Article

Not peer-reviewed version

SDRFPT-Net: A Spectral Dual-Stream Recursive Fusion Network for Multispectral Object Detection

Peida Zhou , Xiaoyong Sun * , Bei Sun , Runze Guo , Zhaoyang Dang , Shaojing Su

Posted Date: 15 May 2025

doi: 10.20944/preprints202505.0884.v2

Keywords: multispectral object detection; spectral feature representation; recursive progressive fusion; hybrid attention mechanism; target perception enhancement

Article

SDRFPT-Net: A Spectral Dual-Stream Recursive Fusion Network for Multispectral Object Detection

Peida Zhou, Xiaoyong Sun *, Bei Sun, Runze Guo, Zhaoyang Dang and Shaojing Su

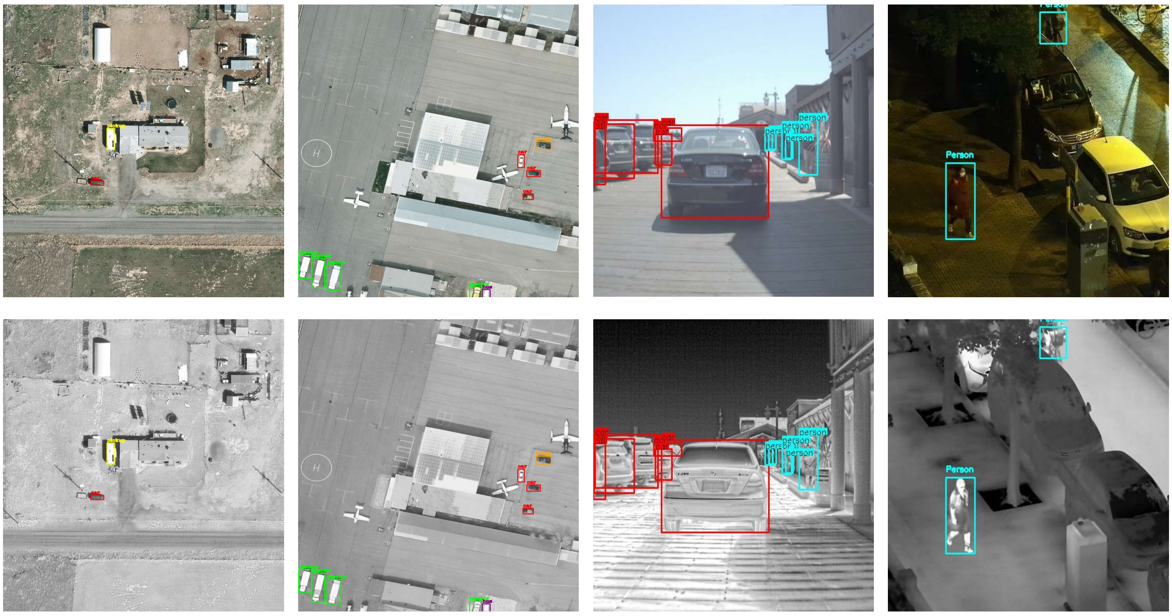
College of Intelligence Science and Technology, National University of Defense Technology,
Changsha 410073, China
* Correspondence: sunxiaoyong14@nudt.edu.cn

Abstract: Multispectral object detection faces challenges in effectively integrating complementary information from different modalities in complex environmental conditions. This paper proposes SDRFPT-Net (Spectral Dual-stream Recursive Fusion Perception Target Network), a novel architecture that integrates three key innovative modules: 1) Spectral Hierarchical Perception Architecture (SHPA), which adopts a dual-stream separated structure with independently parameterized feature extraction paths for visible and infrared modalities; 2) Spectral Recursive Fusion Module (SRFM), which combines hybrid attention mechanisms with recursive progressive fusion strategies to achieve deep feature interaction through parameter-sharing multi-round recursive processing; and 3) Spectral Target Perception Enhancement Module (STPEM), which adaptively enhances target region representation and suppresses background interference. Extensive experiments on VEDAI, FLIR-aligned, and LLVIP datasets demonstrate that SDRFPT-Net significantly outperforms state-of-the-art methods, with improvements of 2.5% in mAP50 and 5.4% in mAP50:95 on VEDAI, 11.5% in mAP50 on FLIR-aligned, and 9.5% in mAP50:95 on LLVIP. Ablation studies further validate the effectiveness of each proposed module. The proposed method provides an efficient and robust solution for multispectral object detection in remote sensing image interpretation, particularly suitable for all-weather monitoring applications from aerial and satellite platforms, as well as in intelligent surveillance and autonomous driving domains.

Keywords: multispectral object detection; spectral feature representation; recursive progressive fusion; hybrid attention mechanism; target perception enhancement

1. Introduction

As one of the core tasks in computer vision, object detection plays a crucial role in remote sensing image interpretation, intelligent surveillance, autonomous driving, and urban planning 1–3]. Remote sensing image interpretation specifically requires robust detection algorithms to identify and locate various objects on the Earth's surface from data acquired by different platforms including drones, aircraft, and satellites. With the rapid development of remote sensing technology, the capability to acquire high-resolution remote sensing images has significantly improved, providing rich data support for the identification and localization of various targets on the Earth's surface [4]. However, due to the unique characteristics of remote sensing platforms, such as varying acquisition angles, diverse imaging conditions, and complex ground scenes, traditional single-modality object detection methods often demonstrate limited performance under complex environmental conditions, especially when targets are in low-light conditions, adverse weather, or cluttered backgrounds [5,6].



(a) (b) (c) (c)

Figure 1. Comparison of multispectral object detection advantages under different lighting conditions. The figure shows detection results for visible (top row) and infrared (bottom row) imaging in daytime (left three columns) and nighttime (rightmost column) scenes. It clearly demonstrates that visible images (top row) provide richer color and texture information for better detection in daylight, while infrared images (bottom row) provide clearer object contours by capturing thermal radiation, showing significant advantages in low-light conditions. This complementarity proves the necessity of multispectral fusion for all-weather object detection, especially in complex and variable environmental conditions.

In practical applications, visible light sensors can capture rich color, texture, and shape information, but they are susceptible to the "same object but different spectrum" phenomenon, exhibiting unstable performance especially under varying lighting conditions [7]. This limitation is particularly evident in remote sensing imagery where atmospheric conditions and diurnal changes can significantly impact image quality. In contrast, infrared sensors are more sensitive to temperature and radiation, performing well in low-light environments, but their low resolution and indistinct edge features make fine-grained target representation difficult [8]. Despite significant advancements in deep convolutional neural networks (CNNs), detection technologies utilizing only a single data source still face enormous challenges in increasingly complex environments [9].

Multi-spectral fusion object detection provides an effective solution for all-weather, all-time target detection by integrating complementary information from different sensors. However, existing CNN-based fusion methods are primarily limited to simple element-wise addition, multiplication, and feature concatenation operations [10–13]. While these strategies improve single-modality detection performance to some extent, they fail to adequately consider deep interactions and correlations between modalities, resulting in poor adaptability [14].

To address the above issues, this paper proposes the Spectral Dual-stream Recursive Fusion Perception Target Network (SDRFPT-Net), a novel multispectral object detection architecture designed to effectively integrate visible and infrared modal information to improve detection performance in complex environments. Unlike existing methods, SDRFPT-Net innovatively proposes a Spectral Hierarchical Perception Architecture (SHPA) based on YOLOv10, providing a solid foundation for multimodal feature extraction, and achieves deep feature interaction and efficient fusion through the Spectral Recursive Fusion Module (SRFM), finally using the Spectral Target Perception Enhancement Module (STPEM) to enhance target region representation and suppress background interference.

Compared to existing research, the main contributions of this paper are:

We propose a spectral dual-stream separated architecture (SHPA) developed based on YOLOv10, with independently parameterized feature extraction paths for visible and infrared modalities, effectively preserving modality-specific information while adapting to the unique characteristics of each spectral domain;

(1) We develop a novel spectral recursive fusion module (SRFM) that combines hybrid attention mechanisms with parameter-sharing recursive processing, achieving deep feature interaction while maintaining computational efficiency through cyclic weight reuse;

(2) We design a spectral target perception enhancement module (STPEM) that adaptively enhances target region representation and suppresses background interference through lightweight mask prediction and similarity-based feature weighting;

(3) We conduct extensive experiments on three benchmark datasets (VEDAI, FLIR-aligned, and LLVIP), demonstrating that our SDRFPT-Net significantly outperforms state-of-the-art methods in multispectral object detection across various environmental conditions and application scenarios.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 introduces the architecture and key modules of SDRFPT-Net in detail; Section 4 presents experimental results and analysis; Section 5 concludes the paper and indicates future research directions.

2. Materials and Methods

This section provides a comprehensive review of multispectral object detection, feature fusion strategies, and applications of YOLO series algorithms in multispectral object detection, with particular focus on remote sensing implementations.

2.1. Multispectral Object Detection

Multispectral object detection technology addresses the limitations of single-modality imaging by fusing complementary information from different spectral bands, enabling all-weather monitoring capabilities [6,15]. In remote sensing applications, this technology effectively overcomes illumination variations, adverse weather conditions, and complex background interference [4]. Early fusion methods relied on traditional mathematical models such as multi-scale transformation [16], sparse representation [17], and saliency-based approaches [18], which were constrained by manually designed feature extractors and predefined fusion rules.

The advent of deep learning has revolutionized multispectral object detection. Significant advances include Liu et al.'s [11] multispectral neural network for improved correlation learning between modalities, Wagner et al.'s [12] deep fusion CNN for visible-infrared image integration, and König et al.'s [13] fully convolutional region proposal networks. These approaches typically employ dual-stream architectures that process different modalities separately before feature fusion at various network levels.

Remote sensing applications present unique challenges including variable object sizes, diverse viewing angles, and unstable imaging conditions [1,4]. To address these issues, researchers have developed specialized solutions combining optical and SAR imagery. Notable contributions include Pang et al.'s [9] RTV-SIFT method for robust cross-modal image registration and Fang et al.'s [7] cross-modal attentive feature fusion technique, which adaptively weights different modality features to enhance detection performance in complex environments.

Recent research has demonstrated the significant potential of Transformer architectures in this domain. Qing et al.'s [19] cross-modal fusion Transformer effectively captures long-range dependencies between modalities, marking an important advancement in attention-based methods for multispectral object detection in remote sensing applications.

2.2. Feature Fusion Strategies

Feature fusion strategies, as the core of multispectral object detection, directly impact final detection performance and are particularly critical in remote sensing image analysis. Based on the stage where fusion occurs, existing methods can be categorized into early fusion, middle fusion, and late fusion [20]. Early fusion directly merges original inputs at the pixel level, offering high computational efficiency but potentially losing modality-specific information; middle fusion occurs after feature extraction, preserving more modal features, also known as feature-level fusion; late fusion integrates outputs from different modalities after detection results are generated [21]. In remote sensing applications, selecting appropriate fusion strategies requires consideration of characteristics from different sensor data and specific application requirements.

Traditional fusion strategies include weighted averaging, maximum/minimum value selection, and principal component analysis [22]. However, these fixed rules struggle to adapt to complex and variable terrain scenes and imaging conditions in remote sensing images. Recently, deep learning-based adaptive fusion strategies have gained widespread attention in remote sensing. Li et al. [23] proposed a multi-granularity attention network that improved fusion effects of infrared and visible images by learning feature correlations at different levels. Wang et al. [24] developed the Res2Fusion architecture, using multi-receptive field aggregation blocks to generate multi-level features and designing non-local attention models for effective fusion, which demonstrates good adaptability to multi-scale characteristics of objects in remote sensing images.

Cross-modal attention mechanisms provide new perspectives for feature fusion in remote sensing imagery. Zhang et al. [25] proposed a cross-stream and cross-scale adaptive fusion network that improved detection performance of objects in multimodal images by establishing connections between different modules and scales. Li et al. [26] designed an attention-based generative adversarial network that achieved efficient fusion of infrared and visible images through adversarial training. In the remote sensing domain, Zhao et al. [27] developed an attention receptive pyramid network specifically for ship detection in SAR images, significantly improving detection accuracy by suppressing background interference. These attention-based methods can adaptively emphasize key information in different modalities, suppress background clutter and noise common in remote sensing images, and achieve more precise feature fusion, particularly suitable for object detection tasks in complex terrain scenes.

2.3. YOLO Series in Multispectral Object Detection

YOLO (You Only Look Once) series algorithms have achieved remarkable success in object detection with their efficient single-stage detection framework [28]. From YOLOv1 [29] to YOLOv10 [30], this series of algorithms has continuously evolved, improving detection accuracy while maintaining efficient inference speed. In remote sensing image analysis, YOLO has attracted significant attention due to its real-time performance and high accuracy characteristics. Chang et al. [31] developed a ship detection method for SAR images based on YOLOv2, while Van Etten [32] proposed the YOLT framework specifically designed for satellite imagery, achieving rapid object detection in large-scale remote sensing images.

In multispectral object detection for remote sensing, YOLO applications primarily utilize dual-stream network structures, processing different modal inputs through parallel backbone networks. Sharma et al. [33] proposed YOLOrs for object detection in multimodal remote sensing imagery, improving detection stability under different imaging conditions through mid-level feature fusion. For SAR ship detection, Ren et al. [6] developed YOLO-Lite, which reduced parameters from 47.1M to 7.64M while increasing frame rate to 103.5 FPS through feature enhancement networks and attention mechanisms without sacrificing detection accuracy. Chen [34] and Guo et al. [35] incorporated attention mechanisms into YOLO frameworks, improving ship detection performance in optical and SAR images respectively.

Current research trends indicate promising directions in combining Transformer with YOLO frameworks. Wang et al. [36] proposed SwinFuse, applying residual Swin Transformer fusion

Peer-reviewed version available at Remote Sens. 2025, 17, 2312; doi:https://doi.org/10.3390/rs17132312

networks to multimodal image fusion, combining CNN's local feature extraction capability with Transformer's global modeling capability. This hybrid architecture provides new technical approaches for multispectral object detection in remote sensing, particularly suitable for processing high-resolution remote sensing data. For small object detection in aerial imagery such as vehicles, Razakarivony et al. [37] developed the VEDAI dataset, providing a standard platform for evaluating different detection algorithms in remote sensing applications.

3. Methodology

This section will detail the SDRFPT-Net algorithm, explaining in order according to the system data flow. The overall architecture of SDRFPT-Net is shown in Figure 2, with its dual-stream design based on YOLOv10 capable of supporting multi-scale spectral feature extraction for visible and infrared modalities, and achieving significant improvement in detection performance through spectral self-adaptive recursive fusion mechanisms and target perception enhancement modules.

