuning or test-time adaptation.

Table 12. Cross-dataset generalisation results (mAP@0.5 in %). The target dataset uses a different sensor pair from the source; all categories except Lamp are present in both.

Setting	mAP@0.5
Proposed EFID trained on M³FD, tested on target (zero-shot)	84.1
Proposed EFID with 10% target fine-tuning (126 paired frames)	86.0
YOLOv8-L trained directly on full target dataset	87.2

5. Discussion

This section interprets the quantitative results, contextualizes the proposed EFID within practical surveillance deployments, and identifies the residual limitations that motivate future research directions.

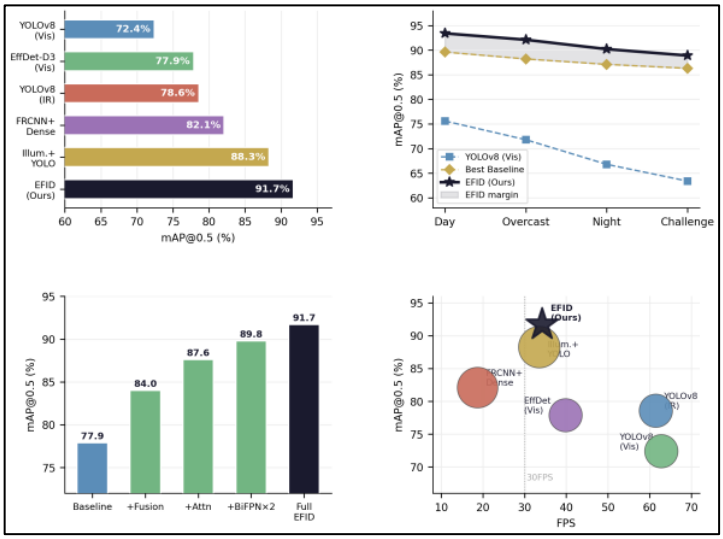


Fig. 8. Consolidated summary: (a) overall ranking, (b) scenario robustness, (c) ablation gains, and (d) accuracy–efficiency trade-off.

5.1. Performance Interpretation

The quantitative evidence presented across the result tables supports three principal observations regarding the design choices of the proposed EFID. The comparison between EfficientDet D3 on visible input alone (77.9%) and Faster R-CNN combined with DenseFuse (82.1%) confirms that fusion provides a measurable benefit over single-modality detection; however, the substantially larger gap to the proposed EFID (91.7%) demonstrates that detection-oriented fusion with attention-based modality weighting provides far greater benefit than appending a perceptual fusion front-end to a standard detector, which aligns with the argument in the literature [15, 16] that fusion and detection objectives must be jointly optimised rather than treated as independent stages. The gradient ablation in Table 10 further shows that the recall improvement from adding the activity-weighted pixel fusion term (89.8 vs. 88.9%) corresponds to improved boundary-level evidence for small categories such as Motorcycle and Lamp, whose fine bounding edges are more strongly represented in the visible channel and would otherwise be lost during activity-weighted blending. The scenario-wise results confirm that Night and Challenge conditions benefit most from fusion, with the proposed EFID degrading by only 4.5 percentage points from Day to Challenge (93.4 to 88.9%), compared with a 12.2 percentage point drop for the visible-only YOLOv8-L (75.6 to 63.4%), demonstrating that the attention module effectively up-weights infrared contributions when visible contrast is insufficient. As shown in Figure 8, the proposed EFID achieves superior overall performance, strong scenario robustness, consistent ablation gains, and a favorable accuracy–efficiency balance.

5.2. Application Contexts

The combination of 34.2 FPS throughput, 93.4% precision, and 89.8% recall sustained under night conditions positions EFID as a practical solution for several real-world surveillance deployments. In perimeter monitoring, the high recall will ensure missed intrusion events are eliminated and the high precision will reduce operator alert fatigue due to false alarms, which are essential features in around the clock unattended operation. People are observed concurrently with a variety of vehicle types such as Bus, Truck, Motorcycle and Car in the same fused stream, which in convoy-corridor observation saves sensor infrastructure costs over running modality-specific detectors. In industrial facility protection, the Challenge scenario result of 88.9% mAP@0.5 indicates robustness under mixed artificial lighting and smoke conditions that are commonly encountered in manufacturing and energy environments. The cross-dataset fine-tuning result of 86.0% achieved with only 126 labelled target pairs is particularly relevant for operational deployment scenarios where full dataset re-annotation is impractical, as it demonstrates that EFID can be rapidly adapted to a new sensor pair with minimal annotation effort while retaining near-fully-supervised detection performance.

