ITC 2/54
Information Technology and Control
Vol. 54 / No. 2/ 2025
pp. 365-379
DOI 10.5755/j01.itc.54.2.39803

ADFN: Adaptive Dynamic Fusion Network for Real-time Multispectral Object Detection

Received 2024/12/13 | Accepted after revision 2025/02/05

HOW TO CITE: Yang, L., Qiao, G. (2025). ADFN: Adaptive Dynamic Fusion Network for Real-time Multispectral Object Detection. *Information Technology and Control*, 54(2), 365-379. https://doi.org/10.5755/j01.itc.54.2.39803

ADFN: Adaptive Dynamic Fusion Network for Real-time Multispectral Object Detection

Lin Yang, Gangzhu Qiao*
School of Computer Science and Technology, North University of China, Taiyuan, China

Corresponding author: qiaogz@nuc.edu.cn

Multispectral object detection leverages the complementary strengths of infrared (IR) and visible (VIS) modalities to improve detection accuracy. However, existing approaches often lack adaptability to dynamic lighting conditions, or fail to achieve real-time performance due to complexity. We propose the Adaptive Dynamic Fusion Network (ADFN), a novel architecture that integrates adaptive multi-path computation and attention-guided feature fusion to address these challenges. ADFN incorporates the Collaborative and Alternating Attention (CAA) modules for efficient feature alignment and the Adaptive Dynamic Pathway (ADP) strategy to dynamically adjust computational pathways based on lighting conditions, optimizing the balance between accuracy and efficiency. Experiments on the FLIR2 and LLVIP datasets demonstrate that ADFN achieves superior mAP@50-95 and real-time performance, showcasing its robustness and efficiency across diverse environments. ADFN offers a practical solution for dynamic lighting conditions and resource-constrained multispectral object detection tasks.

KEYWORDS: Multispectral Object Detection, Real-time, Feature Fusion, Adaptive Multi-path

1. Introduction

In recent years, multispectral imaging has become a critical tool for object detection in fields such as surveillance, autonomous driving, and security systems [9, 20]. By combining information from both infrared (IR) and visible (VIS) spectra, multispectral detection systems leverage the strengths of each modality—IR's capacity to detect thermal differences and VIS light's detailed texture information—to achieve a more comprehensive detection framework [3, 5, 24, 33]. This integration is particularly valuable in challenging environments or under varying lighting conditions, such as daytime, dusk, or night time, where either modality alone may be insufficient for accurate object identification [22, 25].

Challenges in Robustness and Adaptability. While multispectral object detection holds significant promise, ensuring robustness and adaptability in multispectral object detection remains a significant challenge. Lighting conditions—ranging from strong illumination to complete darkness—introduce unique difficulties for maintaining detection accuracy. Current models often employ straightforward IR and VIS feature fusion strategies that lack adaptability, potentially leading to erroneous predictions when both modalities are equally weighted in all settings [10, 24, 32]. This fixed-weight approach may overlook the need for adaptive weighting according to environmental context, risking performance degradation in fluctuating conditions.

Lighting variations are among the most critical factors influencing the performance of multispectral detection systems. In real-world applications such as surveillance, autonomoueqs driving, and security, lighting conditions can vary drastically within a short period due to natural factors (e.g., day-to-night transitions, shadows, or glare) or artificial influences (e.g., headlight reflections, urban lighting). Such variations directly affect the visibility of objects in the VIS spectrum and the thermal contrast captured by IR sensors. Unlike other challenges, such as sensor noise or occlusion, lighting changes occur frequently and unpredictably, making it imperative for detection systems to adapt dynamically. Figure 1 illustrates pairs of IR and VIS images under different conditions, underscoring the need to adjust the contribution of IR and VIS features dynamically to ensure effective detection.

Real-time Performance and Fusion Complexity. In real-world applications, achieving real-time performance in multispectral object detection is often constrained by hardware limitations, especially in resource-limited environments. Although certain models are optimized for speed, their performance can suffer under restricted computational resources, leading to increased latency [29, 30]. Additionally, many existing fusion techniques introduce complex and often redundant processes that increase model complexity and hinder processing speed. These methods frequently rely on intricate cross-modality fusion operations that add computational burden without proportionally improving detection accuracy [9, 20, 24]. An efficient fusion strategy is therefore crucial for reducing model complexity while enabling robust, real-time detection across varied environments.

Figure 1

VIS-IR paired examples from FLIR2&LLVIP. The first row shows ideal lighting conditions with clear VIS and IR features. The second row illustrates intense backlighting, where VIS information is degraded despite ample light, making IR features crucial for accurate detection. The third row depicts low-light conditions where VIS features are minimal, requiring strong reliance on IR.

Motivation for ADFN. Addressing these challenges, this paper introduces the Adaptive Dynamic Fusion Network, designed to improve detection robustness and efficiency across varying lighting conditions. By dynamically balancing IR and VIS contributions, ADFN offers a solution that bridges the gap between adaptability and computational efficiency. The primary contributions are as follows:

1. We introduce a computationally efficient Cooperative and Alternating Attention (CAA) module for IR and VIS feature fusion, backed by thorough theoretical and experimental analysis.
2. We propose an Adaptive Dynamic Pathway (ADP) strategy that adjusts feature fusion weights based on scene context and on this basis, we have designed a dynamic multi-path network, enhancing detection in varying lighting conditions.

3 A comprehensive evaluation of ADFN's performance, demonstrating significant improvements in both detection accuracy and real-time efficiency, making it well-suited for practical applications even in resource-limited environments.

2. Related Work

2.1 IR and VIS Fusion

The fusion of IR and VIS data is fundamental to multispectral object detection due to the complementary strengths of these modalities: IR is sensitive to thermal variations, while VIS captures spatial detail [9, 20].

Traditional methods, such as early concatenation, integrate IR and VIS features but fail to exploit cross-modal dependencies effectively, limiting their performance in complex environments [2, 16]. Advanced methods like Cross-Modality Fusion Transformers use cross-attention mechanisms to enhance feature integration. However, these approaches are computationally expensive and less suitable for real-time applications [14, 22, 26, 27]. Similarly, attention-based methods have been employed to guide feature fusion, but many lack robust guidance mechanisms and often struggle to achieve high accuracy [14, 29, 30].

2.2 Multispectral Object Detection

Recent advancements in multispectral object detection predominantly rely on well-established backbone networks, while incorporating specialized detection heads such as YOLO and Faster R-CNN to perform object localization and classification [18, 22, 23]. These approaches aim to effectively leverage the complementary features of IR and VIS data.

For instance, some methods build upon YOLOV5 by introducing a dual-stream backbone to separately process IR and VIS features, followed by a Transformer-based cross-modal fusion module [4]. The fused features are then passed through YOLOV5's Neck and Head components for detection, enhancing the network's ability to capture cross-modal dependencies while retaining the simplicity.

Similarly, Faster R-CNN can be adapted for multispectral object detection by integrating cross-modal fusion mechanisms into its feature extraction stages, effectively combining the strengths of its region-based proposals with multimodal feature interactions [18].

These methods demonstrate the versatility of adapting existing object detection frameworks to multispectral data. However, the integration of additional fusion modules, such as Transformers, often increases computational overhead, posing challenges for real-time applications and scenarios with constrained hardware resources. Additionally, due to the complexity of real-world scenes, these methods often fail to adaptively balance the contributions of each modality, resulting in reduced model robustness.