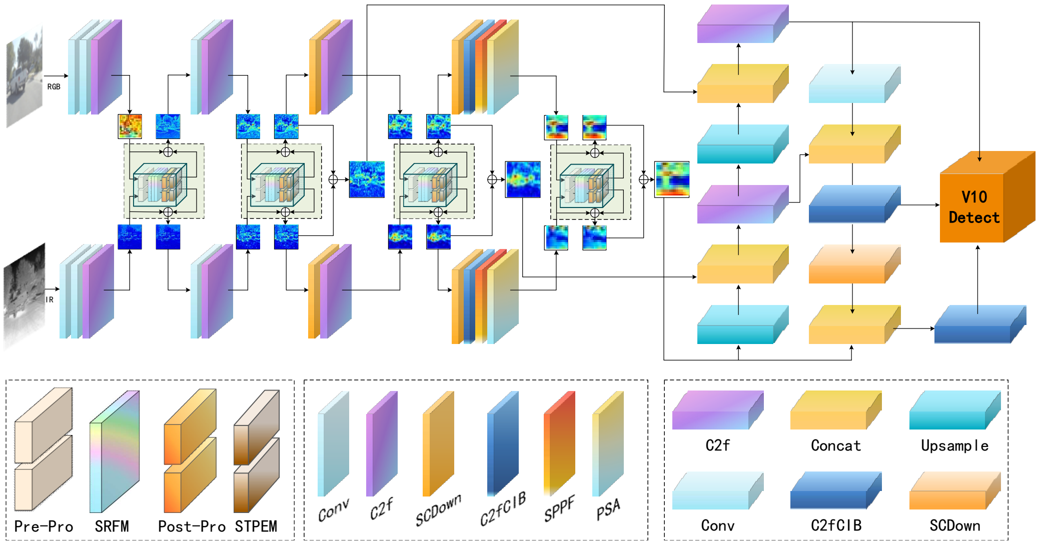


Figure 2. Overall architecture of SDRFPT-Net (Spectral Dual-stream Recursive Fusion Perception Target Network). The architecture employs a dual-stream design with parallel processing paths for visible and infrared input images. The network consists of three key innovative modules: Spectral Hierarchical Perception Architecture (SHPA) for extracting modality-specific features, Spectral Recursive Fusion Module (SRFM) for deep cross-modal feature interaction, and Spectral Target Perception Enhancement Module (STPEM) for enhancing target region representation and suppressing background interference. The feature pyramid and detection head (V10 Detect) enable multi-scale object detection.

The system's data processing flow is as follows: First, the input visible and infrared images are processed separately through dual-stream feature extraction networks, generating feature maps of different scales; Then, these feature maps undergo deep interaction and fusion through the spectral self-adaptive recursive fusion module; Next, the fused features are further enhanced by the self-adaptive target perception enhancement module to strengthen the representation of target regions; Finally, the enhanced multi-scale features are aggregated through feature aggregation and input to the detection head, generating the final detection results.

Compared to traditional single-modality object detection methods, this architecture can more effectively utilize the complementary information of RGB and infrared images, especially showing greater detection accuracy and robustness in challenging scenarios such as low light, adverse weather, and complex backgrounds.

3.1. Spectral Hierarchical Perception Architecture (SHPA)

The SHPA architecture, as the core design of this algorithm, effectively processes visible and infrared spectral domain information through a dual-stream structure, laying the foundation for hierarchical perception and fusion of multi-scale features. This architecture is based on YOLOv10's excellent features and has been systematically improved for multi-modal perception.

3.1.1. Dual-Stream Separated Spectral Architecture Design

Compared to YOLOv10's single backbone network feature extraction mechanism, the dual-stream separated spectral architecture proposed in this paper can effectively process RGB-IR dual-modal data's heterogeneous properties, as shown in Figure 3.

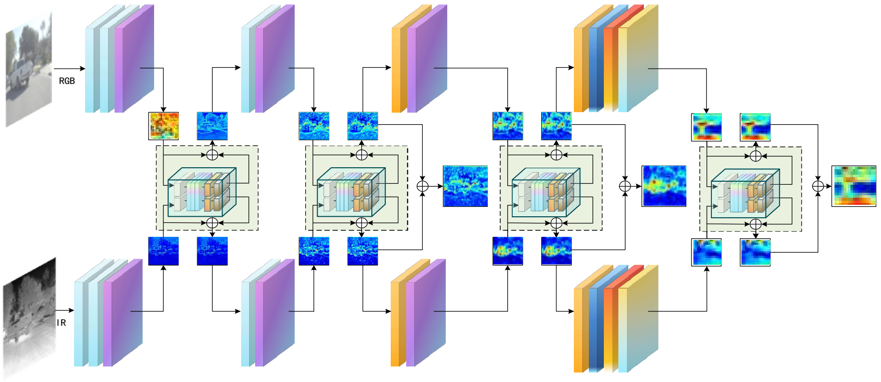


Figure 3. Dual-stream separated spectral architecture design in SDRFPT-Net. The architecture expands a single feature extraction network into a dual-stream structure, where the upper stream processes visible spectral information while the lower stream handles infrared spectral information. Although both processing paths share similar network structures, they employ independent parameter sets for optimization, allowing each stream to specifically learn the feature distribution and representation of its respective modality.

This architecture expands a single feature extraction network into a dual-stream network, processing visible spectral and infrared spectral information separately. The two feature extraction streams share similar network structures but use independent parameters, and the feature extraction process of the dual-stream network can be formalized as:

$$F_{rgb} = \mathcal{F}_{rgb}(I_{rgb};\theta_{rgb}) \tag{1}$$

$$F_{ir} = \mathcal{F}_{ir}(I_{ir};\theta_{ir}) \tag{2}$$

where, I_{rgb} and I_{ir} represent RGB and infrared input images, $\mathcal{F}_{rgb}$ and $\mathcal{F}_{ir}$ represent the corresponding feature extraction functions, θ_{rgb} and θ_{ir} represent their respective network parameters.

The main advantages of the dual-stream architecture are:

(1) It can design specific extraction strategies for the characteristics of different spectral domains, thereby better adapting to the characteristics of data from each modality;

(2) It preserves the unique information of each spectral domain, avoiding the potential loss of information that might occur when processing in a single network;

(3) It captures the feature distributions of different spectral domains through independent parameters, improving the diversity of feature representations.

Compared to YOLOv10's single feature extraction path, the dual-stream architecture shows greater robustness in complex environments, especially when the quality of information from one modality decreases (such as insufficient RGB information at night or reduced infrared contrast during the day), the system can still maintain detection performance by relying on stable information provided by the other modality.

3.1.2. Multi-Scale Spectral Feature Expansion

To comprehensively capture the multi-scale representation of targets, this paper designs a multi-scale spectral feature expansion mechanism. In each spectral stream, features form a multi-scale feature pyramid through progressive downsampling. For each spectral domain $s \in \{rgb, ir\}$, the feature expansion process can be represented as:

$$F_i^s = \mathcal{H}i(F_{i-1}^s; \theta_i^s), \quad i \in \{1,2,3,4\} \tag{3}$$

Where, F_i^s represents the level i feature, $\mathcal{H}_i$ represents the downsampling function, θ_i^s is the corresponding parameter. Specifically, the spatial resolution and channel number of each level feature are:

$$F_1^s \in \mathbb{R}^{B\times128\times\frac{H}{4}\times\frac{W}{4}} \quad (\text{P2/4}) \tag{4}$$

$$F_2^s \in \mathbb{R}^{B\times256\times\frac{H}{8}\times\frac{W}{8}} \quad (\text{P3/8}) \tag{5}$$

$$F_3^s \in \mathbb{R}^{B\times512\times\frac{H}{16}\times\frac{W}{16}} \quad (\text{P4/16}) \tag{6}$$

$$F_4^s \in \mathbb{R}^{B\times1024\times\frac{H}{32}\times\frac{W}{32}} \quad (\text{P5/32}) \tag{7}$$

3.1.3. Feature Aggregation and Detection

After multi-scale expansion, images go through Pre-Pro, SRFM, Post-Pro, STPEM models for fusion, thereby obtaining high-quality multi-scale fusion features. These features need to be further aggregated and processed to generate the final object detection results, as shown in Figure 4.

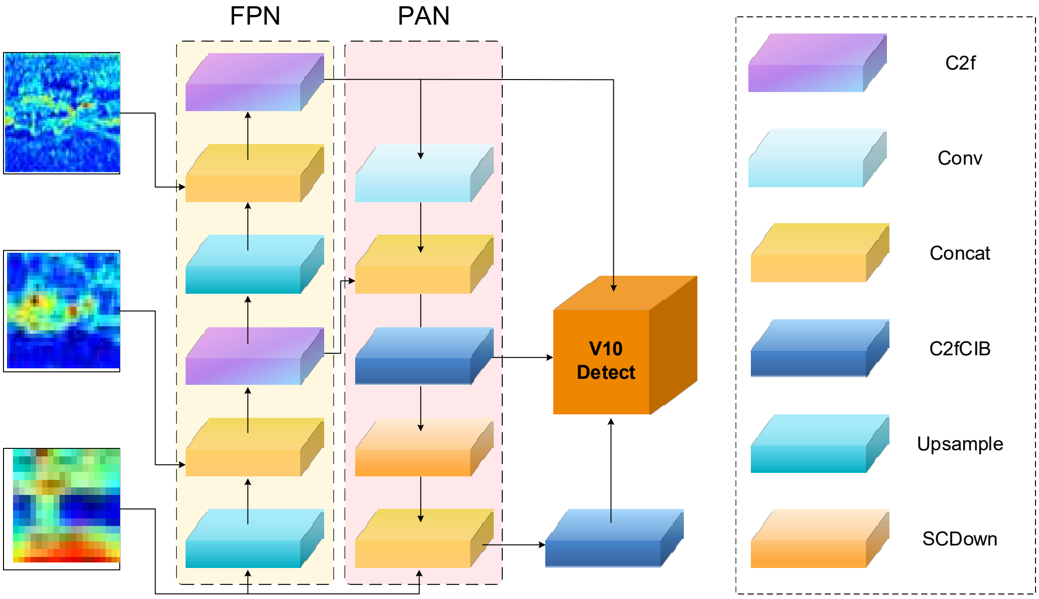


Figure 4. Multi-scale fusion feature aggregation and detection process in SDRFPT-Net. The figure shows features from three different scales (P3, P4, P5) that already contain fused information from visible and infrared

modalities. The middle section presents two complementary information flow networks: Feature Pyramid Network (FPN) and Path Aggregation Network (PAN). FPN (light blue background) follows a top-down path, transferring high-level semantic information to low-level features, while PAN (light pink background) follows a bottom-up path, transferring low-level spatial details to high-level features. This bidirectional feature flow mechanism ensures that features at each scale incorporate both fine spatial localization information and rich semantic representation.

First, multi-scale fusion features are aggregated through the feature pyramid (FPN) and path aggregation network (PAN), enhancing information exchange between features of different scales: :

$$P_i = FPN_i(F_{fused}) \tag{8}$$

$$M_i = \begin{cases} PAN_i(P_i), & \text{if } i = 3 \\ PAN_i(M_{i-1}, P_i), & \text{if } i > 3 \end{cases} \tag{9}$$

Where, P_i represents the FPN output of level i feature, M_i represents the PAN output of level i feature, F_{fused} represents the feature after fusion, containing complementary information from RGB and IR.

FPN transmits semantic information from high levels to low levels, while PAN transmits spatial details from low levels to high levels, forming a powerful feature representation. This bidirectional feature flow mechanism ensures that features at each scale can incorporate both rich semantic information and fine spatial details.

Finally, the aggregated features pass through the v10Detect detection head for object detection:

$$D = \text{Detect}(M_3, M_4, M_5) \tag{10}$$

Where $Detect(\cdot)$ represents the detection function, with outputs including object class, bounding box coordinates, and confidence information. v10Detect adopts a more efficient feature decoding method, including dynamic convolution and branch specialization design, further improving detection accuracy and efficiency. v10Detect employs branch specialization design, designating specialized branches for bounding box regression, feature processing, and classification tasks, further improving detection accuracy and efficiency.

Compared to YOLOv10, our feature aggregation and detection stage utilizes the advantages brought by modal fusion and target perception enhancement, allowing the detection head to perform object detection based on richer and more accurate feature representations. This is particularly important in low light, adverse weather, and complex background conditions, as single-modal information is often unreliable in these scenarios.

3.2. Spectral Recursive Fusion Module (SRFM)

The SRFM module achieves deep interaction and optimized integration of RGB-IR dual-modal features through innovative fusion mechanisms, significantly improving detection performance in complex environments. Unlike traditional fusion methods, SRFM combines hybrid attention mechanisms with recursive progressive fusion strategies organically, achieving deep multi-modal feature interaction while maintaining parameter efficiency, providing powerful feature representation capabilities for multispectral object detection.

As shown in Figure 5, SRFM receives dual-stream features from SHPA and outputs fused enhanced features after cyclic progressive fusion. This section will introduce the design principles, key components and workflow of the mechanism in detail.

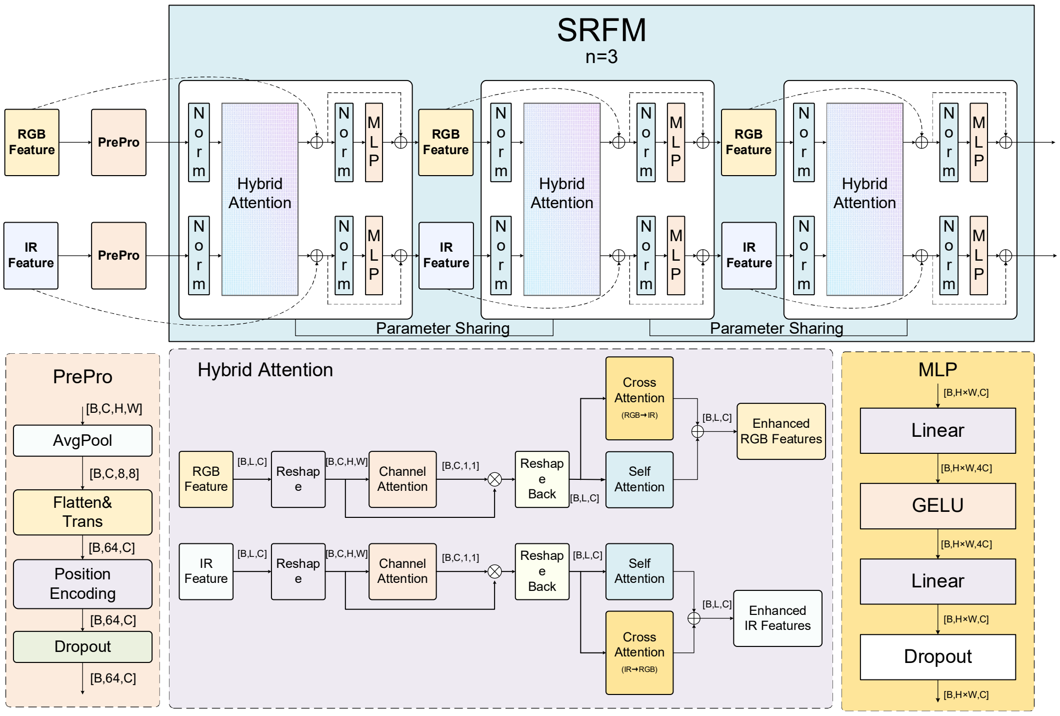


Figure 5. Detailed architecture of the Spectral Recursive Fusion Module (SRFM). The framework is divided into two main parts: the upper light blue background area shows the overall recursive fusion process, labeled as 'SRFM n=3', indicating a three-round recursive fusion strategy. This part receives RGB and IR dual-stream features from SHPA and processes them through three cascaded hybrid attention units with parameter sharing to improve computational efficiency. The lower part shows the detailed internal structure of the hybrid attention unit, including preprocessing components (PrePro) such as AvgPool, Flatten, and Position Encoding; the hybrid attention mechanism implementation including Channel Attention, Self Attention, and Cross Attention; and the MLP module with Linear layers, GELU activation, and Dropout.

3.2.1. Hybrid Attention Mechanism

The hybrid attention mechanism builds a comprehensive feature enhancement system by integrating three complementary mechanisms: self-attention, cross-modal attention, and channel attention, capturing complex feature dependencies from spatial, modal relationship, and channel importance dimensions. This multi-dimensional feature enhancement design significantly improves the model's processing capability for different scenes.