5.3. Limitations

Whereas the suggested EFID proves to be highly effective in terms of detection in all of the considered situations, some of the boundary cases reveal some remaining flaws that should be taken into account. The degree of heavy occlusion compromises the accuracy of the infrared activity measure resulting in fragmented thermal signatures and poor recall of dense cluttered scenes. Cross-sensor calibration drift caused by temperature-induced lens expansion or mechanical vibration introduces sub-pixel misalignment that a fixed homography cannot compensate for, with disproportionate impact on small object categories such as Motorcycle. The absence of explicit illumination state supervision in the attention module limits its ability to respond adaptively during thermal crossover periods, where both modalities simultaneously exhibit reduced discriminability. Furthermore, the current memory footprint of 6.2 GB VRAM constrains deployment to high-end GPU hardware, and post-training quantisation or structured knowledge distillation will be necessary before the pipeline can be reliably executed on embedded surveillance platforms.

6. Conclusion

This paper introduced the EfficientDet Fusion Intrusion Detector (EFID), an attention-guided infrared and visible fusion pipeline for night-time multi-class intrusion detection. The fusion module combines spatially adaptive activity-weighted blending with a visible activity-weighted pixel fusion term, producing a three-channel fused image that preserves infrared thermal salience and visible structural edges simultaneously. An EfficientDet D3 detector with a three-iteration BiFPN head processes the fused representation for scale-consistent multi-class localisation. Experiments on the M³FD benchmark demonstrate 91.7% mAP@0.5 and 67.3% mAP@0.5:0.95, with 93.4% precision and 89.8% recall at 34.2 FPS. Per-class results confirm consistent gains across all six categories, and scenario-wise analysis shows that the Night and Challenge degradations are substantially smaller than those of visible-only or single-IR detectors. Ablation confirms that each component (fusion, attention, activity-weighted pixel fusion, and BiFPN depth) contributes measurably to final performance. Cross-dataset transfer retains 84.1% mAP@0.5 in zero-shot mode and recovers to 86.0% with 10% target fine-tuning, indicating practical data efficiency in new deployment environments. Although EFID achieves 93.4% mAP@0.5 in the Day scenario, its performance decreases to 88.9% in the Challenge scenario, representing a 4.5 percentage-point reduction. This degradation indicates sensitivity to difficult conditions such as fog, mixed illumination, weak thermal contrast, and occlusion. The individual effect of calibration drift was not separately quantified, while the 6.2 GB VRAM requirement continues to restrict embedded deployment. The computational profile and real-time throughput support the practical deployment potential of the proposed framework rather than representing its primary methodological contribution. Future work will explore alignment-aware training with a learned homography module, explicit illumination-state supervision for the attention module, and post-training quantisation to enable EFID on edge hardware without accuracy regression.

All the Declarations and Statements

Author Contributions Statement

Rithick S: Conceptualization, methodology, and supervision. Proposed the research ideas, constructed the overall framework, and supervised the execution of the project.

Jenefa A: Data curation and software implementation. Handled data acquisition, dataset preprocessing, and implementation of the proposed research model.

Abirami M K: Model training, validation, and performance evaluation. Led the model training process, validated the results using standard evaluation metrics, and benchmarked the performance against existing methods.

Sheeba Merlin: Formal analysis, visualization, and statistical analysis. Performed detailed analysis of the experimental

results, prepared performance charts, and ensured the statistical robustness of the evaluation.
Antony Taurshia: Writing, original draft preparation. Drafted the initial manuscript, contributed to the literature survey, and documented the technical background of the study.
Lincy A: Writing, review and editing, and project management. Reviewed and edited the manuscript, improved clarity and coherence, and helped coordinate project milestones and deadlines.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement
The authors declare no conflicts of interest.