2.3 Adaptivity of Lighting Variations

Many current efforts utilize multi-path networks to achieve adaptivity [2, 16, 17], such as adaptively adjusting multi-path networks for semantic segmentation across different scenes [16]. This adaptability is also crucial for multispectral object detection under varying lighting conditions. Adaptive fusion strategies can adjust the importance of IR and VIS features based on scene context, thereby enhancing detection capabilities across diverse environments [5, 10, 32]. For example, Illumination Attention-Guided Transformers prioritize IR features in dark environments and VIS features in bright conditions to improve system robustness [5]. However, these adaptive methods often come with high computational costs, emphasizing the need for lightweight frameworks that strike a balance between adaptability and efficiency.

Therefore, ADFN addresses this gap by introducing an adaptive dynamic pathway that adjusts feature contributions based on environmental context, enabling robust detection while maintaining computational efficiency.

3.Methodology

3.1 Framework Overview

The proposed Adaptive Dynamic Fusion Network (ADFN) is specifically designed for efficient multispectral object detection across varying environmental conditions. As shown in Figure 2, our network primarily consists of four parts:

Dual-Stream Feature Extraction: Extended from the YOLOV8 framework [12], the Dual-Stream Feature Extraction separates streams for VIS and IR images to extract modality-specific features, producing

feature maps (F_i^V for VIS and F_i^I for IR) that capture both spatial detail and thermal information.

Cooperative and Alternating Attention Module: The module ($\varphi_1, \varphi_2, \varphi_3$) fuses features from the VIS and IR streams by combining Cooperative Self Attention and Alternating Attention mechanisms.

Adaptive Pathway Assessment Module: This module analyzes the input image's lighting condition and generates a lighting variable E (indicating Bright, Dim, or Dark). This variable is used to control the pathway and fusion strategy throughout the network.

Final Detection Output: The fused feature map M_i is passed to the detection head, compatible with YOLOV8's framework, for object detection and classification.

3.2 Feature Fusion

We propose the Cooperative and Alternating Attention (CAA) Module, as shown in Figure 3, which combines Cooperative Self Attention [6] and Alternating Attention mechanisms [21], forming a more efficient and adaptive feature fusion module.

Cooperative Self Attention: This approach significantly enhances feature representation while reducing parameter complexity, addressing limitations in traditional self-attention mechanisms, as shown in Figure 3(a). Traditional attention mechanisms allocate separate weights for queries (W^Q), keys (W^K), and values (W^V) across multiple heads. While effective, this approach incurs high computational costs due to the need for independent parameter sets for each head.

In contrast, Cooperative Self Attention employs shared weights for W^Q and W^K across heads, while retaining unique value matrices (W_n^V) for each head to preserve modality-specific information. This shared weight mechanism reduces redundancy, allowing the model to achieve comparable or superior representation efficiency with fewer parameters. The computation process of Cooperative Self Attention is illustrated in Equation 1:

$$Q = XW^Q, K = XW^K, V_n = XW_n^V$$
$$\text{Head}_n = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V_n \quad (1)$$
$$Z = \text{Concat}(\text{Head}_1, \text{Head}_2, ..., \text{Head}_n)W^O$$

Figure 2

Framework of Adaptive Dynamic Fusion Backbone (ADFB). It processes VIS and IR inputs through parallel branches, with lighting conditions dynamically influencing feature fusion using the Lighting Assessment Module. The CAA modules adaptively integrate features, enhancing detection robustness across varying environmental conditions.

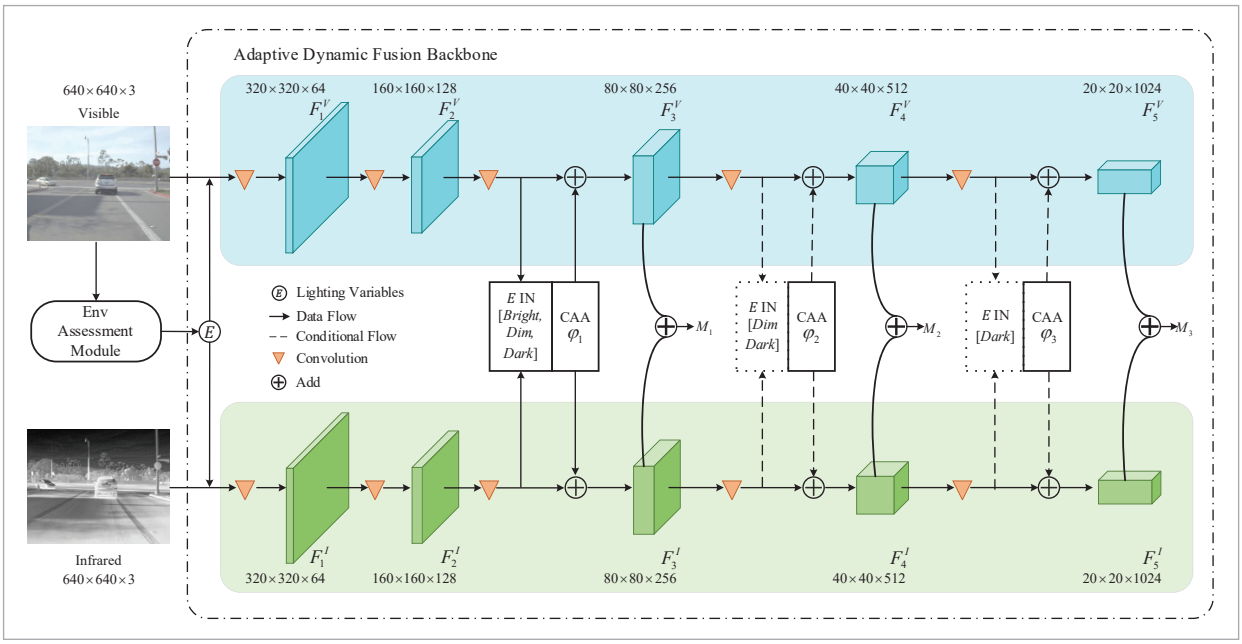
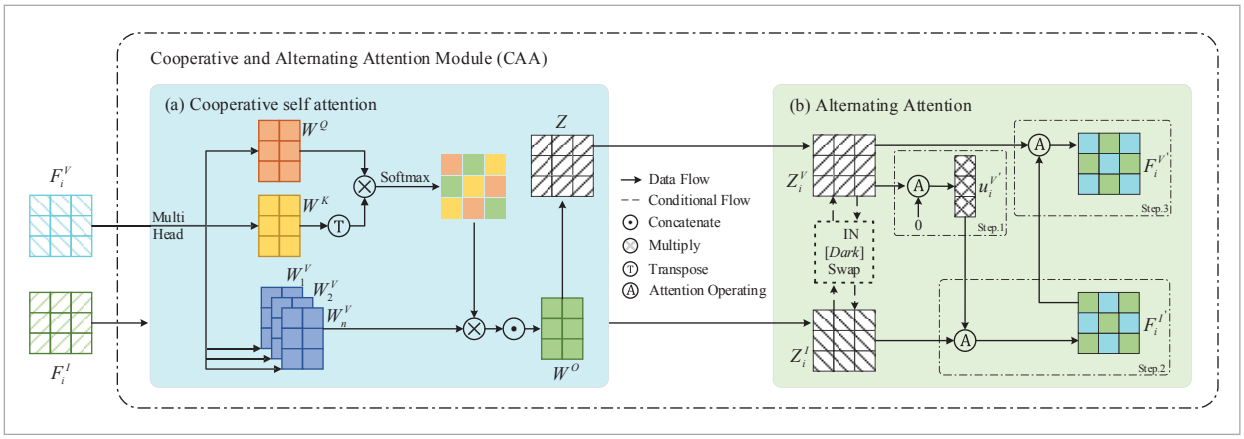


Figure 3

The structure of the CAA Module. It consists of two key components: (a) Cooperative Self-Attention, which enhances feature representation by sharing query and key projections across modalities, and (b) Alternating Attention, which dynamically fuses features by leveraging guidance from either modality based on lighting conditions (E).



