

Article

JFDet: Joint Fusion and Detection for Multimodal Remote Sensing Imagery

Wenhao Xu * and You Yang 

School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China; yangyou@hust.edu.cn

* Correspondence: mlxtj@hust.edu.cn

Highlights

What are the main findings?

- A Joint Fusion and Detection Network (JFDet) is proposed, which realizes the joint optimization of low-level image fusion and high-level object detection through a dual-loss feedback strategy (fusion loss + detection loss), effectively addressing the “task mismatch” issue between the two in traditional methods. On the VEDAI dataset, its mean Average Precision (mAP) reaches 79.6%, which is 1.7% higher than that of the current state-of-the-art method MMFDet (77.9%). On the FLIR-ADAS dataset, its mAP reaches 76.3%, and the detection accuracy of key categories (person, car, bicycle) ranks first among the compared methods.
- A collaborative architecture consisting of a Gradient-Enhanced Residual Module (GERM) and detection enhancement modules (SOCA + MCFE) is designed: GERM combines dense feature connections with Fourier phase enhancement to strengthen the preservation of edge structures and fine-grained texture details; SOCA captures higher-order feature dependencies, and MCFE expands the receptive field. The combination of these two modules significantly improves the detection robustness of small and variably scaled targets in remote sensing imagery.

What are the implication of the main findings?

- This study achieves high radiometric consistency fusion of optical and infrared images, overcoming the limitations of single-modality data (e.g., degradation of visible light under low-light or adverse weather conditions, and the lack of texture information in infrared data). It provides a reliable detection scheme for scenarios with high timeliness requirements and complex environments, such as disaster monitoring and emergency search and rescue, helping to improve the efficiency of emergency decision-making.
- The modular design (GERM, SOCA, MCFE) and adaptive training strategy exhibit good compatibility, which can be adapted to different detection frameworks such as Yolov5s and Faster RCNN. Additionally, they provide high-quality multimodal feature support for downstream remote sensing tasks like semantic segmentation and change detection, expanding the application boundaries of multimodal fusion technology.



Academic Editor: Salah Bourennane

Received: 25 November 2025

Revised: 25 December 2025

Accepted: 30 December 2025

Published: 5 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Abstract

Multimodal remote sensing imagery, such as visible and infrared data, offers crucial complementary information that is vital for time-sensitive emergency applications like search and rescue or disaster monitoring, where robust detection under adverse conditions is essential. However, existing methods’ object detection performance is often suboptimal due to task-independent fusion and inherent modality inconsistency. To address this issue, we propose a joint fusion and detection approach for multimodal remote sensing imagery

(JFDet). First, a gradient-enhanced residual module (GERM) is introduced to combine dense feature connections with gradient residual pathways, effectively enhancing structural representation and fine-grained texture details in fused images. For robust detection, we introduce a second-order channel attention (SOCA) mechanism and design a multi-scale contextual feature-encoding (MCFE) module to capture higher-order semantic dependencies, enrich multi-scale contextual information, and thereby improve the recognition of small and variably scaled objects. Furthermore, a dual-loss feedback strategy propagates detection loss to the fusion network, enabling adaptive synergy between low-level fusion and high-level detection. Experiments on the VEDAI and FLIR-ADAS datasets demonstrate that the proposed detection-driven fusion framework significantly improves both fusion quality and detection accuracy compared with state-of-the-art methods, highlighting its effectiveness and high potential for mission-critical multimodal remote sensing and time-sensitive application.

Keywords: multimodal remote sensing; object detection; image fusion; deep learning

1. Introduction

Remote sensing object detection (RSOD) plays a pivotal role in both fundamental research and practical applications, enabling the automatic identification and precise localization of ground targets within vast remote sensing imagery [1,2]. This technology provides essential data support for critical tasks such as land-use analysis, urban planning, environmental monitoring, and disaster emergency management. However, relying solely on a single-modality sensor presents inherent limitations: visible images offer rich texture and color details but suffer severe degradation under low-light or adverse weather conditions [3], while infrared (IR) images, which capture thermal radiation, perform well under weak illumination but inherently lack fine-grained texture and color information [4,5]. This single-modality deficit is acutely felt in emergency scenarios. For instance, during earthquake disasters, visible light may be reduced by smoke, resulting in insufficient illumination and hindering the clear identification of trapped persons and infrastructure; IR images can be used to locate heat sources but often fail to effectively distinguish target types. Similarly, during floods, visible light is reduced by reflections from the water, while IR images lack the structural characteristics needed for precise damage assessment. To overcome these constraints and meet the growing demand for precision and robustness, fusing complementary data from IR and visible modalities has emerged as a crucial solution for advancing remote sensing information processing [6]. Multimodal fusion offers invaluable support in the context of disaster emergency response: in fire scenes, it integrates two modalities to allow responders to quickly locate fire sources and trapped persons via IR features while assessing combustible types and building damage in visible images. In post-typhoon image assessment, fused images clearly show the structural integrity of roads and thermal anomalies in power facilities. This comprehensive information support significantly improves the efficiency and accuracy of emergency decision-making, conserving valuable time to save lives and reduce property losses [7].