Funding Declaration
This research received no external funding.

Data Availability Statement
The results and processed data generated during the study will be made available by the corresponding author upon reasonable request.

Ethical Declarations
This study did not involve any new experiments on human participants or animals.

Acknowledgments
We sincerely thank the experts for their professional evaluation and valuable recommendations, which have contributed to improving the quality of the experiment and the reliability of its results.

Declaration of Generative AI in Scholarly Writing
The authors declare that no generative AI tools or AI-assisted technologies were used in the preparation, writing, editing, or revision of this manuscript.

Abbreviations
The following abbreviations are used in this manuscript:

AI - Artificial Intelligence
AP - Average Precision
BiFPN - Bidirectional Feature Pyramid Network
CIoU - Complete Intersection over Union
CNN - Convolutional Neural Network
DL - Deep Learning
EFID - EfficientDet Fusion Intrusion Detector
FPS - Frames Per Second
GFLOPs - Giga Floating Point Operations
GPU - Graphics Processing Unit
IR - Infrared
IoU - Intersection over Union
LR - Learning Rate
LWIR - Longwave Infrared
mAP - Mean Average Precision
M³FD - Multi-scenario Multi-modality Fusion and Detection
NMS - Non-Maximum Suppression
RGB - Red, Green, and Blue
VRAM - Video Random Access Memory

Appendix A/B/C⋯, with appendix title
None.

References

[1] Liu, Jinyuan, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5792-5801. 2022.

[2] Tan, Mingxing, Ruoming Pang, and Quoc V. Le. "Efficientdet: Scalable and efficient object detection." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10778-10787. 2020.

[3] Tang, Linfeng, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware." Information Fusion 83 (2022): 79-92.doi: 10.1016/j.inffus.2022.03.007.

[4] Roy, Swalpa Kumar, et al. "Multimodal fusion transformer for remote sensing image classification." IEEE Transactions on Geoscience and Remote Sensing 61 (2023): 1-20.doi: 10.1109/TGRS.2023.3286826.

[5] Ma, Rui, Yi Hou, Chenxuan Li, Huizhu Jia, and Xiaodong Xie. "Scene-adaptive unsupervised crowd counting for video surveillance." IEEE Transactions on Circuits and Systems for Video Technology (2025).doi: 10.1109/TCSVT.2025.3540850

[6] Ni, Peizhou, Weiming Hu, Jinchao Hu, Xiong Chen, and Yan Liu. "Illumination robust semantic segmentation based on cross-dimensional multispectral edge fusion in dynamic traffic scenes." IEEE Access 12 (2024): 171589-171600.doi: 10.1109/ACCESS.2024.3498896.

[7] Santhiya, P., Immanuel JohnRaja Jebadurai, Getzi Jeba Leelipushpam Paulraj, and E. Donald Jawahar. "Implementing ViT Models for Traffic Sign Detection in Autonomous Driving Systems." In 2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST), pp. 382-387. IEEE, 2024.

[8] Yang, Sen, Ze Feng, Zhicheng Wang, Yanjie Li, Shoukui Zhang, Zhibin Quan, Shu-tao Xia, and Wankou Yang. "Detecting and grouping keypoints for multi-person pose estimation using instance-aware attention." Pattern recognition 136 (2023): 109232.doi: 10.1016/j.patcog.2022.109232.

[9] Yang, Guang, Jie Li, Hanxiao Lei, and Xinbo Gao. "A multi-scale information integration framework for infrared and visible image fusion." Neurocomputing 600 (2024): 128116.doi: 10.1016/j.neucom.2024.128116.

[10] Taurshia, A., Kuriakose, B. M., Vijaya Kumar, E. N., & Lincy, A. (2025). Advancing Road Maintenance with EfficientDet-based Pothole Monitoring. Serbian Journal of Electrical Engineering, 22(1), 57.doi: 10.2298/SJEE2501057J

[11] Li, Hui, and Xiao-Jun Wu. "DenseFuse: A fusion approach to infrared and visible images." IEEE transactions on image processing 28.5 (2019): 2614-2623.doi: 10.1109/TIP.2018.2887342.