Where Q, K, and V_n are derived from the input feature X using the learnaable projection weights W^Q, W^K, and W_n^V, respectively. For each head, the compatibility between Q, and K is calculated, scaled by d_k, and normalized via softmax. The resulting weighted summation with V_n produces Head_n. Finally, the heads are concatenated and projected through W^O to form the output Z, which encapsulates information from all attention heads.

Alternating Attention: This approach is specifically designed to enhance cross-modal feature fusion by leveraging adaptive guidance between VIS and IR modalities. Unlike traditional cross-modal fusion techniques, which often rely on equal contributions or static fusion strategies, Alternating Attention dynamically prioritizes one modality based on the context, ensuring effective interaction.

As shown in Figure 3(b), starting with the attention vectors Z^V and Z^I produced by Cooperative Self Attention, A where represents the Attention operation, expressed as Equation 2:

$$\text{Attention} = \sum_k a_k x_k \quad \text{where } a = \text{Softmax}(W_A^{\mathrm{T}} \bullet \tanh(W_X X + W_g g \bullet 1^{\mathrm{T}}) \bullet \alpha), \tag{2}$$

where X represents the feature matrix (either Z^V or Z^I), encoding data from one modality, g is the guidance vector from the complementary modality, enhancing feature representation, W_X and W_g are learnable parameters that transform X and g for interaction, α is a dynamic weight factor that modulates the guidance influence, detailed in Section 3.3, W_A^{T} is a learnable vector used to generate attention scores. α is the attention score vector used to output the attention vector.

$$\begin{aligned} u &= \text{Attention}(Z^V,\ g = 0, \alpha) \\ F_I' &= \text{Attention}(Z^I,\ g = u, \alpha) \\ F_V' &= \text{Attention}(Z^V,\ g = F_I',\ \alpha) \end{aligned} \tag{3}$$

Equation 3 illustrates the three-step calculation process in the Alternating Attention mechanism:

1. Initial Attention Calculation for VIS Features, where the VIS feature Z^V is computed with a zero guidance vector g, producing an intermediate attention vector u that serves as initial guidance for the IR feature fusion in the next step.
2. Guided Attention for IR Features, where the intermediate attention vector u guides the IR feature Z^I to produce the fused IR feature F_I', allowing the VIS features to influence the IR fusion process.
3. Final Attention Calculation for VIS Features, where the fused IR feature F_I' provides guidance to further refine the VIS feature Z^V, resulting in the final fused VIS feature F_V', effectively embedding the IR information back into the VIS modality.

Additionally, the adaptive dynamic pathway mechanism allows the model to choose the dominant modality based on lighting conditions E. This selection enables the model to adaptively emphasize one modality over the other in specific scenarios (e.g., prioritizing IR features in low-light conditions), as detailed in Section 3.3.

In summary, traditional attention mechanisms treat modalities independently or employ static fusion strategies, which fail to adapt to the dynamic nature of real-world conditions. The CAA module introduces two key innovations: shared-weight Cooperative Self Attention for parameter-efficient feature extraction and Alternating Attention for guided cross-modal fusion. These improvements align each modality's strengths to address varying environmental challenges, ensuring robust and adaptive multispectral feature integration.

3.3 Adaptive Dynamic Pathway

IR and VIS features are highly complementary, but certain scenarios may require reliance on only one modality for accurate detection. Using a uniform fusion strategy in such cases can result in unnecessary computational overhead. To address this, we introduce an Adaptive Dynamic Pathway (ADP) wstrategy, as illustrated by the dashed flows and dashed boxes in Figure 2-3. This strategy incorporates multiple conditional modules to dynamically adjust the data flow and network pathways based on lighting conditions.

Step1: Lighting Condition Assessment.

The ADP strategy begins by evaluating the lighting condition of the input image. We calculate the proportion of "dark" pixels in the grayscale image using the formula:

$$dark\_prop = \frac{dark\_sum}{r \times c}$$
$$\text{where } dark\_sum = \sum_{\forall \text{pixel } p \in \text{gray\_img}} I(p < 40), \quad (4)$$

where r and c represent the dimensions (rows and columns) of the grayscale image (gray_img). We define a threshold value (e.g., 40) to classify dark pixels, and where $I(\cdot)$ is an indicator function that counts pixels with intensity values below the threshold (0–39 gray scale).

Based on the value of *dark_prop*, the lighting condition E is classified into Bright, Dim, or Dark as shown in:

$$E = \begin{cases} \text{Dark} & \text{if } dark\_prop \geq 0.75 \\ \text{Dim} & \text{if } 0.4 \leq dark\_prop < 0.75 \\ \text{Bright} & \text{otherwise} \end{cases} \quad (5)$$

Step 2: Pathway Configuration.

Once classified, the lighting condition E determines the active CAA modules, the activation of the Swap Guide, and the feature fusion weights (α^I for IR and α^V for VIS). These configurations are summarized in Table 1.

Step 3: Dynamic Pathway Execution.

In Bright, only φ_1 is activated, while Swap Guide is disabled. The weights for IR and VIS features are set to $\alpha^I = 0.3$ and $\alpha^V = 0.7$, respectively. This configuration emphasizes VIS features, which are typically more effective in well-lit environments, while still incorporating a small contribution from IR features to handle potential challenges, such as glare or backlit scenarios.

In Dim, both φ_1 and φ_2 are activated, and Swap Guide is enabled to allow IR features to guide VIS feature representation during the fusion process. The weights are adjusted to $\alpha^I = 0.6$ and $\alpha^V = 0.4$, prioritizing thermal information while leveraging VIS features for supplementary details. This setup effectively addresses the reduced visibility and lighting inconsistencies often encountered in dim environments.

In Dark, all three CAA modules (φ_1, φ_2, and φ_3) are activated, enabling a comprehensive fusion process. The weights are heavily biased toward IR features, with $\alpha^I = 0.7$ and $\alpha^V = 0.3$, ensuring that the model relies primarily on IR information, which is more reliable in extreme low-light conditions. Swap Guide remains enabled, allowing IR features to dominate and guide the fusion process, enhancing robustness and accuracy in such challenging scenarios.

Table 1
Configuration of the Adaptive Dynamic Pathway under different lighting conditions

E	Active CAA	Swap	α^I	α^V
Bright	φ_1	Disable	0.3	0.7
Dim	φ_1, φ_2	Enable	0.6	0.4
Dark	$\varphi_1, \varphi_2, \varphi_3$	Enable	0.7	0.3

3.4 Multi-Path and Cross-talk Mitigation

The adaptive multi-path design optimizes computational efficiency based on varying lighting conditions. However, multi-path networks introduce a well-documented issue known as Cross-talk [11, 19], where pathways interfere during training, this interference often results in inconsistent optimization directions for shallow and deep layers, particularly when pathways prioritize different features under varying lighting conditions (e.g., Bright, Dim, Dark). Without mitigation, this conflict can lead to erratic updates in shared parameters, reducing the overall network efficiency and robustness. To address this challenge, we propose a unified approach that incorporates Weight Smoothing and Gradient Averaging, targeting the shared dual-backbone layers, while excluding CAA modules. These techniques are aimed at aligning pathway updates and ensuring stable convergence.