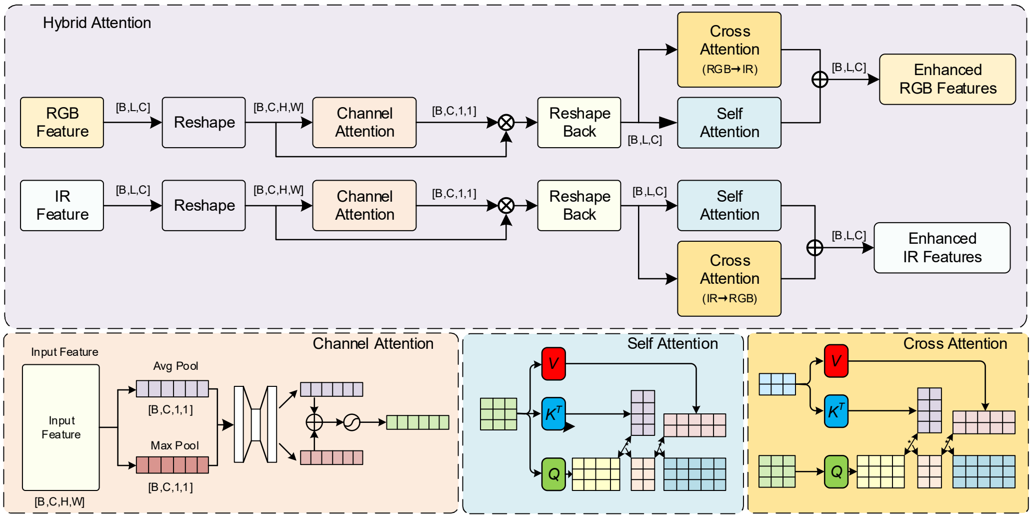


Figure 6. Detailed structure of the hybrid attention mechanism in SDRFPT-Net. The mechanism integrates three complementary attention computation methods to achieve multi-dimensional feature enhancement. The upper part shows the overall processing flow: RGB and IR features first undergo reshaping and enter the Channel Attention module, which focuses on learning the importance weights of different feature channels. After reshaping back, the features simultaneously enter both Self Attention and Cross Attention modules, capturing intra-modal spatial dependencies and inter-modal complementary information. Finally, the outputs from both attention modules are added to generate enhanced RGB and IR feature representations.

According to the data flow shown in the figure, the overall calculation process of the hybrid attention mechanism can be expressed as:

$$F_{out}^{RGB} = SelfAtt(F_{chan}^{RGB}) + CrossAtt(F_{chan}^{RGB}, F_{chan}^{IR}) \tag{11}$$

$$F_{out}^{IR} = SelfAtt(F_{chan}^{IR}) + CrossAtt(F_{chan}^{IR}, F_{chan}^{RGB}) \tag{12}$$

Where, $F_{chan}^{RGB} = ChanAtt(F_{in}^{RGB})$ and $F_{chan}^{IR} = ChanAtt(F_{in}^{IR})$ respectively represent RGB and infrared features after channel attention processing.

Channel attention mechanism. The channel attention sub-module learns channel dependencies through global information modeling, providing all-around enhanced features for the spectral recursive progressive fusion strategy. Given input feature map $F \in \mathbb{R}^{B\times C\times H\times W}$, where B, C, H, W respectively represent batch size, channel number, height, and width, the calculation process of channel attention can be expressed as:

$$F_{chan} = ChanAtt(F) = F \cdot \sigma(W_2 \cdot ReLU(W_1 \cdot F_{avg}) + W_2 \cdot ReLU(W_1 \cdot F_{max})) \tag{13}$$

Where, $F_{avg} = AvgPool(F) \in \mathbb{R}^{B\times C\times 1\times 1}$ and $F_{max} = MaxPool(F) \in \mathbb{R}^{B\times C\times 1\times 1}$ represent global average pooling and global maximum pooling operations; $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $W_2 \in \mathbb{R}^{C\times\frac{C}{r}}$ are shared weight fully connected layer parameter matrices, where r is the reduction rate. The weights are mapped to the $(0,1)$ interval ; finally, the channel attention weights are applied to the original features through element-wise multiplication.

Self-attention mechanism. The self-attention mechanism focuses on capturing spatial dependencies within a modality, allowing features to attend to related regions within the same modality, providing richer contextual information for the spectral hierarchical perception architecture. For input feature F_{chan}, the calculation process of self-attention can be expressed as:

$$F_{self} = SelfAtt(F_{chan}) = Softmax\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \tag{14}$$

Where, $Q = W_Q \cdot F$, $K = W_K \cdot F$, $V = W_V \cdot F$ are query, key, and value matrices obtained through learnable parameter matrices W_Q, W_K, W_V; d_k is the feature dimension, serving as a scaling factor to avoid gradient vanishing problems.

Cross-modal attention mechanism. The cross-modal attention mechanism is used to capture complementary information between different modalities, establishing connections between visible and infrared features, and is the core component for achieving spectral information exchange. The unique aspect of cross-modal attention is that it uses the query from one modality to interact with the keys and values from another modality, thereby enabling information flow between modalities. For RGB and IR features, the calculation of cross-modal attention can be expressed as:

$$F_{cross}^{RGB} = CrossAtt(F_{chan}^{RGB}, F_{chan}^{IR}) = \alpha \cdot Softmax\left(\frac{Q^{RGB} \cdot (K^{IR})^{T}}{\sqrt{d_k}}\right) \cdot V^{IR} \tag{15}$$

$$F_{cross}^{IR} = CrossAtt(F_{chan}^{IR}, F_{chan}^{RGB}) = \alpha \cdot Softmax\left(\frac{Q^{IR} \cdot (K^{RGB})^{T}}{\sqrt{d_k}}\right) \cdot V^{RGB} \tag{16}$$

Where, $Q^{RGB} = W_Q \cdot F_{chan}^{RGB}$, $K^{IR} = W_K \cdot F_{chan}^{IR}$, $V^{IR} = W_V \cdot F_{chan}^{IR}$ are the cross-modal attention calculation from RGB to IR; $Q^{IR} = W_Q \cdot F_{chan}^{IR}$, $K^{RGB} = W_K \cdot F_{chan}^{RGB}$, $V^{RGB} = W_V \cdot F_{chan}^{RGB}$ are the matrices for IR to RGB calculation; α is a learnable scaling factor that controls the strength of cross-modal information fusion.

Finally, the outputs of self-attention and cross-modal attention are added to obtain the final enhanced features:

$$F_{out}^{RGB} = F_{self}^{RGB} + F_{cross}^{RGB} \tag{17}$$

$$F_{out}^{IR} = F_{self}^{IR} + F_{cross}^{IR} \tag{18}$$

Through this design, the hybrid attention mechanism can simultaneously attend to channel importance, spatial dependency relationships, and modal complementary information, building a more comprehensive and robust feature representation.

3.2.2. Recursive Progressive Fusion Strategy

Multi-modal feature fusion is a key challenge in RGB-T object detection. Traditional multi-modal feature fusion methods typically enhance performance by stacking multiple Transformer blocks, but this approach leads to dramatic increases in parameter count and computational complexity. Inspired by the "review-consolidate" mechanism in human learning processes, this paper proposes a spectral hierarchical recursive progressive fusion strategy, achieving feature progressive refinement through repeatedly applying the same feature transformation operations, thereby enhancing fusion effects without increasing model parameters.

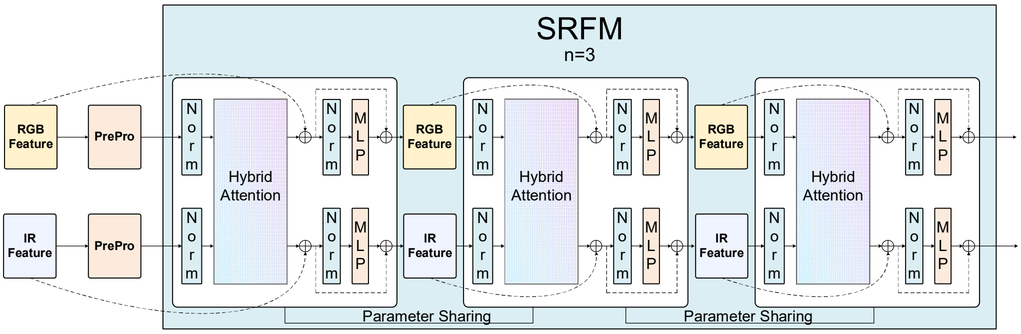


Figure 7. Spectral recursive progressive fusion architecture in SDRFPT-Net. The light blue background area (labeled as 'SRFM n=3') shows the parameter-sharing three-round recursive fusion process. The left side includes RGB and IR input features after preprocessing (PrePro), which flow through three cascaded hybrid attention units. The key innovation is that these three processing units share the exact same parameter set (indicated by 'Parameter Sharing' connections), achieving deep recursive structure without increasing model complexity. Each processing unit contains normalization (Norm) components and MLP modules, forming a complete feature refinement path.

Parameter Cycling Reuse Structure. The core idea of the spectral hierarchical recursive progressive fusion strategy is to use the same set of parameters for multiple rounds of feature refinement. Each refinement builds on the results of the previous round, forming a continuous, progressive feature fusion process. This process can be expressed as:

$$[F_{RGB}^{t+1}, F_{IR}^{t+1}] = T(F_{RGB}^{t}, F_{IR}^{t}; \theta) \tag{19}$$

Where, F_{RGB}^{t} and F_{IR}^{t} respectively represent the visible and infrared features after the t round of cycling, T represents the feature transformation function, θ is the reused model parameter.

Through multiple cycles, the feature representation ability is continuously enhanced:

$$[F_{RGB}^{fianl}, F_{IR}^{final}] = T^{n}(F_{RGB}^{0}, F_{IR}^{0}; \theta) \tag{20}$$

Where, T^{n} represents applying the transformation function T continuously n times, F_{RGB}^{0} and F_{IR}^{0} are the initial features.

Compared to traditional methods, the cyclic weight reuse structure significantly reduces the model parameter count, while achieving deep feature interaction through multiple refinements.This design not only improves the model's representation ability but also alleviates the risk of overfitting.

Spectral Feature Progressive Fusion. Spectral feature progressive fusion is the core characteristic of this strategy, progressively fusing different spectral domain features. This progressive fusion process operates in the spectral dimension, ensuring each spectral property is fully preserved and mutually enhanced. The fusion process includes the following key steps:

1. Spectral feature normalization: Normalization is performed separately on visible and infrared features, expressed as follows.

$$\hat{F}_{RGB} = \mathrm{LN}(F_{RGB}) \tag{21}$$

$$\hat{F}_{IR} = \mathrm{LN}(F_{IR}) \tag{22}$$

2. Hybrid attention calculation: Apply hybrid attention mechanism to process normalized features, expressed as follows, $HybridAttention(\cdot)$ represents hybrid attention calculation.

$$F'_{RGB}, F'_{IR} = \mathrm{HybridAttention}(\hat{F}_{RGB}, \hat{F}_{IR}; \theta_{attn}) \tag{23}$$

3. Spectral residual connection: Combine attention outputs with original spectral features, expressed as follows.

$$F''_{RGB} = F_{RGB} + F'_{RGB} \tag{24}$$

$$F''_{IR} = F_{IR} + F'_{IR} \tag{25}$$

4. Spectral feature enhancement: Further enhance each spectral feature through multilayer perceptron and residual connection, expressed as follows.

$$F'''_{RGB} = F''_{RGB} + \mathrm{MLP}(\mathrm{LN}(F''_{RGB})) \tag{26}$$

$$F'''_{IR} = F''_{IR} + \mathrm{MLP}(\mathrm{LN}(F''_{IR})) \tag{27}$$

Where, $LN(\cdot)$ represents layer normalization operation, $MLP(\cdot)$ represents multilayer perceptron.

Progressive feature refinement process. The progressive feature refinement process can be viewed as a "feature distillation" mechanism, where each round of cycling makes the feature representation more pure and effective. In this research, we adopt a fixed 3-round cycling structure, a design based on extensive experimental validation.

The refinement process can be divided into three stages:

1. First round of cycling: Initial fusion stage. Mainly captures basic intra-modal and inter-modal relationships, establishing initial feature interaction;

2. Second round of cycling: Feature reinforcement stage. Based on the already established initial relationships, further strengthens important feature connections, suppressing noise and irrelevant information;

3. Third round of cycling: Feature refinement stage. Performs final optimization and fine-tuning on features, forming high-quality fusion representations.

This three-round progressive refinement process can be expressed as:

$$F_{RGB}^{3}, F_{IR}^{3} = T^{3}(F_{RGB}^{0}, F_{IR}^{0}; \theta) \tag{28}$$

The progressive refinement mechanism creates a "deep cascade" effect, achieving deep network feature representation capabilities within a fixed parameter space, which is fundamentally different from traditional "multi-layer stacking" approaches. Traditional methods require introducing new parameter sets for each additional layer, while our method achieves deeper effective network depth through parameter reuse while maintaining parameter efficiency.

Spectral Multi-scale Fusion Mechanism. The spectral multi-scale fusion mechanism is an important component of the recursive progressive fusion strategy, applying recursive progressive fusion on features of different scales to achieve comprehensive multi-scale feature optimization. This mechanism includes the following key designs:

1. Multi-scale feature selection: The fusion strategy is applied separately on three scales—P3/8, P4/16, and P5/32—ensuring thorough fusion of features at all three scales;

2. Inter-scale information flow: Information exchange between features of different scales is achieved through FPN and PAN structures;

The multi-scale fusion process can be expressed as:

$$F_{s}^{fusion} = T_{s}^{N}(F_{s}^{init}; \theta_{s}), s \in \{P3/8, P4/16, P5/32\} \tag{29}$$

Where s represents the feature scale index, and T_s represents the feature transformation function for the s-th scale. By applying recursive progressive fusion across multiple scales, the system comprehensively enhances the representation capability of features at different scales, providing a solid foundation for detecting targets of various sizes.

3.3. *Spectral Target Percpetion Enhancement Module (STPEM)*

The STPEM module focuses on enhancing target regions in features while reducing background interference. Through mask generation and feature enhancement mechanisms, this module significantly improves the model's detection capability for small and low-contrast targets, providing more precise feature representation for object detection in complex environments.

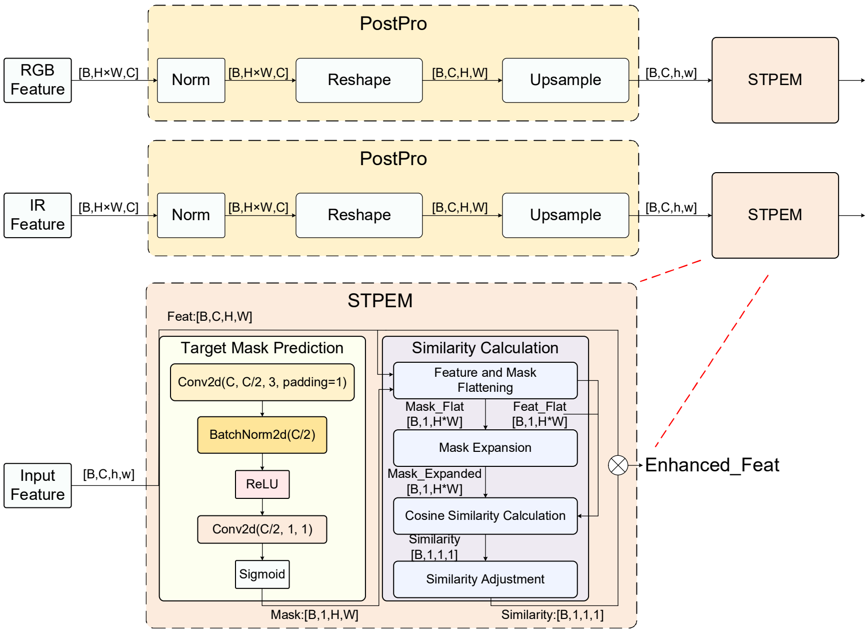


Figure 8. Spectral Target Perception Enhancement Module (STPEM) structure and data flow. The module aims to enhance target region representation while suppressing background interference to improve detection accuracy. The figure is divided into three main parts: the upper and middle parts show parallel processing paths for features from RGB and IR modalities. Both feature paths first go through post-processing (PostPro) modules, including feature normalization, reshaping, and upsampling, before entering the STPEM module for enhancement processing.