[12] Li, Hui, Xiao-Jun Wu, and Josef Kittler. "RFN-Nest: An end-to-end residual fusion network for infrared and visible images." Information Fusion 73 (2021): 72-86.doi: 10.1016/j.inffus.2021.02.023.

[13] Ma, Jiayi, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion." IEEE Transactions on Image Processing 29 (2020): 4980-4995.doi: 10.1109/TIP.2020.2977573.

[14] Xu, Han, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. "Fusiondn: A unified densely connected network for image fusion." In Proceedings of the AAAI conference on artificial intelligence, vol. 34, no. 07, pp. 12484-12491. 2020.

[15] Santhiya, P., Jebadurai, I. J., Paulraj, G. J. L., & Karan, S. K. (2024, June). Deep Vision: Lane Detection in ITS: A Deep Learning Segmentation Perspective. In 2024 Second International Conference on Inventive Computing and Informatics (ICICI) (pp. 21-26). IEEE.

[16] Xiang, Xiantai, Guangyao Zhou, Ben Niu, Zongxu Pan, Lijia Huang, Wenshuai Li, Zixiao Wen, Jiamin Qi, and Wanxin Gao. "Infrared-visible image fusion meets object detection: Towards unified optimization for multimodal perception." Remote Sensing 17, no. 21 (2025): 3637.doi: 10.3390/rs17213637.

[17] Gallagher, James E., and Edward J. Oughton. "Surveying you only look once (YOLO) multispectral object detection advancements, applications, and challenges." IEEE Access 13 (2025): 7366-7395.doi: 10.1109/ACCESS.2025.3526458.

[18] Wang, Qinghua, et al. "DICFusion: Infrared and Visible Image Fusion via a Deep Integrated and Semantic-Coordinated Network." IEEE Transactions on Circuits and Systems for Video Technology (2025).doi: 10.1109/TCSVT.2025.3625990

[19] Madhasu, Nithya, and Sagar Dhanraj Pande. "Revolutionizing wildlife protection: A novel approach combining deep learning and night-time surveillance." Multimedia Tools and Applications 84.21 (2025): 23465-23499.doi: 10.1007/s11042-024-19876-4.

[20] Joshua Samuel, Giri Balan, and Joshua Premkumar. "Enhancing public safety through license plate recognition for counter terrorism through deep learning technique." In 2023 4th International Conference on Signal Processing and Communication (ICSPC), pp. 96-100. IEEE, 2023.

[21] Tu, Haiyan, Xiaoyue Tan, Zhongping Yin, Xuegang Zhang, Kang Yang, Zhengkun Qiu, and Xiujuan Zheng. "Non-intrusive fatigue detection based on multidomain features of sitting pressure and machine learning." IEEE Sensors Journal (2025).doi: 10.1109/JSEN.2025.3544299.

[22] Vasantharaj, A., J. Immanuel Jebaraj, K. Kavinraj, and S. Sebin. "An Intelligent Deep Learning Approach for Wild Animal Detection and Alert System Using YOLOv11." In 2025 10th International Conference on Communication and Electronics Systems (ICCES), pp. 1519-1526. IEEE, 2025.

[23] Wang, Jiali, Zhengyu Xie, Yong Qin, Xiaoqiang Zhang, Zengqing Wang, and Li Wang. "Boundary Dynamic Perception Detection Network for Low Light Railway Environment." IEEE Sensors Journal (2025).doi: 10.1109/JSEN.2025.3543860.

[24] Rocha, Pedro Daniel, Fernando JP Lopes, and Luís A. Da Silva Cruz. "Automating Electrical Grid Asset Inspection: From Current Challenges to Future Directions." IEEE Access 13 (2025): 201392-201438.doi: 10.1109/ACCESS.2025.3636718.

[25] Aman, Serge Stéphane, Behou Gérard N'guessan, Kouadio Prosper Kimou, and Tiemoman Kone. "AHOD: adaptive hybrid object detector for context-aware and real-time object detection in complex environments." Discover Applied Sciences 7, no. 11 (2025): 1254.doi: 10.1007/s42452-025-07784-7

[26] Ande, A., Mounikuttan, T., Anuj, M. D., Makros, G. J., Rejoice, G. R., & Shalini, T. M. (2023, March). Real-time rail safety: A deep convolutional neural network approach for obstacle detection on tracks. In 2023 4th International Conference on Signal Processing and Communication (ICSPC) (pp. 101-105). IEEE.