Weight Smoothing: To ensure consistency across pathways, we apply $L2$ regularization to minimize the differences between the weights of shared layers in the dual-backbone network. This encourages the shared parameters to converge toward a common representation while maintaining pathway-specific nuances. The smoothing loss is defined as:

$$L_{smooth} = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} \| W_i^{p1} - W_i^{p2} \|_2^2 \\ + \| W_i^{p2} - W_i^{p3} \|_2^2 \\ + \| W_i^{p1} - W_i^{p3} \|_2^2 \end{pmatrix}, \quad (6)$$

where W_i^{p1}, W_i^{p2}, W_i^{p3} is the weights of the i-th layer in pathways $p1$ (Bright), $p2$ (Dim), $p3$ (Dark), respectively. N is the total number of shared layers in the dual-stream backbone.

Gradient Averaging: To further mitigate inconsistencies during backpropagation, we compute a balanced gradient for shared parameters by averaging contributions from all active pathways. This ensures that updates are proportionally weighted by the number of samples in each pathway, avoiding dominance by any single pathway. The balanced Gradient for a shared parameter W is defined as:

$$G^{avg} = \frac{\sum_{p=1}^{P} |P_p| \cdot G^p}{\sum_{p=1}^{P} |P_p|}, \quad (7)$$

where $|P_p|$ is the number of samples in pathway p, and G^p is the gradient contribution from pathway p.

4.Experiments

4.1 Dataset

All experiments for our model are conducted on the FLIR2 [9] and LLVIP [13] datasets, as shown in Figure 1, the former primarily used in autonomous driving and surveillance applications and the latter primarily used in person detection under low-light conditions.

Table 2
Distribution of lighting conditions across FLIR2 and LLVIP datasets

Dataset	Total	Bright	Dim		Dark
			Normal	Glare	
FLIR2	15113	9601	3102	1325	1125
LLVIP	15490	443	10500	11	4536

Table 2 displays the distribution of data under various lighting conditions within each dataset, with classifications determined by the processes outlined in Equations 5-7. Notably, scenes with Intense Lighting or Glare are classified under the Dim category by analyzing brightness, shape, and rounded edges.

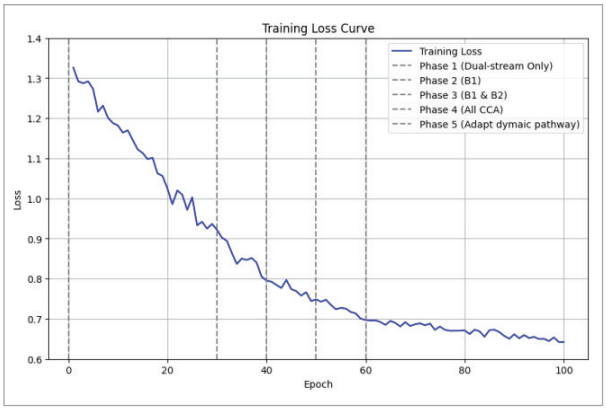
In FLIR2, it includes a large portion of images under Bright conditions (9601), making it well-suited for evaluating model performance in well-lit environments. Additionally, FLIR2 has 3102 images in normal dim conditions and 1325 images with Glare, providing a good representation of challenging low-light scenes affected by light reflection. The Dark category has 1125 images, allowing some testing under extreme low-light conditions.

In LLVIP, it predominantly focuses on low-visibility scenarios, with 10,500 images classified as Dim (mainly Normal), which are ideal for testing the model's robustness in low-light scenes. The Dark

category is also well represented, with 4536 images, making LLVIP particularly valuable for testing IR reliance in very low-light settings.

Figure 4

The training loss curve across 100 epochs. This illustrating the phased training strategy: Phase 1 involves training the dual-stream network without CAA modules. In Phase 2, φ_1 is introduced, followed by φ_1, φ_2 in Phase 3. Phase 4 activates all CAA modules, and Phase 5 implements the ADP. The steady decrease in loss demonstrates effective model convergence and the benefits of the phased training approach.



4.2 Setup and Detail

The experiments were conducted on a system equipped with Ubuntu 22, Python 3.9, Pytorch 1.14, and an NVIDIA RTX 4090 GPU 24GB. Additionally, the model's FPS metrics were evaluated on an NVIDIA Jetson Xavier NX 8G.

Training Strategy: The model was initialized with YOLOv8m pre-trained weights on the COCO dataset. Training was carried out for a total of 100 epochs, with a batch size of 32, using the Adam optimizer with an initial learning rate of 0.05 and a weight decay of 0.0005. Data augmentation was performed using the Mosaic method to enhance the model's robustness to varied input patterns.

A progressive unfreezing approach was employed to stabilize training and ensure effective convergence:

Epochs 1–30: Only the dual-stream network was trained, with the CAA module deactivated to focus on initial feature extraction and alignment.

Epochs 31–40: The φ_1 were unfrozen, allowing early-stage cross-modality attention while maintaining stable learning dynamics.

Epochs 41–50: Both φ_1 and φ_2 modules were unfrozen, enhancing feature fusion through mid-level cross-modality interactions.

Epochs 51–60: All CAA modules were activated, enabling full cross-modality attention throughout the network.

Epochs 61–100: The ADP was activated to fine-tune the model parameters, allowing it to adaptively adjust across different environmental conditions.

4.3 Evaluation Metrics

To evaluate the detection accuracy, we use Mean Average Precision at 50% IoU (mAP@50), which evaluates the model's ability to detect objects with moderate overlap, and Mean Average Precision across 50-95% IoU thresholds (mAP@50-95), which metric provides a more comprehensive assessment, measuring detection precision across multiple overlap thresholds.

To evaluate computational efficiency, we measure the Giga Floating Point Operations (GFLOPs) and Frames Per Second (FPS). GFLOPs represent the computational load for a single forward pass, reflecting the model's memory footprint. for resource-constrained environments. FPS measures the number of frames the model can process per second, indicating real-time performance. Higher FPS values signify a faster model, which is essential for real-time applications.

4.4 Ablation Study

In Table 3, YOLOv8 is used as the baseline model, and detection accuracy is compared across two datasets, FLIR2 and LLVIP. Bolded values highlight the best results for each metric, while improvements are marked with an ↑.

On the FLIR2 dataset, the baseline achieved the best results under VIS light conditions, while on the LLVIP dataset, it performed best under IR conditions. This indirectly validates the effectiveness of our proposed lighting assessment module in addressing varied scenarios. However, using a simple dual-stream architecture for feature fusion results in slightly lower performance compared to the baseline's best scores. This suggests that the model does not fully exploit the complementary nature of IR and VIS features.