3.3.1. Lightweight Mask Prediction

Lightweight mask prediction is the core component of STPEM. Given input feature $F \in \mathbb{R}^{B \times C \times H \times W}$, mask prediction first predicts target region masks through a lightweight convolutional network:

$$M = \sigma(M_{pred}(F)) \tag{30}$$

Where M_{pred} represents the mask prediction network, and σ represents the sigmoid activation function. The mask prediction network adopts a two-layer convolutional structure:

$$M_{pred}(F) = \text{Conv}_{1\times1}(\text{ReLU}(\text{BN}(\text{Conv}_{3\times3}(F)))) \tag{31}$$

The first layer is a 3×3 convolution that reduces the number of channels from C to C/2, followed by batch normalization and ReLU activation; the second layer is a 1×1 convolution that reduces the number of channels from C/2 to 1, outputting a single-channel mask. Finally, the sigmoid function maps values to the [0,1] range, representing the probability that each position contains a target.

The mask prediction network is essentially learning "what feature patterns might correspond to target regions." For example, in RGB images, targets typically have distinct edges and texture features; in infrared images, targets often appear as regions with significant temperature differences from the background. The mask prediction network captures these feature patterns through convolutional operations to generate masks representing potential target regions.

3.3.2. Similarity Calculation and Adjustment

After mask generation, the module calculates the cosine similarity between features and masks to evaluate the correlation between each feature channel and the target region, thereby establishing explicit associations between feature channels and potential target regions:

Peer-reviewed version available at *Remote Sens.* **2025**, *17*, 2312; doi:https://doi.org/10.3390/rs17132312

$$F_{flat} = \mathrm{Flatten}(F) \in \mathbb{R}^{B\times C\times(H\times W)} \tag{32}$$

$$M_{flat} = \mathrm{Flatten}(M) \in \mathbb{R}^{B\times 1\times(H\times W)} \tag{33}$$

$$M_{expanded} = \mathrm{Expand}(M_{flat}, C) \in \mathbb{R}^{B\times C\times(H\times W)} \tag{34}$$

$$S = \mathrm{CosineSimilarity}(F_{flat}, M_{expanded}, \mathrm{dim}=2) \in \mathbb{R}^{B\times C} \tag{35}$$

Where the $Flatten(\cdot)$ operation flattens the spatial dimensions of the features, the $Expand(\cdot)$ operation expands the mask to the same number of channels as the features, and $\mathrm{CosineSimilarity}(\cdot)$ calculates the cosine similarity between two vectors.

After calculating the similarity between each channel and the mask, further processing is done through averaging operations and a learnable adjustment layer:

$$S_{avg} = \mathrm{Mean}(S, \mathrm{dim}=1, \mathrm{keepdim}=True) \in \mathbb{R}^{B\times 1} \tag{36}$$

$$S_{adjusted} = \sigma(S_{adjust}(S_{avg})) \in \mathbb{R}^{B\times 1\times 1\times 1} \tag{37}$$

Where S_{adjust} is a 1×1 convolutional layer for adjusting similarity, and σ is the sigmoid activation function. This learnable similarity adjustment mechanism enables the module to adaptively adjust similarity calculations according to different scenes, improving the flexibility and adaptability of the module.

3.3.3. Feature Enhancement Mechanism

Finally, the enhanced feature $F_{enhanced}$ is achieved through similarity weighting:

$$F_{enhanced} = F \times S_{adjusted} \tag{38}$$

The core idea of this weighting mechanism is: if a feature has high similarity with the predicted target region, it is preserved or enhanced; if the similarity is low, the feature is suppressed. In this way, features of target regions are effectively enhanced while features of background regions are suppressed, thereby improving the signal-to-noise ratio of the features.

The STPEM module significantly improves the performance of multispectral object detection by effectively identifying and enhancing potential target regions, showing excellent performance especially when processing complex background scenes.

4. Experiments

This section will detail the experimental results of SDRFPT-Net on the VEDAI, FLIR-aligned and LLVIP datasets, verifying the effectiveness of our proposed algorithm. First, we introduce the experimental setup and evaluation datasets; second, we compare SDRFPT-Net with current state-of-the-art multispectral object detection methods; finally, we analyze the contribution of each innovative module through comprehensive ablation experiments.

4.1. Datasets and Evaluation Metrics

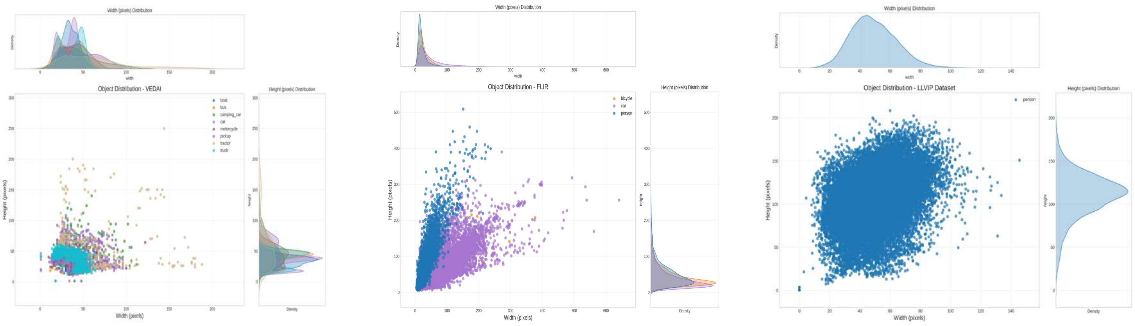
4.1.1. Datasets

This study employs two widely used multispectral object detection benchmark datasets: VEDAI[38], FLIR-aligned [39] and LLVIP [40].

VEDAI. A benchmark dataset for aerial remote sensing vehicle detection with 1,210 images (1024×1024 pixels) at 12.5-25.0 cm resolution, captured from aircraft platforms. Contains eight vehicle classes across diverse terrain backgrounds (woods, cities, roads) with 3,700+ annotated targets. Specifically designed for remote sensing applications, featuring complex terrain textures and variable imaging perspectives common in earth observation tasks. Particularly challenging due to small target size (0.7% of image pixels), significant scale variations, and complex ground backgrounds typical in aerial imagery acquisition. Evaluation uses 10-fold cross-validation with mAP and FPPI metrics.

FLIR-aligend. An aligned dataset derived from FLIR ADAS [41] containing 4,129 training and 1,013 testing image pairs of spatially aligned thermal infrared and visible light images. Features three object classes (person, car, bicycle) captured in diverse environments (urban roads, highways, residential areas) under various conditions (day, night, dusk). Valuable for evaluating multispectral detection algorithms in real driving scenarios.

LLVIP. A visible-infrared paired dataset for low-light visual tasks with 16,836 image pairs captured at night (6-10 PM) across 26 locations. All pairs are time-space aligned containing annotated pedestrians. Targets difficult to identify in visible images are clearly visible in infrared. Images registered via semi-automatic method ensuring identical field of view, processed to uniform 1080×720 resolution.



(**a**) (**b**) (c)

Figure 9. Object size distribution characteristics in multispectral detection datasets. (a) VEDAI dataset showing eight vehicle classes (boat, bus, camping_car, car, motorcycle, pickup, tractor, truck), with most vehicles concentrated in width range of 20-100 pixels and height range of 10-100 pixels, while tractors extend to greater heights (up to 250 pixels); (b) FLIR-aligned dataset displaying three object classes (blue: person, purple: car, orange: bicycle), where persons exhibit slender features (narrow width, greater height up to 400 pixels), cars show a triangular distribution pattern (width 20-300 pixels, height 10-300 pixels); (c) LLVIP dataset illustrating pedestrian distribution with a highly concentrated circular clustering pattern (width 20-80 pixels, height 40-150 pixels). The density curves at the top and right of all figures show the statistical distribution of width and height, providing important reference for network design.

4.1.2. Metrics

To comprehensively evaluate the performance of object detection models, this study adopts the following standard evaluation metrics:

Precision (P). Precision is a key metric for measuring detection accuracy, defined as the ratio of correctly detected targets (true positives) to all detected targets (true positives and false positives). This metric reflects the accuracy of model target recognition, calculated as follows:

$$P = \frac{TP}{TP + FP} \tag{39}$$

Recall (R). Recall measures the model's ability to detect all targets, defined as the ratio of correctly detected targets (true positives) to all actually existing targets (true positives and false negatives). This metric reflects the completeness of the model's capture of all targets in the image, calculated as follows:

$$R = \frac{TP}{TP + FN} \tag{40}$$

Mean Average Precision at IoU=0.50 (mAP50). Mean Average Precision at IoU=0.50 (mAP50): mAP50 is the average precision calculated at an Intersection over Union (IoU) threshold of 0.50. This metric primarily measures the model's performance on "simple" detection tasks, i.e., the detection accuracy when the overlap area between the predicted bounding box and the ground truth bounding box accounts for at least 50% of the total area.

Mean Average Precision across IoU=0.50:0.95 (mAP50-95). mAP50-95 is a more comprehensive evaluation metric that calculates the average precision at different IoU thresholds (from 0.50 to 0.95, with a step size of 0.05), and then takes the average of these values. Compared to mAP50, mAP50-95 better reflects the model's localization accuracy by considering a range of stricter IoU thresholds. A high mAP50-95 score indicates that the model can maintain good performance under stricter localization standards, which is particularly important for applications requiring high localization accuracy.

4.2. Experimental Setup

All experiments in this study were conducted on a server equipped with an NVIDIA RTX 4090 GPU (24GB memory) and an Intel Core i7-13700 processor (24 cores), with 62GB system memory. The experimental environment was based on Linux 20.04 operating system, PyTorch 2.0.1 deep learning framework, CUDA 11.7, cuDNN 8.7.0, and Python 3.9.21.

During the training process, we used the SGD optimizer with a momentum parameter of 0.937 and a weight decay coefficient of 0.0005. The initial learning rate was set to 0.01, and a cosine annealing strategy was adopted to reduce the learning rate to 0.01 times its initial value by the end. The batch size was fixed at 4, input image dimensions were uniformly adjusted to 640×640 pixels, and the maximum number of training epochs was 300. An early stopping strategy was also implemented—automatically terminating the training process when there was no performance improvement for 30 consecutive epochs.

4.3. Comparison with State-of-the-Art Methods

4.3.1. On the VEDAI Dataset

Table 1 presents the performance comparison between SDRFPT-Net and current state-of-the-art methods on the VEDAI dataset. The VEDAI dataset serves as a significant benchmark for evaluating small vehicle detection performance in aerial remote sensing imagery, and is used to test the effectiveness of multispectral object detection algorithms under complex terrain backgrounds.

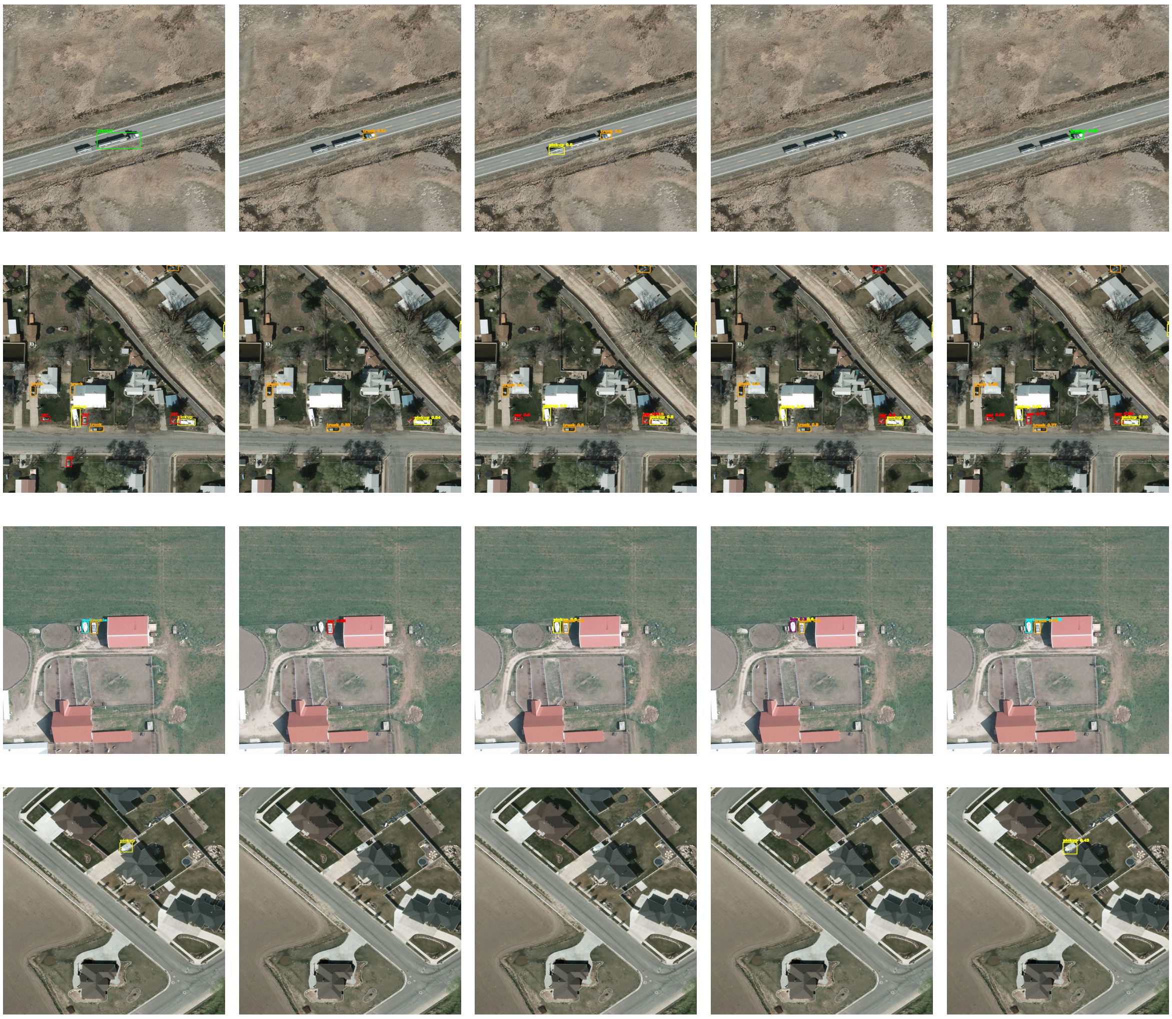
Table 1. Performance comparison of SDRFPT-Net with state-of-the-art methods on the VEDAI dataset. The table presents Precision (P), Recall (R), mean Average Precision at IoU threshold of 0.5 (mAP50), and mean Average Precision across IoU thresholds from 0.5 to 0.95 (mAP50:95). V-I indicates the fusion of visible and infrared modalities. The best results are highlighted in bold.