The field of image fusion has seen significant development and can be broadly categorized into pixel-, feature-, and decision-level fusion based on the integration stage [8]. In particular, pixel-level infrared–visible fusion has garnered extensive attention, with methods evolving from traditional techniques (e.g., multi-scale transform, sparse representation) to sophisticated deep learning models (e.g., auto-encoder (AE)-based [9], convolutional neural network (CNN)-based [10], and generative adversarial network (GAN)-based ap-

proaches) [11,12]. In the AE category, Li et al. [9] proposed DenseFuse, enhancing feature extraction via densely connected blocks, but manual fusion strategies limited their model's performance. Subsequent work [13] replaced manual fusion with convolutional operations, improving performance at the cost of increased model complexity. Wang et al. [10] introduced a multi-scale encoder–decoder with nested connections and attention mechanisms, capturing both global and local features. In CNN-based methods, Long et al. [11] proposed a dense residual network for feature reuse. Xu et al. [14] designed U2Fusion to enhance mutual benefits across tasks, and Zhang et al. [15] developed a universal loss function to optimize gradient and intensity ratios. Li et al. [16] incorporated Transformers to combine global and local features, improving fusion results but increasing parameters. In GAN-based methods, Ma et al. [12] first applied GANs for adversarial fusion, while Zhou et al. [17] introduced a dual-discriminator to preserve features from both modalities, increasing complexity. Wang et al. [18] proposed FusionGRAM, using multi-scale attention to focus on salient regions and reduce redundant interference. While these approaches have successfully enhanced the visual quality and information content of the fused images, a key macroscopic challenge remains unaddressed: the optimization goal of most existing fusion methods is limited to low-level visual quality metrics, e.g., preserving intensity or structure. Critically, these methods typically operate independently of high-level downstream tasks like object detection. This lack of semantic awareness creates a significant Task Mismatch, in which visually compelling fusion results often fail to translate into substantial performance gains for object detection [19]. The absence of a mechanism to explicitly optimize the fusion process based on semantic relevance and detection loss feedback severely limits the full potential of multimodal data in real-world remote sensing applications.

To address the challenge of task mismatch, where low-level fusion objectives fail to benefit high-level object detection in remote sensing imagery sufficiently, we propose the Joint Fusion and Detection network (JFDet), a dual-driven multimodal framework that tightly integrates image fusion and object detection. This framework explicitly aligns the image fusion process with the object detection objective, ensuring that feature generation is driven by downstream detection performance rather than by visual quality alone. Our contributions are summarized below.

- We propose a novel dual-Loss feedback strategy that achieves adaptive synergy between low-level fusion and high-level detection tasks. The fusion process is explicitly guided by feedback from the detection loss, ensuring the fused images are optimized specifically for improving detection accuracy.
- We introduce a Gradient-Enhanced Residual Module (GERM) within the fusion network. This module effectively combines dense feature connections with gradient residual pathways to significantly strengthen edge, structural representation, and fine-grained texture details, providing high-quality, task-aware input for the detection stage.
- We design a robust detection network that integrates two key components to boost recognition: a Second-Order Channel Attention (SOCA) module that models higher-order feature dependencies to highlight semantically important regions and a Multi-scale Contextual Feature Encoding (MCFE) module that captures features across different scales, significantly enhancing the network's robustness for detecting the small and variably scaled targets common in remote sensing.
- Experimental results on the VEDAI and FLIR-ADAS dataset demonstrate that our method achieves state-of-the-art performance, confirming the effectiveness of our joint optimization strategy, particularly in enhancing detection under challenging conditions.

The structure of the rest of this paper is arranged as follows: Section 2 will review the related research work, including the attention mechanisms in multimodal image fusion,

as well as image fusion methods in both the spatial domain and the transform domain. Section 3 will elaborate on the proposed JFDet in detail, covering its overall network architecture, the Phase-Aware Gradient-Enhanced Residual Network for Image Fusion, the Visual Perception-Oriented Network for Object Detection, along with the loss functions and adaptive training strategies in the detection-driven image fusion framework. Section 4 will present and discuss the experimental results on the VEDAI and FLIR-ADAS datasets. Finally, Section 5 will summarize the research work of this paper and briefly prospect the potential future research directions.

2. Related Work

2.1. The Attention Mechanism in Multimodal Image Fusion

The attention mechanism, inspired by the human visual system's ability to focus adaptively on salient regions, has been widely applied in multimodal image fusion to enhance the quality of fused images by emphasizing key information. In multimodal fusion, different image modalities often contain complementary features, and effective fusion requires both preserving modality-specific features and reducing redundancy. Attention mechanisms improve feature representation in spatial and channel dimensions by adaptively weighting important regions.