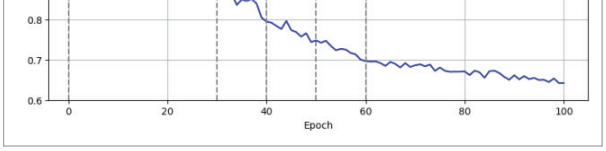


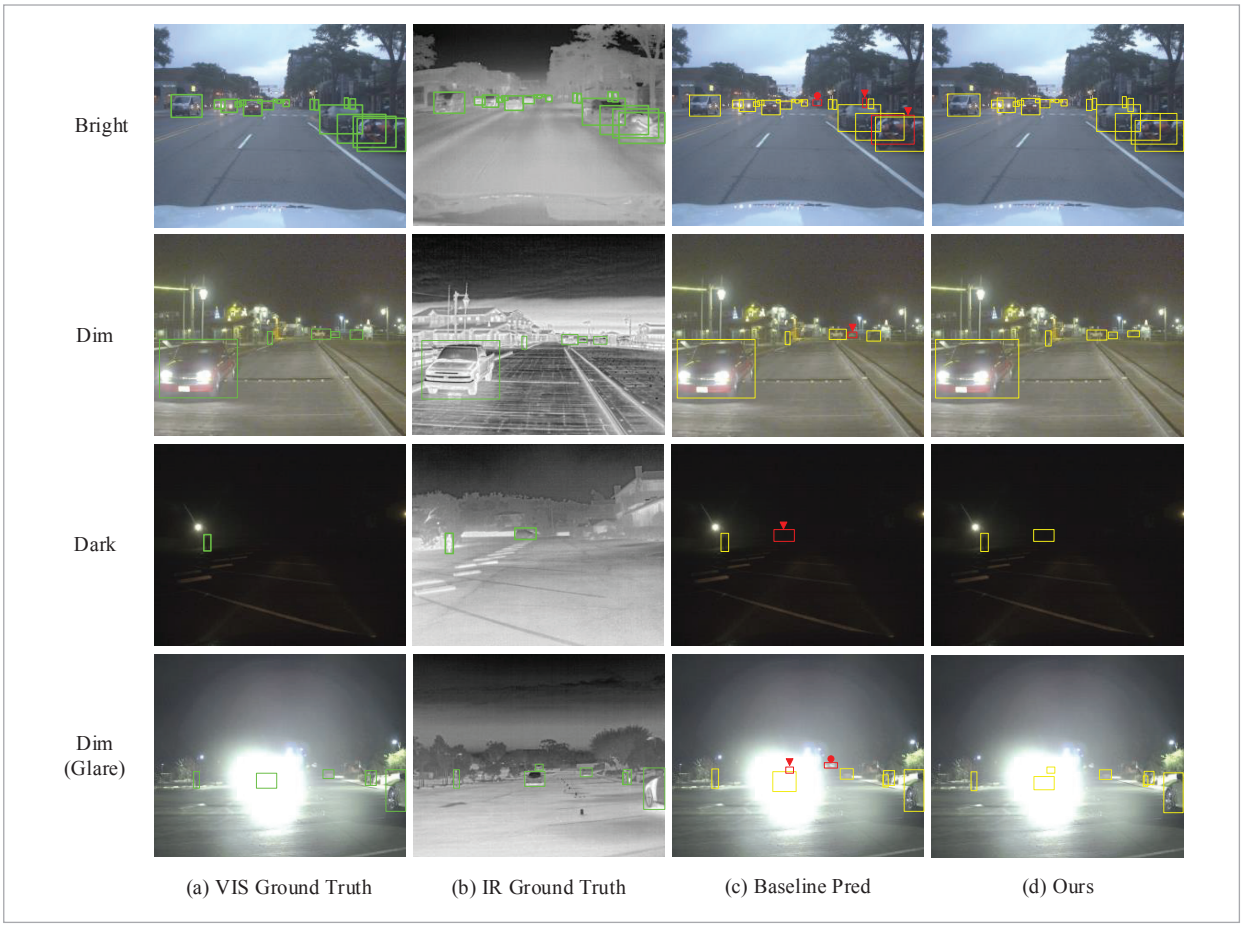
Table 3

Comparison of accuracy as model components are incrementally added

Dataset	Spectrum	Method	mAP@50	mAP@50-95
FLIR2	VIS	YOLOV8	75.7	51.6
	IR	YOLOV8	51.3	32.3
	VIS+IR	+Dual-Stream	72.3	48.3
		+CAA	**76.2(↑0.5)**	51.2
		+ADP	**78.1(↑2.4)**	**53.1(↑1.5)**
LLVIP	VIS	YOLOV8	91.3	58.5
	IR	YOLOV8	96.4	63.7
	VIS+IR	+Dual-Stream	95.9	60.3
		+CAA	**96.8(↑0.4)**	63.5
		+ADP	**97.3(↑0.9)**	**65.5(↑1.8)**

Figure 5

Qualitative comparison of multispectral object detection in the FLIR2 dataset under varying lighting conditions. The rows represent varying lighting conditions, ordered from top to bottom as Bright, Dim, Dark, and Dim with Glare. The first two columns display the Ground Truth for the images, while the third and fourth columns present the detection results from the baseline model and our model, respectively. Note that a red triangle (▲) indicates a False Negative, and a red (●) indicates a False Positive.



The introduction of the CAA module resulted in slight improvements in mAP@50 on both datasets (0.5 and 0.4, respectively). Although modest, these gains indicate that the CAA module effectively begins enhancing feature fusion between IR and VIS modalities.

When the ADP strategy was incorporated, significant improvements were observed. On the FLIR2 dataset, mAP@50 increased by 2.4, and mAP@50-95 improved by 1.5. Similarly, on LLVIP, mAP@50 increased by 1.2, and mAP@50-95 increased by 0.4. These results demonstrate that the ADP strategy enables purposeful adjustments to the network based on input conditions, enhancing both the model's performance and robustness across varying environments.

Figure 5 further illustrates these findings, showcasing detection results across various lighting conditions (Bright, Dim, Dark, and Dim with glare). Compared to the baseline predictions, our proposed method achieves more precise object localization and reduced missed detections, particularly under challenging conditions such as Dim and Dark scenarios. This visual evidence supports the qualitative numerical results, highlighting the improvements in accuracy and robustness brought by CAA and ADP.

In addition, we used YOLOv8, YOLOv5, and Faster R-CNN as baselines on the FLIR2 dataset to evaluate the computational efficiency and real-time performance of our proposed methods, CAA and ADP, applied to fused VIS and IR features. Table 4 summarizes the experimental results.

Across all models, our method consistently improves detection accuracy. Notably, YOLOv8 achieves significant gains in mAP while maintaining real-time performance across all lighting conditions, with FPS values above 30 even in the more computationally demanding Dark environment. YOLOv5 also demonstrates notable accuracy improvements with our method, achieving real-time performance on average, though it struggles to maintain sufficient FPS in Dark conditions. In contrast, Faster R-CNN, due to its inherently higher computational complexity, fails to achieve real-time performance even with our optimizations, despite the observed gains in detection accuracy.

Overall, the results confirm that our approach not only improves detection accuracy but also ensures real-time capability for models like YOLOv8, making it highly practical for real-world applications that require both precision and efficiency.