Methods	Modality	P	R	mAP50	mAP50:95
YOLOv5	Visible	0.704	0.604	0.675	0.398
	Infrared	0.521	0.514	0.498	0.280
YOLOv8	Visible	0.727	0.431	0.537	0.333
	Infrared	0.520	0.523	0.494	0.291
YOLOv10	Visible	0.610	0.523	0.587	0.329
	Infrared	0.421	0.463	0.447	0.244
YOLOv10-add	V-I	0.500	0.527	0.537	0.276
CFT	V-I	0.701	0.627	0.672	0.427
CMAFF	V-I	0.616	0.508	0.452	0.275

SuperYOLO	V-I	0.790	0.678	0.716	0.425
SDRFPT-Net (ours）	V-I	**0.796**	**0.683**	**0.734**	**0.450**

Experimental results demonstrate that the proposed SDRFPT-Net achieves excellent performance across all key evaluation metrics. Compared to single-modality baseline methods, SDRFPT-Net realizes significant performance improvements. Specifically, in terms of mAP50, SDRFPT-Net reaches 0.734, showing a 2.5% improvement over the second-best performing SuperYOLO (0.716). Under the more stringent mAP50:95 evaluation criterion, SDRFPT-Net achieves 0.450, outperforming the second-best method CFT (0.427) by 5.4%, indicating that our method not only enhances target detection rates but also maintains high-precision bounding box localization capabilities.

As shown in Figure 10, under complex terrain textures and multi-scale remote sensing imaging conditions, SDRFPT-Net can accurately detect all targets and provide precise bounding box localization. In comparison, YOLOv10-add exhibits notable missed detections for small targets, while CMAFF and CFT demonstrate issues with false detections and imprecise bounding box localization. These visualization results intuitively validate the detection advantages of SDRFPT-Net in complex remote sensing scenarios where spatial resolution, imaging angle, and environmental factors pose significant challenges for object detection tasks.



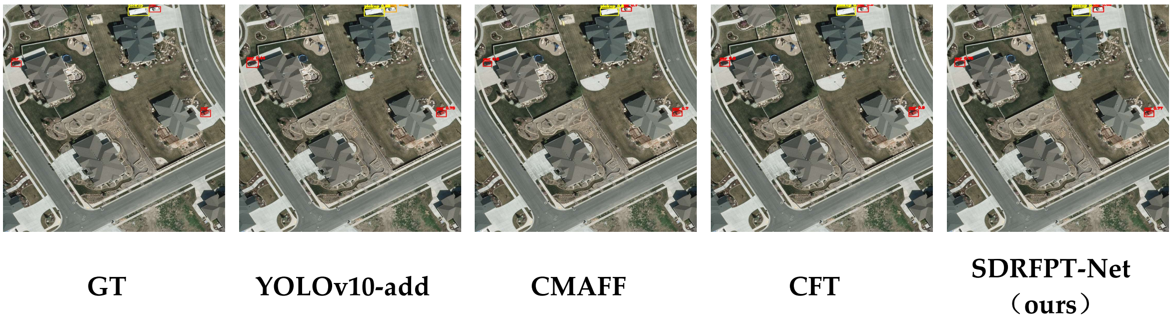


Figure 10. Comparison of detection performance for different multispectral object detection models on the VEDAI dataset. Images are organized by columns from left to right: Ground Truth (GT), YOLOv10-add, CMAFF, CFT, and our proposed SDRFPT-Net model. Each row displays typical remote sensing scenarios with different environmental conditions and object distributions, including roads, building clusters, and open areas from an aerial perspective. The visualization results clearly demonstrate the advantages of SDRFPT-Net: under complex terrain textures and multi-scale remote sensing imaging conditions, SDRFPT-Net can accurately detect all targets with more precise bounding box localization. In contrast, other methods such as YOLOv10-add, CMAFF, and CFT exhibit certain limitations in small target detection and bounding box localization accuracy.

Comprehensive analysis indicates that SDRFPT-Net significantly improves vehicle detection performance in aerial remote sensing imagery by effectively integrating complementary information from visible and infrared modalities, particularly excelling in detecting small-sized targets and targets in complex terrain backgrounds. This capability is crucial for remote sensing applications where targets often occupy only a fraction of the pixel space and must be distinguished from heterogeneous ground textures and shadows. This outstanding performance can be attributed to the synergistic effect of three innovative modules: Spectral Hierarchical Perception Architecture (SHPA), Spectral Recursive Fusion Module (SRFM), and Spectral Target Perception Enhancement Module (STPEM). These modules collectively enhance the network's ability to process multi-source remote sensing data and extract discriminative features across different spectral domains, addressing the unique challenges of earth observation imagery.

4.3.2. On the FLIR-Aligned Dataset

Table 2 shows the comparison results of SDRFPT-Net with other state-of-the-art methods on the FLIR-aligned dataset. This dataset is widely used as a benchmark for evaluating the performance of multispectral object detection systems under various environmental conditions.

Table 2. Performance comparison of SDRFPT-Net with state-of-the-art methods on the FLIR-aligned dataset. The table presents Precision (P), Recall (R), mean Average Precision at IoU threshold of 0.5 (mAP50), and mean Average Precision across IoU thresholds from 0.5 to 0.95 (mAP50:95). The best results are highlighted in bold.

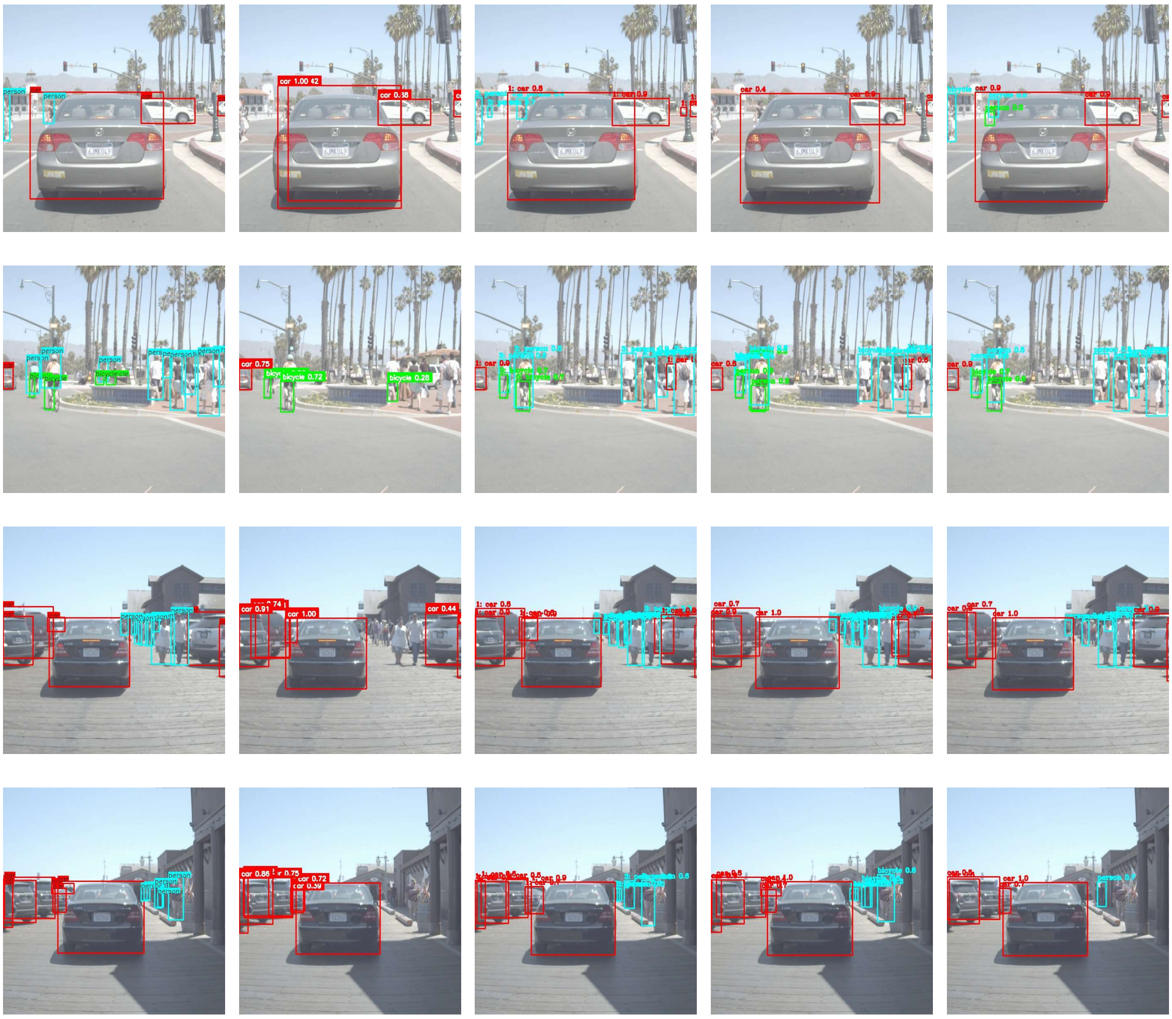
Methods	Modality	P	R	mAP50	mAP50:95
YOLOv5	Visible	0.531	0.395	0.441	0.202
	Infrared	0.625	0.468	0.539	0.272
YOLOv8	Visible	0.532	0.396	0.448	0.218
	Infrared	0.559	0.514	0.549	0.288
YOLOv10	Visible	0.727	0.538	0.620	0.305
	Infrared	0.773	0.618	0.727	0.424
YOLOv10-add	V-I	0.748	0.623	0.701	0.354
CMA-Det	V-I	0.812	0.468	0.518	0.237
TFDet	V-I	0.827	0.606	0.653	0.346
CMAFF	V-I	0.792	0.550	0.558	0.302
BA-CAMF Net	V-I	0.798	0.632	0.704	0.351
SDRFPT-Net (ours）	V-I	**0.854**	**0.700**	**0.785**	**0.426**

The experimental results show that the proposed SDRFPT-Net outperforms existing methods in all key metrics including precision, recall, and mAP. Compared to the best-performing single-modality method YOLOv10-infrared (with mAP50 of 0.727), SDRFPT-Net's mAP50 improved by 8.0% (from 0.727 to 0.785). This improvement demonstrates that our proposed multi-modal fusion strategy can effectively integrate complementary information from different spectral domains.

Compared to other multispectral fusion methods, SDRFPT-Net achieves precision (P) and recall (R) of 0.854 and 0.700 respectively, significantly outperforming other methods. Particularly in terms of mAP50, SDRFPT-Net (0.785) improved by 11.5% compared to the second-best performing BA-CAMF Net (0.704), demonstrating the superiority of our proposed spectral dual-stream recursive fusion perception architecture in multispectral object detection tasks.

Notably, under the more stringent mAP50:95 evaluation criterion, SDRFPT-Net achieves 0.426, comparable to the single-modality baseline YOLOv10-infrared (0.424), while significantly outperforming other multi-modal fusion methods (with the highest being BA-CAMF Net's 0.351). This indicates that SDRFPT-Net not only improves target detection rate but also maintains high-precision bounding box localization capability.

As shown in Figure 11, in complex lighting and occlusion conditions, SDRFPT-Net (j) can accurately detect all targets with more precise bounding boxes. In contrast, YOLOv10-add (g) has some missed detections on small targets, while TFDet (h) and CMAFF (i) have some false detections and inaccurate bounding box issues. These visualization results intuitively demonstrate the detection advantages of SDRFPT-Net in complex scenes.



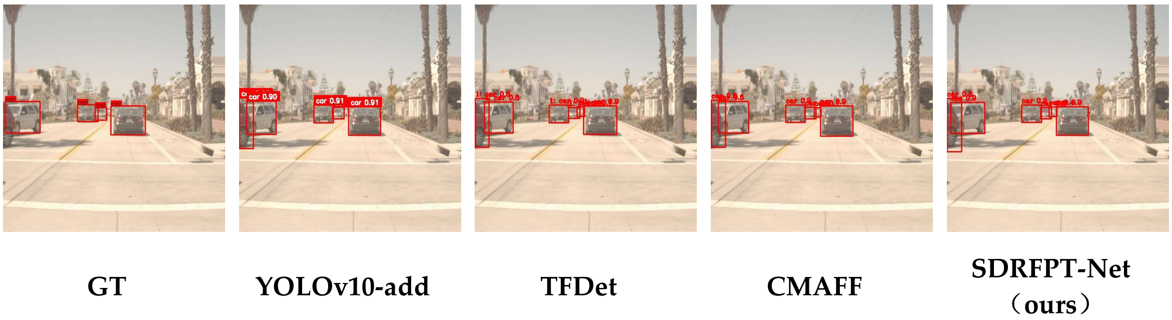


Figure 11. Comparison of detection performance for different multispectral object detection models on various scenarios in the FLIR-aligned dataset. Images are organized by columns from left to right: Ground Truth (GT), YOLOv10-add, TFDet, CMAFF, and our proposed SDRFPT-Net model. Each row shows typical scenarios with different environmental conditions and object distributions, including close-range vehicles, multiple roadway targets, parking areas, narrow streets, and open roads. The visualization results clearly demonstrate the advantages of SDRFPT-Net: in the first row's close-range vehicle scene, SDRFPT-Net's bounding boxes almost perfectly match GT; in the second row's complex multi-target scene, it successfully detects all pedestrians and bicycles without obvious misses; in the third row's parking lot scene, it accurately identifies multiple closely parked vehicles with precise bounding box localization.

4.3.3. On the LLVIP Dataset

Table 3 shows the comparison results of SDRFPT-Net with other state-of-the-art methods on the LLVIP dataset. The LLVIP dataset focuses on pedestrian detection in low-light environments and is an important benchmark for evaluating algorithm robustness in nighttime scenes.

Table 3. Performance comparison of SDRFPT-Net with state-of-the-art methods on the LLVIP dataset. The table presents Precision (P), Recall (R), mean Average Precision at IoU threshold of 0.5 (mAP50), and mean Average Precision across IoU thresholds from 0.5 to 0.95 (mAP50:95). V-I indicates the fusion of visible and infrared modalities. The best results are highlighted in bold.