[27] Rampriya, R. S., Sahaya Beni Prathiba, L. Jerart Julus, Arikumar K. Selvaraj, and Joel JPC Rodrigues. "Fault detection and semantic segmentation on railway track using deep fusion model." IEEE Access 12 (2024): 136183-136201.doi: 10.1109/ACCESS.2024.3462822.
[28] Santhiya, P., Immanuel Johnraja Jebadurai and Getzi Jeba Leelipushpam Paulraj, "BIPOOLNET: An advanced UNet architecture for enhanced lane detection in autonomous vehicles." Intell. Decis. Technol. 18, no. 2 (2024): 743-757.doi: 10.3233/IDT-240162.
[29] Isaac J, Santhiya P, Archpaul J. HearTheSigns: AI-Based Road Sign Detection with Real-Time Voice Feedback for Blind. In2025 International Conference on Computing Technologies (ICOCT) 2025 Jun 13 (pp. 1-6). IEEE.
[30] RS, R., Al-Shehari, T., Nathan, S., A, J., R, S., P, S.P., Alfakih, T. and Alsalman, H., 2024. An unmanned aerial vehicle captured dataset for railroad segmentation and obstacle detection. Scientific Data, 11(1), p.1315.doi: 10.1038/s41597-024-03952-3.
[31] J. Hu, L. Shen, G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE CVPR, 2018, pp. 7132-7141.
[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, "Focal loss for dense object detection," in Proc. IEEE ICCV, 2017, pp. 2980-2988.
[33] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in Proc. AAAI, vol. 34, no. 7, pp. 12993-13000, 2020.

Authors' Profiles



Rithick S. has completed a Bachelor of Technology degree in Artificial Intelligence and Machine Learning. His academic interests include intelligent systems, machine learning, and emerging AI technologies. He is interested in applying AI-based methods to practical and real-world problems.



Dr. A. Jenefa is an Associate Professor in the Division of Data Science and Cyber Security at Karunya Institute of Technology and Sciences. She completed her Ph.D. at Anna University, Chennai, in 2022. Her research interests include artificial intelligence, computer networks, machine learning-based diagnostics, network traffic classification, and AI for public safety. She has published in international journals and conferences and has received recognition for her teaching and research.



Abirami M.K. serves as an Assistant Professor with academic and research interests in emerging computing technologies. Her research focuses on Artificial Intelligence, Machine Learning, Agentic AI, and Quantum Cryptography. She is particularly interested in developing intelligent and secure computational solutions for real-world applications.



Dr. G. Sheeba Merlin serves as an Assistant Professor of Mathematics at Karunya Institute of Technology and Sciences, Coimbatore, with over 12 years of experience in teaching and research. Her research interests include Algebra, Graph Theory, and Fixed-Point Theory, with a strong publication record in Scopus and Web of Science indexed journals.



Antony Taurshia is an Assistant Professor in the Department of Data Science and Cyber Security at Karunya Institute of Technology and Sciences. She is pursuing her Ph.D. and holds an M.Tech and B.E. Her research interests include cybersecurity, IoT security, and software-defined network-based security solutions.



Lincy A. is an Assistant Professor in the Department of Computer Science and Engineering at National Engineering College, Tamil Nadu. She holds bachelor's and master's degrees in computer science and is pursuing her Ph.D. under Anna University, Chennai. Her research interests include deep learning, generative AI, and explainable AI for healthcare**.**

How to cite this paper

Rithick S., Jenefa A., Abirami M. K., G. Sheeba Merlin, Antony Taurshia, Lincy A., "Spatially Adaptive Infrared–visible Fusion with EfficientDet for Night-time Multi-Class Intrusion Detection", International Journal of Engineering and Manufacturing (IJEM), Vol.16, No.5, pp. 312-329, 2026. DOI:10.5815/ijem.2026.05.17