Table 4

Comparison of accuracy and real-time performance using different detection models with Our Method

Method	Spectrum	Lighting	Parameters	GFLOPs	FPS	mAP@50	mAP@50-95
YOLOV8	VIS	/	25.9M	79.1	56	75.7	51.6
	IR					61.3	42.3
	VIS+IR	Bright	49.4M	138.7	**45**	**78.1**	**53.1**
		Dim	50.5M	139.1	42		
		Dark	54.1M	140.2	39		
YOLOV5	VIS	/	25.1M	64.2	52	71.1	46.2
	IR					59.7	40.8
	VIS+IR	Bright	46.9M	191.1	**41**	**74.9**	**50.8**
		Dim	48.1M	191.3	36		
		Dark	50.4M	192.5	31		
Faster R-CNN	VIS	/	41.1M	180.6	24	72.7	47.2
	IR					60.8	42.3
	VIS+IR	Bright	66.7M	229.1	**19**	**73.3**	**49.8**
		Dim	67.8M	229.7	14		
		Dark	71.9M	230.6	12		

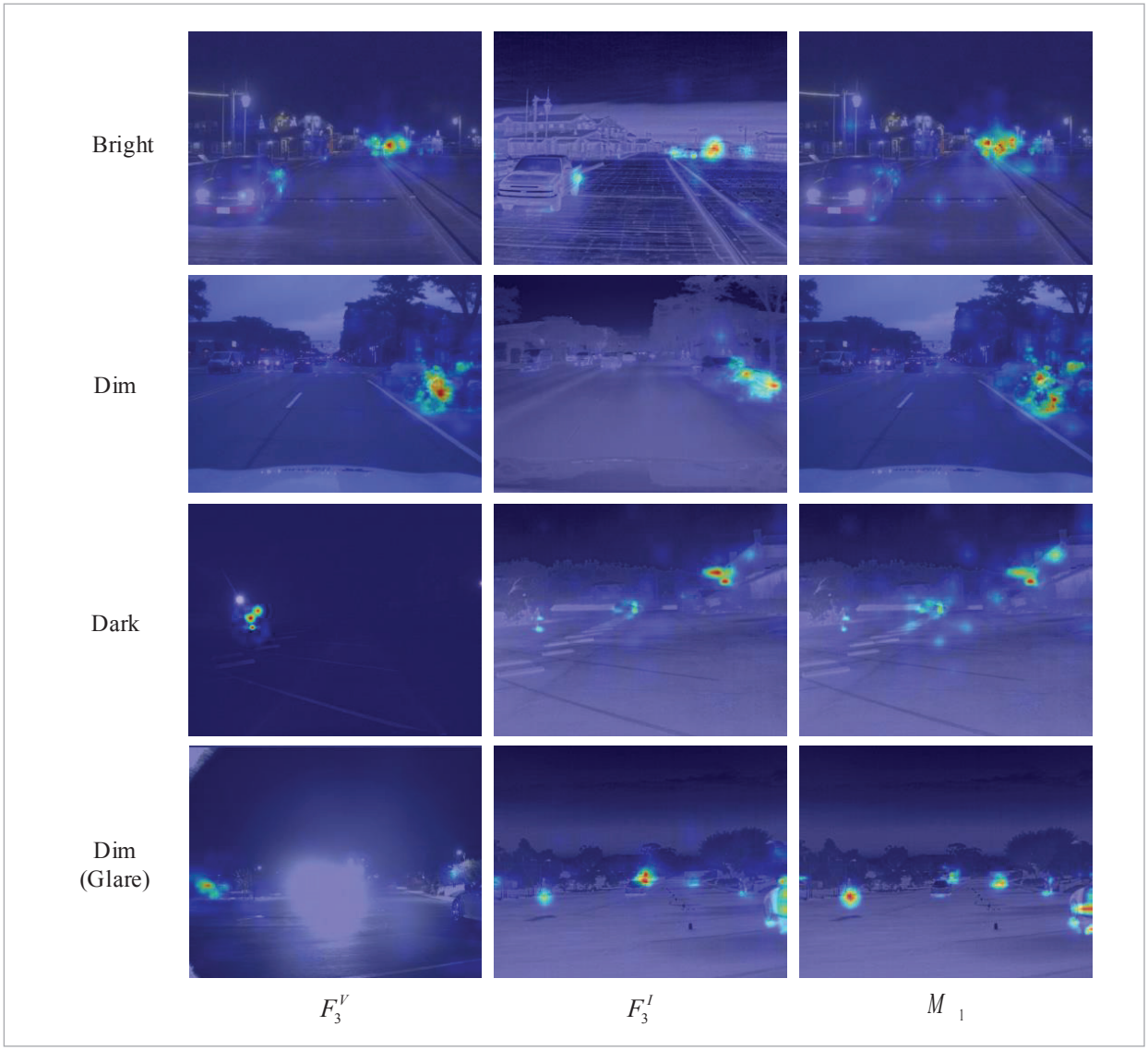
4.5 Visualization of Feature Representation

To better understand the effectiveness of our proposed CAA module and ADP strategy, we visualize the feature maps at different stages of the network. Figure 6 illustrates the feature maps F_3^V, F_3^I, and the fused representation M_1 across various lighting conditions, including Bright, Dim, Dark, and Dim with Glare.

In Bright, F_3^V focuses strongly on high-texture areas such as lane markings and vehicle outlines, while F_3^I exhibits lower activation, as thermal information is less critical in these conditions. After fusion, M_1 effectively highlights key objects, preserving VIS details while subtly incorporating thermal information for enhanced robustness against potential challenges like glare.

Figure 6

Feature representation visualization in the FLIR2 dataset under varying lighting conditions. The first two columns represent the pre-fusion VIS (F_3^V) and IR (F_3^I) features, as shown in Figure 2. The third column illustrates the fused features (M_1) after applying our CAA module, demonstrating improved feature integration across various lighting conditions, including Bright, Dim, Dark, and Dim with Glare.



In Dim, F_3^V activations weaken, particularly in shadowed regions, while F_3^I shows stronger activations around IR objects such as vehicles. M_1 balances these modalities, effectively combining the spatial detail of VIS features with the thermal reliability of IR features, ensuring robust object representation.

In Dark, F_3^V are almost entirely suppressed due to the lack of light, while IR features F_3^I dominate, with clear activations around objects such as vehicles and pedestrians. M_1 relies heavily on thermal information, while still incorporating minimal VIS features for supplementary contextual details, ensuring accurate detection in low-visibility scenarios.

In Dim with Glare, such as headlights or reflective surfaces, F_3^V are overwhelmed by noise, showing spurious activations. Conversely, IR features F_3^I remain unaffected, maintaining clear activations around key objects. M_1 effectively suppresses noise from the VIS modality while leveraging the reliability of IR, resulting in a robust and focused feature representation.

In summary, the visualization highlights the complementary strengths of VIS and IR modalities, with VIS features excelling in well-lit environments and IR dominating in low-light or noisy scenarios. The fused representation effectively uses these contributions, enhancing object focus while suppressing irrelevant noise. This demonstrates the robustness and adaptability of our approach, particularly in challenging conditions like Dim with Glare, where the model prioritizes reliable thermal information over noisy VIS features.

4.6 Comparison with State-of-the-art Methods

To ensure fairness in our comparisons, we selected a diverse range of benchmark methods with publicly available implementations. These include traditional deep learning-based methods (e.g., U2Fusion, ICAFusion) and transformer-based approaches (e.g., CFT, TFDet). These methods were chosen due to their focus on multispectral fusion for object detection, representing the state-of-the-art in the field. We conducted experiments these methods on the FLIR2 and LLVIP datasets. Additionally, we evaluated each method specifically on images classified under the Dim lighting condition in both datasets to verify the effectiveness of our modules in handling low-light or glare condition. The results, summarized in Table 5, where bold values denote comparison items and improvements are marked with an (↑), show that our method achieves competitive performance across both datasets.