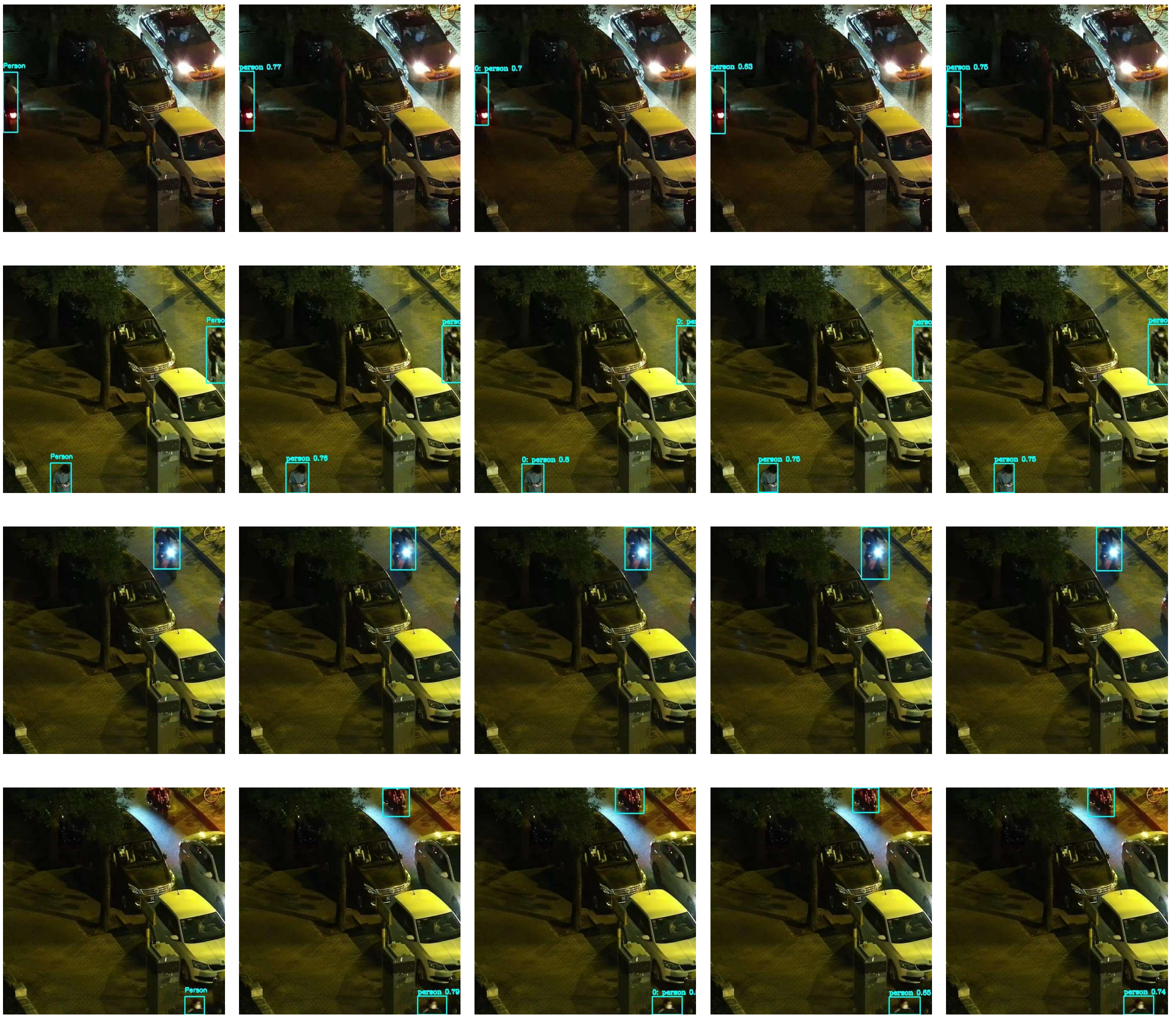
Methods	Modality	P	R	mAP50	mAP50:95
YOLOv5	Visible	0.906	0.820	0.895	0.504
	Infrared	0.962	0.898	0.960	0.631
YOLOv8	Visible	0.933	0.829	0.896	0.513
	Infrared	0.956	0.901	0.961	0.645
YOLOv10	Visible	0.914	0.833	0.892	0.512
	Infrared	0.962	0.909	0.961	0.637
YOLOv10-add	V-I	0.961	0.893	0.957	0.628
TFDet	V-I	0.960	0.896	0.960	0.594
CMAFF	V-I	0.958	0.899	0.915	0.574
BA-CAMF Net	V-I	0.866	0.828	0.887	0.511
SDRFPT-Net (ours）	V-I	**0.963**	**0.911**	**0.963**	**0.706**

As can be seen from Table 3, SDRFPT-Net also achieves excellent performance on the LLVIP dataset. In terms of mAP50, SDRFPT-Net reaches 0.963, showing a slight improvement (0.2%) compared to the closest-performing single-modality method YOLOv10-infrared and multi-modal method TFDet (both at 0.961). Although this improvement is modest, achieving further improvement at an already near-saturated performance level is still significant. Under the more stringent mAP50:95 evaluation criterion, SDRFPT-Net reaches 0.706, significantly outperforming all comparison methods. Compared to the second-best performing YOLOv8-infrared (0.645), it improves by 9.5%, indicating that the proposed method has significant advantages in precise bounding box localization. This result confirms that SDRFPT-Net can not only detect target locations but also more accurately describe target boundaries.

Notably, under the low-light conditions of the LLVIP dataset, the infrared modality alone can achieve high performance (e.g., YOLOv8-infrared achieves an mAP50 of 0.961). In this case, SDRFPT-Net still achieved performance improvements through effective integration of complementary information from visible light, especially with significant improvement in mAP50:95 (from 0.645 to 0.706). This indicates that the proposed spectral recursive fusion mechanism can still effectively extract and integrate valuable features from the visible light modality even when infrared information is dominant.

SDRFPT-Net's recall reaches 0.911, higher than all comparison methods, indicating it has stronger target detection capability and can find pedestrian targets that might be missed by other methods, which is particularly important for practical application scenarios.

Figure 12 provides visualized detection results of various methods in typical nighttime scenes from the LLVIP dataset. Qualitative analysis shows that SDRFPT-Net can accurately locate all pedestrian targets in low-contrast environments with high bounding box matching accuracy. In contrast, other multi-modal fusion methods such as YOLOv10-add, TFDet, and CMAFF show varying degrees of detection instability in complex scenes, including missed detections, false detections, or bounding box localization deviations. These visualization results further confirm the advantages of SDRFPT-Net demonstrated in the quantitative evaluation.



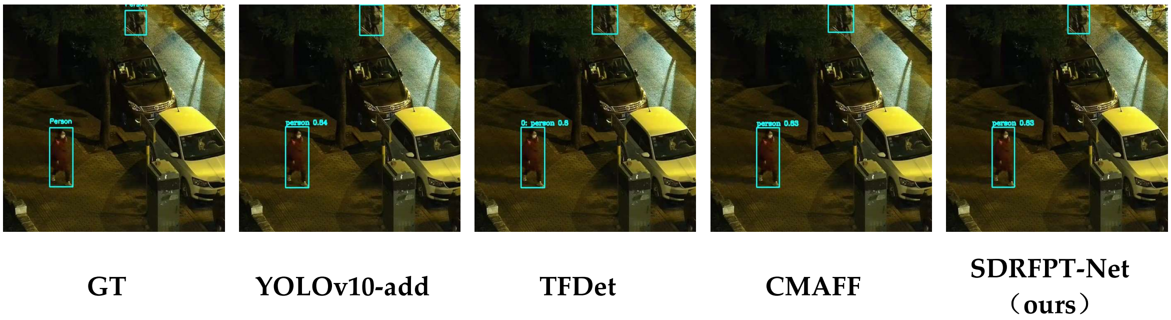


Figure 12. Comparison of pedestrian detection performance for different detection models on nighttime low-light scenes from the LLVIP dataset. Images are organized by columns from left to right: Ground Truth (GT), YOLOv10-add, TFDet, CMAFF, and our proposed SDRFPT-Net model. The rows display five typical nighttime scenes representing challenging situations with different lighting conditions, viewing angles, and target distances. In all low-light scenes, SDRFPT-Net demonstrates excellent pedestrian detection capability: accurately identifying distant pedestrians with precise bounding boxes in the first row's street lighting scene; maintaining stable detection performance despite strong light interference in the second and fourth rows; successfully detecting distant pedestrians that other methods tend to miss in the fifth row's dark area.

Figure 12 shows the visualization detection results of various algorithms on the LLVIP dataset. In typical nighttime low-light scenes, SDRFPT-Net (j) can accurately detect all pedestrians with more precise bounding boxes. In contrast, YOLOv10-add (g), TFDet (h), and CMAFF (i) exhibit missed detections or inaccurate bounding box issues in some complex scenes. These visualization results further confirm the detection advantages of SDRFPT-Net in low-light environments.

Combining the experimental results from both the FLIR-aligned and LLVIP datasets, SDRFPT-Net demonstrates powerful detection capability and robustness under various environmental conditions. This is attributed to the collaborative work of our three innovative modules: the spectral hierarchical perception architecture provides a solid foundation for multi-modal feature extraction; the spectral adaptive recursive fusion module achieves deep interaction and efficient fusion; and the spectral adaptive target perception enhancement module further improves target region feature representation. The organic combination of these three modules enables SDRFPT-Net to achieve excellent multispectral object detection performance while maintaining low computational complexity.

4.4. Ablation Studies

To verify the effectiveness of each innovative module in SDRFPT-Net, we conducted systematic ablation experiments on the FLIR-aligned dataset, which can more intuitively reflect the effectiveness of our algorithm. These experiments aim to evaluate the contribution of each component to the overall performance of the network and validate the rationality of our proposed design scheme.

4.4.1. Baseline Model Comparison.

First, we established a baseline model, then gradually added each core component to evaluate the contribution of each module. Table 4 shows the performance comparison of different component combinations.

Table 4. Impact of different attention combinations on detection performance. The table compares the effects of Self Attention, Cross-modal Attention, and Channel Attention in various combinations. The best results are highlighted in bold.

ID	SHPA	SRFM	STPEM	mAP50	mAP50:95
A1	✓			0.701	0.354
A2	✓	✓		0.775	0.373

A3	✓	✓	✓	**0.785**	**0.426**

From the results in Table 4, it is evident that each component we proposed contributes significantly to detection performance. The model based on the Spectral Hierarchical Perception Architecture (SHPA) (A1) achieves an mAP50 of 0.701. After adding the Spectral Adaptive Recursive Fusion Module (SRFM) (A2), the mAP50 increases to 0.775, a relative improvement of 10.6%. With the further addition of the Spectral Adaptive Target Perception Enhancement Module (STPEM) (A3), mAP50 and mAP50:95 reach 0.785 and 0.426 respectively, with mAP50:95 showing a particularly significant 14.2% relative improvement over A2, indicating that STPEM greatly enhances the model's localization accuracy under stricter detection standards.

4.4.2. Ablation Experiments on Hybrid Attention Mechanism

To evaluate the effectiveness of various attention mechanisms in the hybrid attention mechanism, we designed a series of comparative experiments, with results shown in Table 5.

Table 5. Impact of different attention combinations on detection performance. The table compares the effects of Self Attention, Cross-modal Attention, and Channel Attention in various combinations. The best results are highlighted in bold.

ID	Self-attention	Cross-attention	Channel-attention	mAP50	mAP50:95
B1	✓			0.776	0.408
B2		✓		0.749	0.372
B3			✓	0.730	0.384
B4	✓	✓		0.774	0.424
B5	✓		✓	0.763	0.409
B6		✓	✓	0.729	0.362
B7	✓	✓	✓	**0.785**	**0.426**

The results show that different types of attention mechanisms have varying impacts on model performance. When used individually, the self-attention mechanism (B1) performs best with an mAP50 of 0.776 and mAP50:95 of 0.408, indicating that capturing spatial dependencies within modalities is critical for object detection. Although cross-modal attention (B2) and channel attention (B3) show slightly inferior performance when used alone, they provide feature enhancement capabilities in different dimensions.

In combinations of two attention mechanisms, the combination of self-attention and cross-modal attention (B4) performs best, with mAP50:95 reaching 0.424, approaching the performance of the complete model. The complete combination of three attention mechanisms (B7) achieves the best performance, confirming the rationality of the hybrid attention mechanism design, which can comprehensively capture the complex relationships in multi-modal data.

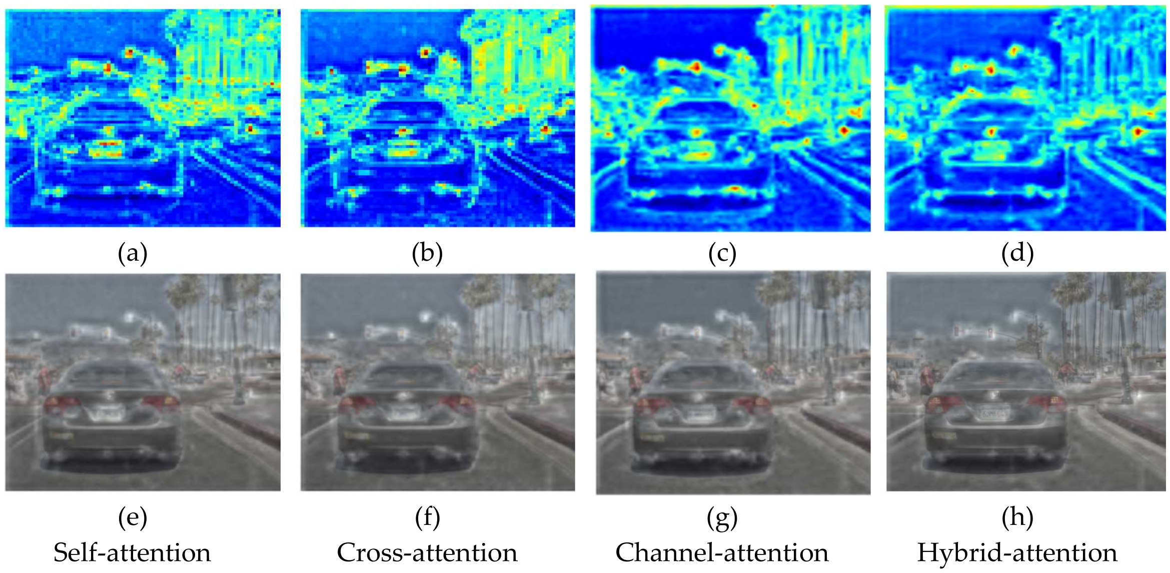
To gain a deeper understanding of the role of different attention mechanisms in multispectral object detection, we analyzed the visualization results of self-attention, cross-modal attention, and channel attention on the P3 feature layer.

From Figure 13, it can be observed that the feature maps of single attention mechanisms present different attention patterns:

1. Self-attention mechanism (B1): Mainly focuses on target contours and edge information, effectively capturing spatial contextual relationships, with strong response to target boundaries, helping to improve localization accuracy;

2. Cross-modal attention mechanism (B2): Presents overall attention to target areas, integrating complementary information from RGB and infrared modalities, but with relatively weak background suppression capability;

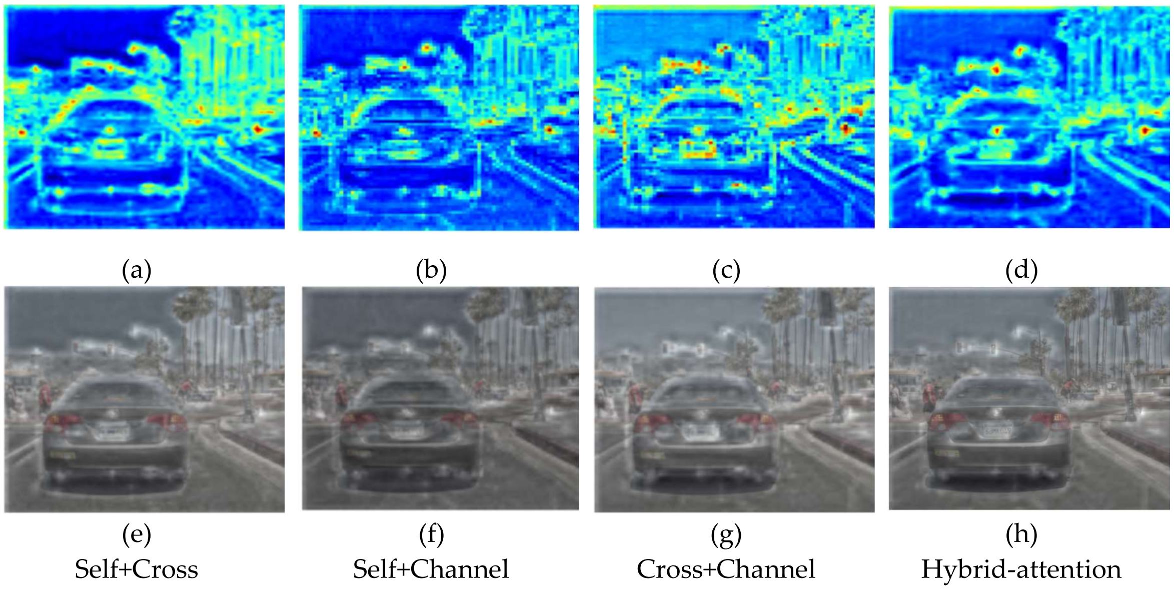
3. Channel attention mechanism (B3): Demonstrates selective enhancement of specific semantic information, highlighting important feature channels, with strong response to specific parts of targets, improving the discriminability of feature representation.


(a) (b) (c) (d)

(e) Self-attention (f) Cross-attention (g) Channel-attention (h) Hybrid-attention

Figure 13. Comparative impact of different attention mechanisms on the P3 feature layer (high-resolution features) in SDRFPT-Net. The top row (a-d) presents feature activation maps, while the bottom row (e-h) shows the corresponding original image heatmap overlay effects, demonstrating the differences in feature attention patterns. Self-attention (a,e) focuses on target contours and edge information; Cross-attention (b,f) presents overall attention to target areas with complementary information from RGB and IR modalities; Channel-attention (c,g) demonstrates selective enhancement of specific semantic information; Hybrid-attention (d,h) combines the advantages of all three mechanisms for optimal feature representation.

Furthermore, we conducted a comparative analysis of the visualization effects of dual attention mechanisms versus the full attention mechanism on the P3 feature layer, as shown in Figure 14.


(a) (b) (c) (d)

(e) Self+Cross (f) Self+Channel (g) Cross+Channel (h) Hybrid-attention

Figure 14. Representational differences between dual attention mechanism combinations and the complete triple attention mechanism on the P3 feature layer. The top row (a-d) shows feature activation maps, while the bottom row (e-h) shows original image heatmap overlay effects, revealing the complementarity and synergistic effects of different attention combinations. Self+Cross attention (a,e) simultaneously possesses excellent boundary localization and target region representation; Self+Channel attention (b,f) enhances specific semantic features

while preserving boundary information; Cross+Channel attention (c,g) enhances channel representation based on multi-modal fusion but lacks spatial context; Hybrid-attention (d,h) achieves the most comprehensive and effective feature representation through synergistic integration of all three mechanisms.

Through the visualization comparative analysis in Figure 14, we observe that the feature maps of dual attention mechanisms present complex and differentiated feature representations:

1. Self-attention + Cross-modal attention (B4): The feature map simultaneously possesses excellent boundary localization capability and overall target region representation capability. The heatmap shows precise response to target regions with significant background suppression effect. This combination fully leverages the complementary advantages of self-attention in spatial modeling and cross-modal attention in multi-modal fusion, enabling it to reach 0.424 in mAP50:95, approaching the performance of the full attention mechanism.

2. Self-attention + Channel attention (B5): The feature map enhances the representation of specific semantic features while preserving target boundary information. The heatmap shows strong response to key parts of targets, enabling the model to better distinguish different categories of targets, achieving 0.409 in mAP50:95, outperforming any single attention mechanism.

3. Cross-modal attention + Channel attention (B6): The feature map enhances specific channel representation based on multi-modal fusion, but lacks the spatial context modeling capability of self-attention. The heatmap shows some response to target regions, but boundaries are not clear enough and background suppression effect is relatively weak, which explains its relatively lower performance.

Although dual attention mechanisms (especially self-attention + cross-modal attention) can improve feature representation capability to some extent, they cannot completely replace the comprehensive advantages of the full attention mechanism.

As shown in Figures 13d and 14d, the full attention mechanism, through the synergistic effect of three attention mechanisms, shows the most precise and strong response to target regions in the heatmap, with clear boundaries and optimal background suppression effect, achieving an organic unification of spatial context modeling, multi-modal information fusion, and channel feature enhancement, obtaining optimal performance in multispectral object detection tasks.

Based on the above experiments and visualization analysis, we verified the effectiveness of the proposed hybrid attention mechanism. The results show that, despite the advantages of single and dual attention mechanisms, the complete combination of three attention mechanisms achieves optimal performance across all evaluation metrics. This confirms the rationality of our proposed "spatial-modal-channel" multi-dimensional attention framework, which creates an efficient synergistic mechanism through self-attention capturing spatial contextual relationships, cross-modal attention fusing complementary information, and channel attention selectively enhancing key features. This multi-dimensional feature enhancement strategy provides a new feature fusion paradigm for multispectral object detection, offering valuable reference for research in related fields.

4.4.3. Ablation Experiments on Spectral Hierarchical Recursive Progressive Fusion Strategy

To verify the effectiveness of the spectral fusion strategy, we conducted ablation experiments from two aspects: fusion position selection and recursive progression iterations.

Impact of introducing fusion positions. Multi-scale feature fusion is a key link in multispectral object detection, and effective fusion of features at different scales has a decisive impact on model performance. This section explores the impact of fusion positions and fusion strategies on detection performance through ablation experiments and visualization analysis.

To systematically study the impact of fusion positions on model performance, we designed a series of ablation experiments, as shown in Table 6. The experiments started from the baseline model (C1, using simple addition fusion at all scales), progressively applying our proposed innovative fusion modules (fusion mechanism combining SRFM and STPEM) at different scales, and finally evaluating the effect of comprehensive application of advanced fusion strategies.

Table 6. Ablation experiments on fusion positions. The table shows the impact of applying advanced fusion modules at different feature scales. P3/8, P4/16, and P5/32 represent feature maps at different scales, with numbers indicating the downsampling factor relative to the input image. The best results are highlighted in bold.

ID	P3/8	P4/16	P5/32	mAP50	mAP50:95
C1	Add	Add	Add	0.701	0.354
C2	SRFM+STPEM	Add	Add	0.769	0.404
C3	SRFM+STPEM	SRFM+STPEM	Add	0.776	0.410
C4	SRFM+STPEM	SRFM+STPEM	SRFM+STPEM	**0.785**	**0.426**

As shown in Table 6, with the increase in application positions of innovative fusion modules, model performance progressively improves. The baseline model (C1) only uses simple addition fusion at all feature scales, with an mAP50 of 0.701. When applying SRFM and STPEM modules at the P3/8 scale (C2), performance significantly improves to an mAP50 of 0.769. With further application of advanced fusion at the P4/16 scale (C3), mAP50 increases to 0.776. Finally, when the complete fusion strategy is applied to all three scales (C4), performance reaches optimal levels with an mAP50 of 0.785 and mAP50:95 of 0.426.

To intuitively understand the effect differences between different fusion strategies, we visualized and compared feature maps of simple addition fusion and advanced fusion strategies (SRFM+STPEM) at three scales: P3/8, P4/16, and P5/32, as shown in Figure 15.

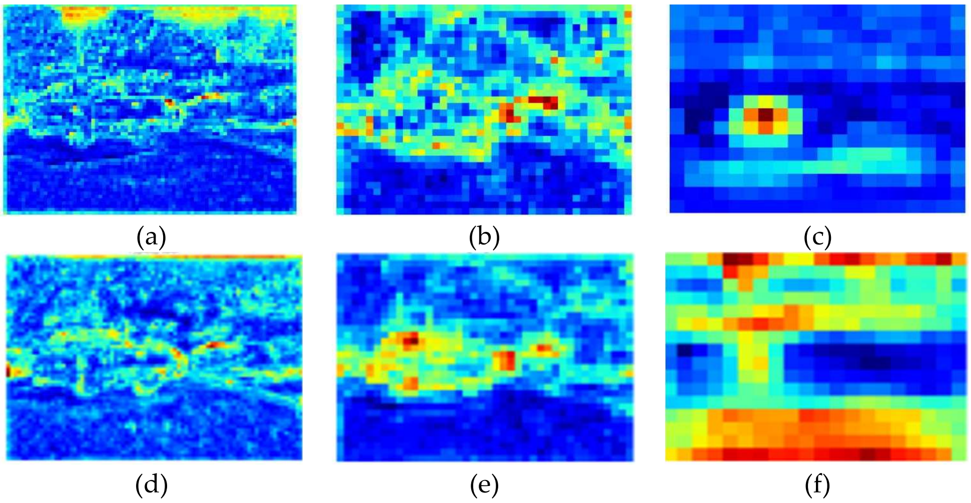


Figure 15. Comparison between simple addition fusion and innovative fusion strategies (SRFM+STPEM) on three feature scale layers. The upper row (a,b,c) presents traditional simple addition fusion at different scales: P3/8 high-resolution layer (a) shows dispersed activation with insufficient target-background differentiation; P4/16 medium-resolution layer (b) has some response to vehicle areas but with blurred boundaries; P5/32 low-resolution layer (c) only has rough response to the central vehicle. The lower row (d,e,f) shows feature maps of the innovative fusion strategy: P3/8 layer (d) provides clearer vehicle contour representation with precise edge localization; P4/16 layer (e) shows more concentrated target area activation; P5/32 layer (f) preserves richer scene semantic information while enhancing central target representation.

To gain a deeper understanding of the performance differences between different fusion strategies, we conducted systematic visualization comparisons at different scales (P3/8, P4/16, P5/32).

1. At the P3/8 scale, simple addition fusion presents dispersed activation patterns with insufficient target-background differentiation, especially with suboptimal activation intensity for small vehicles; whereas the feature map generated by the SRFM+STPEM fusion strategy possesses more precise target localization capability and boundary representation, with significantly improved background suppression effect and activation intensity distribution more concentrated on target regions, effectively enhancing small target detection performance.

2. The comparison of P4/16 scale feature maps shows that although simple addition fusion can capture medium target positions, activation is not prominent enough and background noise interference exists; in contrast, the advanced fusion strategy produces more concentrated activation areas with higher target-background contrast and clearer boundaries between vehicles. As an intermediate resolution feature map, P4 (40×40) demonstrates superior structured representation and background suppression capability under the advanced fusion strategy.

3. At the P5/32 scale, rough semantic information generated by simple addition fusion makes it difficult to distinguish main vehicle targets; whereas the advanced fusion strategy can better capture overall scene semantics, accurately represent main vehicle targets, and effectively suppress background interference. Although the P5 feature map has the lowest resolution (20×20), it has the largest receptive field, and the advanced fusion strategy fully leverages its advantages in large target detection and scene understanding.

Through comparative analysis, we observed three key synergistic effects of multi-scale fusion:

1. Complementarity enhancement: the advanced fusion strategy makes features at different scales complementary, with P3 focusing on details and small targets, P4 processing medium targets, and P5 capturing large-scale structures and semantic information;

2. Information flow optimization: features at different scales mutually enhance each other, with semantic information guiding small target detection and detail information precisely locating large target boundaries;

3. Noise suppression capability: the advanced fusion strategy demonstrates superior background noise suppression capability at all scales, effectively reducing false detections.

Impact of recursive iteration count. The recursive progression mechanism is a key strategy in our proposed SDRFPT-Net model, which can further enhance feature representation capability through multiple recursive progressive fusions. To explore the optimal number of recursive iterations, we designed a series of experiments to observe model performance changes by varying the number of iterations. Table 7 shows the impact of iteration count on model performance.

Table 7. Impact of different iteration counts of the recursive progressive fusion strategy on model detection performance. The experiment compares performance with recursive depths from 1 to 5 iterations (D1-D5), evaluating detection accuracy using mAP50 and mAP50:95 metrics. The best results are highlighted in bold.

ID	times	mAP50	mAP50:95
D1	1	0.769	0.395
D2	2	0.783	0.418
D3	3	**0.785**	**0.426**
D4	4	0.783	0.400
D5	5	0.761	0.417

The experimental results show that the number of recursive iterations significantly affects model performance. When the iteration count is 1 (D1), the model achieves an mAP50 of 0.769, indicating that even a single round of iteration can provide effective feature fusion. As the iteration count increases to 2 (D2) and 3 (D3), performance continues to improve, reaching mAP50 of 0.783 and 0.785, and mAP50:95 of 0.418 and 0.426 respectively. However, when the iteration count further increases to 4 (D4) and 5 (D5), performance begins to decline, with the mAP50 of 5 iterations significantly dropping to 0.761.

To more intuitively understand the impact of iteration count on feature representation, we conducted visualization analysis of feature maps at three feature scales—P3, P4, and P5—with different iteration counts, as shown in Figure 16.

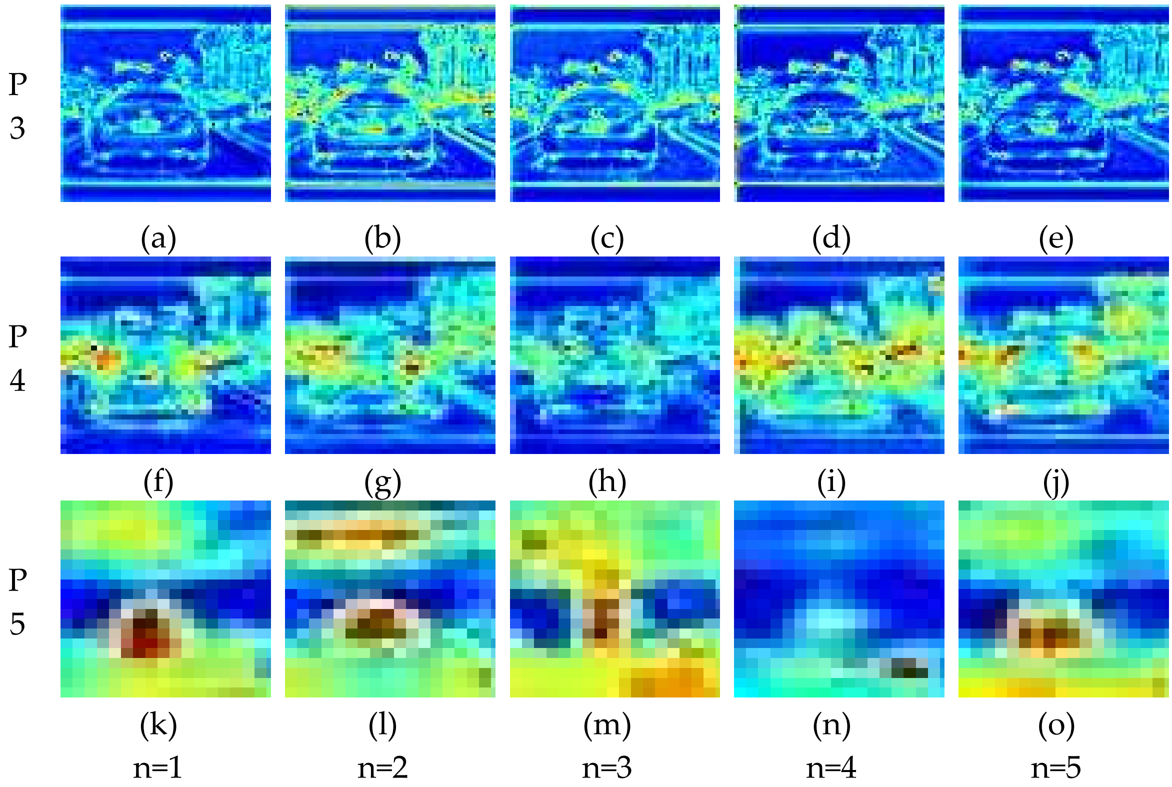


Figure 16. Impact of iteration counts (n=1 to n=5) in the recursive progressive fusion strategy on three feature scale layers of SDRFPT-Net. By comparing the evolution within the same row, changes in features with recursive depth can be observed; by comparing different rows, response characteristics at different scales can be understood. The P3 high-resolution layer (first row) shows feature representation gradually evolving from initial dispersed response (n=1) to more focused target contours (n=2,3), with clearer boundaries and stronger background suppression, but experiencing over-smoothing at n=4,5. The P4 medium-resolution layer (second row) shows optimal target-background differentiation at n=3, followed by feature response diffusion at n=4,5. The P5 low-resolution layer (third row) presents the most significant changes, achieving highly structured representation at n=3 that clearly distinguishes main scene elements, while showing obvious degradation at n=4 and n=5.