In terms of detection Accuracy: On the FLIR2 dataset, our model achieved a mAP@50 of 78.1% and a mAP@50-95 of 53.1%, slightly below U2Fusion (78.3% mAP@50) but outperforming all other methods in mAP@50-95. This highlights the strength of our approach in capturing finer-grained features and adapting to diverse conditions. On the LLVIP dataset, our model achieved a mAP@50 of 97.3% and a mAP@50-95 of 65.5%, again slightly lower than ICAFusion (97.7% mAP@50) but achieving the highest mAP@50-95, surpassing the next best method, U2Fusion, by 0.2%.

The slight decline in mAP@50 compared to U2Fusion and ICAFusion can be attributed to their specialized focus on maximizing modal contributions under specific conditions, such as high-contrast or low-light scenes. Conversely, our model's adaptive design prioritizes consistent performance across a broader range of scenarios, excelling in mAP@50-95 by effectively balancing VIS and IR feature contributions dynamically. The higher mAP@50-95 scores across both datasets demonstrate the robustness and adaptability of our model, particularly in scenarios requiring precise object localization across varying lighting conditions.

Moreover, to further validate the effectiveness of AFDN in challenging lighting conditions, we evaluated all methods specifically on images classified under the dim lighting condition in both FLIR2 and LLVIP datasets. The results demonstrate that our method consistently outperforms other approaches in low-light or glare condition. Notably, in the Dim subset of the FLIR2 dataset, our model achieves a mAP@50 of 75.6% and a mAP@50-95 of 51.9%, reflecting improvements of 1.5% and 1.2%, respectively, over the next best method. Similarly, in the Dim subset of the LLVIP dataset, ADFN achieves a mAP@50 of 93.6% and a mAP@50-95 of 57.7%, outperforming competing methods by 0.7% and 0.6%, respectively. These results highlight the robustness and adaptability of ADFN, particularly in scenarios where traditional methods often struggle to maintain accuracy and further underscores the capability of ADFN to handle low-light and glare conditions effectively, ensuring reliable detection across diverse real-world scenarios.

Table 5
Comparison of performances across multispectral object detection methods on FILR2 and LLVIP dataset

Model	FILR2				LLVIP				FPS
	mAP@50		mAP@50-95		mAP@50		mAP@50-95		
	All	Dim	All	Dim	All	Dim	All	Dim	
CFR_3 [29]	75.4	68.3	51.5	47.9	91.9	87.6	59.3	51.6	/
GAFF [30]	75.1	67.9	50.8	47.5	94.8	88.5	61.9	52.0	/
CFT [22]	77.3	72.3	**52.7**	50.1	97.5	90.6	63.9	**57.1**	28
DetFusion [25]	76.9	70.8	52.6	49.9	96.4	90.5	65.1	54.7	30
TFDet [31]	77.2	71.1	52.4	49.7	95.7	90.1	64.5	54.9	29
U2Fusion [28]	78.3	73.9	51.9	**50.7**	97.2	**92.9**	**65.3**	56.9	27
ICAFusion [24]	77.8	**74.1**	52.6	50.0	97.7	92.1	65.2	55.9	**33**
ADFN(Ours)	78.1	**75.6 (↑1.5)**	**53.1 (↑0.4)**	**51.9 (↑1.2)**	97.3	**93.6 (↑0.7)**	**65.5 (↑0.2)**	**57.7 (↑0.6**	**39(↑6)**

In terms of real-time performance: The FPS values were obtained from experiments conducted on the resource-constrained hardware (NVIDIA Jetson Xavier NX 8G). ADFN achieves the highest FPS of 39, which achieves FPS values of 43, 40, and 35 under Bright, Dim, and Dark conditions, respectively, significantly outperforming all competing methods, including ICAFusion (33 FPS) and transformer-based models such as CFT (28 FPS) and TFDet (27 FPS). This performance demonstrates the computational efficiency of ADFN, which leverages lightweight attention mechanisms and the adaptive pathway design to minimize overhead. Even under resource-constrained environments, ADFN maintains real-time capabilities, making it highly suitable for applications requiring both accuracy and speed.

5.Conclusion

In this paper, we proposed the Adaptive Dynamic Fusion Network (ADFN), a novel architecture designed to address the challenges of multispectral object detection in dynamic and challenging lighting environments. ADFN effectively integrates adaptive multi-path computation with attention-guided feature fusion to dynamically adjust to varying conditions. At its core, the Collaborative and Alternating Attention (CAA) modules enhance feature alignment and cross-modal fusion, while the Adaptive Dynamic Pathway (ADP) strategy ensures the network activates only the most relevant pathways, optimizing both computational efficiency and detection performance.

Comprehensive experiments conducted on the FLIR2 and LLVIP datasets validate the effectiveness of our approach. ADFN achieves competitive detection accuracy, significantly outperforming state-of-the-art methods in terms of mAP@50-95, a metric highlighting its robustness across varying IoU thresholds. Additionally, ADFN maintains real-time inference speeds, making it highly practical for deployment in real-world applications such as surveillance systems, where rapid and accurate object detection is critical for ensuring safety and security. Similarly, in autonomous driving, where vehicles must process diverse environmental conditions with minimal latency, ADFN's ability to adapt to varying lighting scenarios ensures reliable performance in both bright and low-visibility conditions. The results demonstrate that ADFN balances adaptability, accuracy, and efficiency, offering a robust solution for real-world multispectral object detection challenges across diverse lighting conditions.

However, some limitations remain. First, while ADFN demonstrates strong performance across

diverse lighting conditions, its reliance on pre-defined thresholds for pathway activation may not generalize perfectly to highly dynamic or unpredictable environments. Future work could explore learning-based mechanisms to further enhance the adaptability of the pathway selection process. Additionally, the computational efficiency of the network, while sufficient for most real-time applications, could be further optimized for resource-constrained devices, such as edge computing platforms. Lastly, the evaluation datasets primarily focus on low-light and dim environments, leaving room to investigate the performance of ADFN in high-resolution and highly cluttered scenes. Addressing these challenges will pave the way for broader applications of ADFN in more complex and diverse scenarios.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, author-ship, and/or publication of this article.

Data Sharing Agreement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Funding

This work was supported by the Shanxi Provincial Basic Research Program (Industry Development Category) under the joint funding of Taiyuan Heavy Industry, Project No. TZLH20230818007.

References

1. Bahdanau, D. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv Preprint arXiv:1409.0473, 2014.
2. Bakhtiarnia, A., Zhang, Q., Iosifidis, A. Multi-Exit Vision Transformer for Dynamic Inference. arXiv Preprint arXiv:2106.15183, 2021.
3. Bao, W., Huang, M., Hu, J., Xiang, X. Atten-tion-Guided Multi-Modal and Multi-Scale Fusion for Multispectral Pedestrian Detection. In Chinese Con-ference on Pattern Recognition and Computer Vision (PRCV), 2022, 382-393. https://doi.org/10.1007/978-3-031-18907-4_30
4. Castro, F. M., Marin-Jimenez, M. J., Guil, N., Pérez de la Blanca, N. Multimodal Feature Fusion for CNN-Based Gait Recognition: An Empirical Com-parison. Neural Computing and Applications, 2020, 32, 14173-14193. https://doi.org/10.1007/s00521-020-04811-z
5. Chen, K., Liu, J., Zhang, H. IGT: Illumination-Guided RGB-T Object Detection with Transformers. Knowledge-Based Systems, 2023, 268, 110423. https://doi.org/10.1016/j.knosys.2023.110423
6. Cordonnier, J. B., Loukas, A., Jaggi, M. Multi-Head Attention: Collaborate Instead of Concatenate. arXiv Preprint arXiv:2006.16362, 2020.
7. Du, C., Wang, Y., Wang, C., Shi, C., Xiao, B. Selective Feature Connection Mechanism: Concatenating Multi-Layer CNN Features with a Feature Selector. Pattern Recognition Letters, 2020, 129, 108-114. https://doi.org/10.1016/j.patrec.2019.11.015
8. Feng, D., Haase-Schütz, C., Rosenbaum, L., et al. Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges. IEEE Transactions on In-telligent Transportation Systems, 2020, 22(3), 1341-1360. https://doi.org/10.1109/TITS.2020.2972974
9. FLIR. Teledyne FLIR Free ADAS Thermal Dataset v2. FLIR Conservator. https://adas-dataset-v2.flirconservator.com/
10. Guan, D., Cao, Y., Yang, J., Cao, Y., Yang, M. Y. Fu-sion of Multispectral Data Through Illumina-tion-Aware Deep Neural Networks for Pedestrian Detection. Information Fusion, 2019, 50, 148-157. https://doi.org/10.1016/j.inffus.2018.11.017
11. Huang, G., Chen, D., Li, T., Wu, F., van der Maaten, L., Weinberger, K. Multi-Scale Dense Networks for Resource-Efficient Image Classification. In Interna-tional Conference on Learning Representations (ICLR), 2018.
12. Hussain, M. YOLOv1 to v8: Unveiling Each Vari-ant-A Comprehensive Review of YOLO. IEEE Ac-cess, 2024, 12, 42816-42833. https://doi.org/10.1109/ACCESS.2024.3378568
13. Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W. LLVIP: A Visible-Infrared Paired Dataset for Low-Light Vision. In Proceedings of the IEEE/CVF International Con-ference on Computer Vision, 2021, 3496-3504. https://doi.org/10.1109/ICCVW54120.2021.00389
14. Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., Shah, M. Transformers in Vision: A Survey. ACM Computing Surveys (CSUR), 2022, 54(10s), 1-41. https://doi.org/10.1145/3505244

15. Kim, J. U., Park, S., Ro, Y. M. Uncertainty-Guided Cross-Modal Learning for Robust Multispectral Pe-destrian Detection. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(3), 1510-1523. https://doi.org/10.1109/TCSVT.2021.3076466

16. Kouris, A., Venieris, S. I., Laskaridis, S., Lane, N. Multi-Exit Semantic Segmentation Networks. In Eu-ropean Conference on Computer Vision, 2022, 330-349.Lee, Y., Kim, J., Willette, J., Hwang, S. J. MPViT: Multi-Path Vision Transformer for Dense Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 7287-7296.

17. Lee, Y., Kim, J., Willette, J., Hwang, S. J. MPViT: Multi-Path Vision Transformer for Dense Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 7287-7296. https://doi.org/10.1109/CVPR52688.2022.00714

18. Li, C., Song, D., Tong, R., Tang, M. Illumina-tion-Aware Faster R-CNN for Robust Multispectral Pedestrian Detection. Pattern Recognition, 2019, 85, 161-171. https://doi.org/10.1016/j.patcog.2018.08.005

19. Li, H., Zhang, H., Qi, X., Yang, R., Huang, G. Im-proved Techniques for Training Adaptive Deep Networks. In IEEE International Conference on Computer Vision (ICCV), 2019. https://doi.org/10.1109/ICCV.2019.00198

20. Li, K., Wan, G., Cheng, G., et al. Object Detection in Optical Remote Sensing Images: A Survey and a New Benchmark. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 159, 296-307. https://doi.org/10.1016/j.isprsjprs.2019.11.023

21. Lu, J., Yang, J., Batra, D., Parikh, D. Hierarchical Question-Image Co-Attention for Visual Question Answering. Advances in Neural Information Pro-cessing Systems, 2016, 29.

22. Qingyun, F., Dapeng, H., Zhaokui, W. Cross-Modality Fusion Transformer for Multispectral Object Detection. arXiv Preprint arXiv:2111.00273, 2021.

23. Roszyk, K., Nowicki, M. R., Skrzypczyński, P. Adopting the YOLOv4 Architecture for Low-Latency Multispectral Pedestrian Detection in Autonomous Driving. Sensors, 2022, 22(3), 1082. https://doi.org/10.3390/s22031082

24. Shen, J., Chen, Y., Liu, Y., Zuo, X., Fan, H., Yang, W. ICAFusion: Iterative Cross-Attention Guided Feature Fusion for Multispectral Object Detection. Pattern Recognition, 2024, 145, 109913. https://doi.org/10.1016/j.patcog.2023.109913

25. Sun, Y., Cao, B., Zhu, P., Hu, Q. DetFusion: A De-tection-Driven Infrared and Visible Image Fusion Network. In Proceedings of the 30th ACM Interna-tional Conference on Multimedia, 2022, 4003-4011. https://doi.org/10.1145/3503161.3547902

26. Tang, W., He, F., Liu, Y. YDTR: Infrared and Visible Image Fusion via Y-Shape Dynamic Transformer. IEEE Transactions on Multimedia, 2022, 25, 5413-5428.Vaswani, A., Shazeer, N., Parmar, N., Usz-koreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Po-losukhin, I. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017, 30, 5998-6008. https://doi.org/10.1109/TMM.2022.3192661

27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017, 30, 5998-6008.

28. Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H. U2Fusion: A Unified Unsupervised Image Fusion Network. IEEE Transactions on Pattern Analysis and Machine Intel-ligence, 2020, 44(1), 502-518. https://doi.org/10.1109/TPAMI.2020.3012548

29. Zhang, H., Fromont, E., Lefevre, S. Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks. In 2020 IEEE International Conference on Image Processing, 2020, 276-280. https://doi.org/10.1109/ICIP40778.2020.9191080

30. Zhang, H., Fromont, E., Lefèvre, S., Avignon, B. Guided Attentive Feature Fusion for Multispectral Pedestrian Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vi-sion, 2021, 72-80. https://doi.org/10.1109/WACV48630.2021.00012

31. Zhang, X., Zhang, X., Wang, J., et al. TFDet: Tar-get-Aware Fusion for RGB-T Pedestrian Detection. IEEE Transactions on Neural Networks and Learning Systems, 2024. https://doi.org/10.1109/TNNLS.2024.3443455

32. Zhang, Y., Yu, H., He, Y., Wang, X., Yang, W. Illu-mination-Guided RGBT Object Detection with In-ter-and Intra-Modality Fusion. IEEE Transactions on Instrumentation and Measurement, 2023, 72, 1-13. https://doi.org/10.1109/TIM.2023.3251414

33. Zhang, Y., Zeng, W., Jin, S., Qian, C., Luo, P., Liu, W. When Pedestrian Detection Meets Multi-Modal Learning: Generalist Model and Benchmark Dataset. arXiv Preprint arXiv:2407.10125, 2024. https://doi.org/10.1007/978-3-031-73195-2_25