Through visualization, we observed the following feature evolution patterns:

1. P3 feature layer (high resolution): As the iteration count increases, the feature map gradually evolves from initial dispersed response (n=1) to more focused target representation (n=2,3), with clearer boundaries and stronger background suppression effect. However, when the iteration count reaches 4 and 5, over-smoothing phenomena begin to appear, with some loss of boundary details.

2. P4 feature layer (medium resolution): At n=1, the feature map has basic response to targets but is not focused enough. After 2-3 rounds of iteration, the activation intensity of target areas significantly increases, improving target differentiation. Continuing to increase the iteration count to 4-5 rounds, feature response begins to diffuse, reducing precise localization capability.

3. P5 feature layer (low resolution): This layer demonstrates the most obvious evolution trend, gradually developing from blurred response at n=1 to highly structured representation at n=3 that can clearly distinguish main targets. However, obvious signs of overfitting appear at n=4 and n=5, with feature maps becoming overly smoothed and target representation degrading.

These observations reveal the working mechanism of recursive progressive fusion: moderate iteration count (n=3) can achieve progressive optimization of features through multiple rounds of interactive fusion of complementary information from different modalities, enhancing target feature representation and suppressing background interference. However, excessive iteration count may lead to "over-fusion" of features, i.e., the model overfits specific patterns in the training data, losing generalization capability.

Combining quantitative and visualization analysis results, we determined n=3 as the optimal iteration count, achieving the best balance between feature enhancement and computational efficiency. This finding is also consistent with similar observations in other research areas, such as the optimal unfolding steps in recurrent neural networks and the optimal iteration count in message passing neural networks, where similar "performance saturation points" exist.

Through the above ablation experiments, we verified the effectiveness and optimal configuration of each core component of SDRFPT-Net. The results show that the spectral hierarchical perception architecture (SHPA), the complete combination of three attention mechanisms, the full-scale advanced fusion strategy, and three rounds of recursive progressive fusion collectively contribute to the model's superior performance. The rationality of these design choices is not only validated through quantitative metrics but also intuitively explained through feature visualization, providing new ideas for multispectral object detection research.

5. Discussion

SDRFPT-Net demonstrated superior performance across VEDAI, FLIR-aligned, and LLVIP datasets, effectively integrating complementary information from visible and infrared domains with significant improvements over single-modality methods (8.0% in mAP50 on FLIR-aligned and 9.5% in mAP50:95 on LLVIP). The proposed spectral recursive fusion mechanism represents a computationally efficient innovation through cyclic weight reuse with parameter sharing, with ablation experiments confirming three recursive iterations as optimal for the progressive "feature distillation" process. Feature visualization validated the synergistic effects across different scales (P3 capturing details and small targets, P4 providing better target-background differentiation, P5 retaining richer semantic information) and the complementary functions of hybrid attention components (self-attention for spatial context, cross-modal attention for inter-modality exchange, channel attention for semantic enhancement).

Despite these advantages, SDRFPT-Net faces limitations in scenes with densely arranged or occluded targets, exhibits slower inference speed compared to some single-modality detectors, and lacks adaptive adjustment capabilities in its fixed three-round recursive iteration. Future research directions include optimizing dense target detection through specialized loss functions, improving inference speed via model pruning and quantization, designing adaptive recursive mechanisms, extending the framework to incorporate more spectral modalities, and exploring integration with Vision Transformers and other recent architectures.

6. Conclusions

This paper presents SDRFPT-Net, a novel architecture for multispectral object detection that integrates three key innovative modules: Spectral Hierarchical Perception Architecture (SHPA), which adopts a dual-stream separated structure with independently parameterized feature extraction paths for visible and infrared modalities; Spectral Recursive Fusion Module (SRFM), which combines hybrid attention mechanisms with recursive progressive fusion strategies; and Spectral Target Perception Enhancement Module (STPEM), which adaptively enhances target region representations and suppresses background interference. Experimental validations on three benchmark datasets demonstrate the effectiveness of the proposed method, with SDRFPT-Net achieving significant improvements over state-of-the-art methods: on VEDAI (2.5% in mAP50, 5.4% in mAP50:95), FLIR-aligned (11.5% in mAP50), and LLVIP (9.5% in mAP50:95). The outstanding performance on the VEDAI dataset particularly validates the algorithm's capacity to handle the unique challenges of remote sensing imagery, including variable scales, complex terrain backgrounds, and diverse viewing angles. Systematic ablation experiments further verified the contribution of each innovative module and its optimal configuration.

The main contributions of this research include designing a dual-stream architecture adapted to different spectral domain characteristics, proposing a computationally efficient recursive fusion mechanism, developing a target perception enhancement module, and demonstrating superior performance under various environmental conditions. SDRFPT-Net provides an efficient and robust solution for multispectral object detection in remote sensing applications, particularly suitable for processing data from drones, aircraft, and satellite platforms under complex environmental conditions. The network offers new insights for multi-modal information fusion with significant value for remote sensing image interpretation, intelligent surveillance, autonomous driving, and other application domains. The approach represents an important advancement for earth observation technologies, with potential applications in environmental monitoring, resource investigation, urban planning, and disaster management.

Author Contributions: Conceptualization, P.Z. and X.S.; methodology, P.Z. and B.S.; validation, X.S., B.S. and R.G.; formal analysis, P.Z.; investigation, P.Z. and X.S.; resources, P.Z.; data curation, B.S.; software implementation, P.Z.; writing—original draft preparation, P.Z.; writing—review and editing, X.S.; visualization, P.Z. and Z.D.; supervision, X.S. and S.S.; project administration, X.S. P.Z. was responsible for the primary algorithm design and experimental implementation. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Hunan Provincial Postgraduate Research Innovation Programme, grant number XJZH2024033.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the relevance of data to individual privacy.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object Detection in Optical Remote Sensing Images: A Survey and a New Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307, doi:10.1016/j.isprsjprs.2019.11.023.
2. Zhang, C.; Chen, B.Y.; Lam, W.H.K.; Ho, H.W.; Shi, X.; Yang, X.; Ma, W.; Wong, S.C.; Chow, A.H.F. Vehicle Re-Identification for Lane-Level Travel Time Estimations on Congested Urban Road Networks Using Video Images. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 12877–12893, doi:10.1109/TITS.2021.3118206.
3. Feng, D.; Haase-Schutz, C.; Rosenbaum, L.; Hertlein, H.; Glaser, C.; Timm, F.; Wiesbeck, W.; Dietmayer, K. Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 1341–1360, doi:10.1109/TITS.2020.2972974.
4. Li, C.; Cong, R.; Hou, J.; Zhang, S.; Qian, Y.; Kwong, S. Nested Network with Two-Stream Pyramid for Salient Object Detection in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9156–9166, doi:10.1109/TGRS.2019.2925070.
5. Yun Liu, X.-Y.Z. SAMNet: Stereoscopically Attentive Multi-Scale Network for Lightweight Salient Object Detection. *IEEE Trans. Image Process. : Publ. IEEE Signal Process. Soc.* 2021, *Vol.30*, 3804–3814.
6. Ren, X.; Bai, Y.; Liu, G.; Zhang, P. YOLO-Lite: An Efficient Lightweight Network for SAR Ship Detection. *Remote Sens.* 2023, *Vol.15*, 3771.
7. Qingyun, F.; Zhaokui, W. Cross-Modality Attentive Feature Fusion for Object Detection in Multispectral Remote Sensing Imagery. *Pattern Recognit.* **2022**, *130*, 108786, doi:10.1016/j.patcog.2022.108786.
8. Zhang, T.; Wu, H.; Liu, Y.; Peng, L.; Yang, C.; Peng, Z. Infrared Small Target Detection Based on Non-Convex Optimization with Lp-Norm Constraint. *Remote Sens.* **2019**, *11*, 559, doi:10.3390/rs11050559.
9. Pang, S.; Ge, J.; Hu, L.; Guo, K.; Zheng, Y.; Zheng, C.; Zhang, W.; Liang, J. RTV-SIFT: Harnessing Structure Information for Robust Optical and SAR Image Registration. *Remote Sensing* **2023**, *15*, 4476, doi:10.3390/rs15184476.
10. Song, K.; Bao, Y.; Wang, H.; Huang, L.; Yan, Y. A Potential Vision-Based Measurements Technology: Information Flow Fusion Detection Method Using RGB-Thermal Infrared Images. *IEEE Trans. Instrum. Meas.* 2023, *Vol.72*, 1–13.

11. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral Deep Neural Networks for Pedestrian Detection 2016.
12. Wagner, J.; Fischer, V.; Herman, M.; Behnke, S. Multispectral Pedestrian Detection Using Deep Fusion Convolutional Neural Networks. *Comput. Intell.* **2016**.
13. Konig, D.; Adam, M.; Jarvers, C.; Layher, G.; Neumann, H.; Teutsch, M. Fully Convolutional Region Proposal Networks for Multispectral Person Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); IEEE: Honolulu, HI, USA, July 2017; pp. 243–250.
14. Zhang, Y.; Yu, H.; He, Y.; Wang, X.; Yang, W. Illumination-Guided RGBT Object Detection with Inter- and Intra-Modality Fusion. *IEEE Trans. Instrum. Meas.* 2023, *Vol.72*, 1–13.
15. Zhou, W.; Zhu, Y.; Lei, J.; Wan, J.; Yu, L. CCAFNet: Crossflow and Cross-Scale Adaptive Fusion Network for Detecting Salient Objects in RGB-D Images. *IEEE Trans. Multimedia* **2022**, *24*, 2192–2204, doi:10.1109/TMM.2021.3077767.
16. Zhi-she, W.; Feng-bao, Y.; Zhi-hao, P.; Lei, C.; Li-e, J. Multi-sensor image enhanced fusion algorithm based on NSST and top-hat transformation. *Optik* **2015**, *126*, 4184–4190, doi:10.1016/j.ijleo.2015.08.118.
17. Zhang, Q.; Liu, Y.; Blum, R.S.; Han, J.; Tao, D. Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: a review. *Inf. Fusion* **2018**, *40*, 57–75, doi:10.1016/j.inffus.2017.06.005.
18. Ma, J.; Zhou, Z.; Wang, B.; Zong, H. Infrared and Visible Image Fusion Based on Visual Saliency Map and Weighted Least Square Optimization. *Infrared Phys. Technol.* **2017**, *82*, 8–17, doi:10.1016/j.infrared.2017.02.005.
19. Qingyun, F.; Dapeng, H.; Zhaokui, W. Cross-modality fusion transformer for multispectral object detection 2022.
20. Chen, Y.-T.; Shi, J.; Ye, Z.; Mertz, C.; Ramanan, D.; Kong, S. Multimodal object detection via probabilistic ensembling 2022.
21. Zhang, H.; Xu, H.; Tian, X.; Jiang, J.; Ma, J. Image Fusion Meets Deep Learning: A Survey and Perspective. *Inf. Fusion* **2021**, *76*, 323–336, doi:10.1016/j.inffus.2021.06.008.
22. Fu, Y.; Wu, X.-J.; Durrani, T. Image Fusion Based on Generative Adversarial Network Consistent with Perception. *Inf. Fusion* **2021**, *72*, 110–125, doi:10.1016/j.inffus.2021.02.019.
23. Li, J.; Huo, H.; Li, C.; Wang, R.; Sui, C.; Liu, Z. Multigrained Attention Network for Infrared and Visible Image Fusion. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–12, doi:10.1109/TIM.2020.3029360.
24. Wang, Z.; Wu, Y.; Wang, J.; Xu, J.; Shao, W. Res2Fusion: Infrared and Visible Image Fusion Based on Dense Res2net and Double Nonlocal Attention Models. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–12, doi:10.1109/TIM.2021.3139654.
25. Zhang, X.; Wang, J.; Wang, T.; Jiang, R. Hierarchical Feature Fusion with Mixed Convolution Attention for Single Image Dehazing. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 510–522, doi:10.1109/TCSVT.2021.3067062.
26. Li, J.; Huo, H.; Li, C.; Wang, R.; Feng, Q. AttentionFGAN: Infrared and Visible Image Fusion Using Attention-Based Generative Adversarial Networks. *IEEE Trans. Multimedia* **2021**, *23*, 1383–1396, doi:10.1109/TMM.2020.2997127.
27. Zhao, Y.; Zhao, L.; Xiong, B.; Kuang, G. Attention Receptive Pyramid Network for Ship Detection in SAR Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 2020, *Vol.13*, 2738–2756.
28. Redmon, J.; Farhadi, A. YOLO9000: better, faster, stronger 2016.
29. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Las Vegas, NV, USA, June 2016; pp. 779–788.
30. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: real-time end-to-end object detection 2024.
31. Chang, Y.-L.; Anagaw, A.; Chang, L.; Wang, Y.C.; Hsiao, C.-Y.; Lee, W.-H. Ship Detection Based on YOLOv2 for SAR Imagery. *Remote Sens.* 2019, *Vol.11*, 786.

32. Etten, A.V. You Only Look Twice: Rapid Multi-Scale Object Detection In Satellite Imagery. *Comput. Vis. Pattern Recognit.* **2018**.
33. Sharma, M.; Dhanaraj, M.; Karnam, S.; Chachlakis, D.G.; Ptucha, R.; Markopoulos, P.P.; Saber, E. YOLOrs: object detection in multimodal remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 2021, *Vol.14*, 1497–1508.
34. Chen, L. Improved YOLOv3 Based on Attention Mechanism for Fast and Accurate Ship Detection in Optical Remote Sensing Images. *Remote Sens.* 2021, *Vol.13*, 660.
35. Guo, Y.; Chen, S.; Zhan, R.; Wang, W.; Zhang, J. *SAR ship detection based on YOLOv5 using CBAM and BiFPN*; College of Electronic Science and Engineering, National University of Defense Technology, Changsha, 410073, China, 2022;
36. Wang, Z.; Chen, Y.; Shao, W.; Li, H.; Zhang, L. SwinFuse: A Residual Swin Transformer Fusion Network for Infrared and Visible Images. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–12, doi:10.1109/TIM.2022.3191664.
37. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery: a small target detection benchmark(article). *J. Visual Commun. Image Represent.* 2016, *Vol.34*, 187–203.
38. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery : a small target detection benchmark. *Journal of Visual Communication and Image Representation* **2016**, *34*, 187–203, doi:10.1016/j.jvcir.2015.11.002.
39. Zhang, H.; Fromont, E.; Lefevre, S.; Avignon, B. Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks 2020.
40. Jia, X.; Zhu, C.; Li, M.; Tang, W.; Liu, S.; Zhou, W. LLVIP: A Visible-Infrared Paired Dataset for Low-Light Vision 2023.
41. Flir, T. Free FLIR Thermal Dataset for Algorithm Training 2018.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.