Wang et al. [20] proposed non-local neural networks to capture long-range spatial dependencies through non-local operations. By computing pixel-wise similarity, the network models long-range contextual relationships, which is particularly beneficial when the targets and backgrounds in multimodal images are distant, enhancing correspondence between modalities and improving fusion quality. Channel attention is exemplified in Hu et al.'s [21] SENet (Squeeze-and-Excitation Networks), which dynamically reweights channels via a squeeze-and-excitation operation using global average pooling. Such reinforcement of channel modeling strengthens the network's capacity to discriminate features in multimodal fusion. However, relying solely on first-order statistics limits the ability to capture complex inter-modal relationships, leaving room for higher-order attention mechanisms. Recently, Transformer-based architectures have been integrated into multimodal fusion. Ma et al. [22] proposed SwinFusion, incorporating the Swin Transformer with shifted-window self-attention and cross-domain fusion modules. This design captures long-range intra-modal dependencies and establishes cross-modal associations, efficiently integrating global information and enhancing complementary feature fusion between modalities such as visible and infrared images. Transformers excel at modeling global dependencies, making them highly effective for optimizing multimodal fusion performance.

2.2. Image Fusion in Spatial and Transform Domains

Image fusion methods aim to combine complementary information from multiple source images to produce a single image with enhanced interpretability and feature representation. Traditionally, these methods can be broadly classified into spatial domain and transform domain approaches. Spatial domain methods operate directly on the pixel values of the source images, using techniques such as weighting, gradient, or convolution operations [23]. They are structurally simple, computationally efficient, and can preserve overall image structures. Common examples include multi-scale decomposition algorithms like Laplacian, gradient, and morphological pyramids, as well as discrete wavelet transform (DWT) [24] and shift-invariant DWT (SIDWT) [25]. However, spatial domain methods have several limitations: they can be sensitive to lighting variations, noise, and complex imaging conditions and may struggle to model long-range dependencies, leading to potential information loss or artifacts [26].

To overcome the limitations of spatial domain methods in modeling long-range dependencies and capturing global structures—particularly in remote sensing images, where local receptive fields may miss continuous surface features [27]—transform domain methods have become increasingly popular. These methods convert images from the spatial domain to the frequency domain via mathematical transforms, selectively process frequency components, and reconstruct images through inverse transforms, enabling more effective separation and utilization of frequency-specific features [28,29]. Fourier Transform (FFT) decomposes images into amplitude (energy distribution) and phase (structural geometry), facilitating the decoupling of degraded information [27]. Discrete Cosine Transform (DCT) is widely used in image compression, where it directly processes encoded image blocks, reduces computational cost, and facilitates resource-constrained applications [28].

Recent deep learning approaches have combined transform domain features with convolutional networks to enhance fusion and image restoration. For example, in low-light remote sensing image enhancement, dual-domain feature fusion networks (DFFNs) use FFT to separate amplitude and phase: amplitude guides illumination restoration, while phase refines structural details, and both are fused with spatial convolutional features to capture local textures and global structures [27]. In multi-focus image fusion, DCT-based strategies select high-information blocks using spatial frequency metrics, avoiding block artifacts while leveraging convolutional feature extraction [28]. In cross-modal fusion tasks, such as infrared-visible image fusion, FFT's magnitude–phase separation enhances frequency representation: phase preserves structural information while magnitude reflects energy. When combined with convolutional spatial features, this improves the complementarity of different modalities, yielding fused images with clearer details and stronger discriminative power [29].

3. Methodology

This section introduces the overall network architecture of the detection-driven multimodal remote sensing image fusion method, the composition structure and working principle of the phase-aware gradient-enhanced residual image fusion network, and the visual perception-based object detection network, as well as the loss function and training strategy applied in the detection-driven image fusion framework.

3.1. The Overall Architecture of JFDet

We propose a dual-driven multimodal fusion network for object detection. An overview of the network architecture is shown in Figure 1. Our network comprises two main components: an image fusion network and an object detection network. In the image fusion network, we introduce a GERM. By fusing a dense connection structure with high-frequency gradient information, the GERM effectively enhances the representation of fine details and improves fusion efficiency. This design strikes a balance between fusion accuracy and computational cost.

For the object detection network, we employ a second-order channel attention (SOCA) mechanism to improve the modeling of key semantic regions. This module is combined with a Multi-Scale Contextual Feature Encoding (MCFE) module to expand the receptive field, thereby enhancing detection robustness for objects of varying scales.

The entire network is trained using a joint optimization strategy that incorporates both fusion and detection losses. The fusion loss is composed of three components—intensity, texture, and gradient—which guide the fused image to preserve both structural information and detailed features. The detection loss, conversely, improves the accuracy of object localization and recognition through constraints from bounding box regression loss, object confidence loss, and category classification loss. This collaborative feedback from the dual

losses allows the fusion network to adaptively generate fused images that are more suitable for subsequent detection tasks, ultimately achieving a simultaneous improvement in both fusion quality and detection performance.

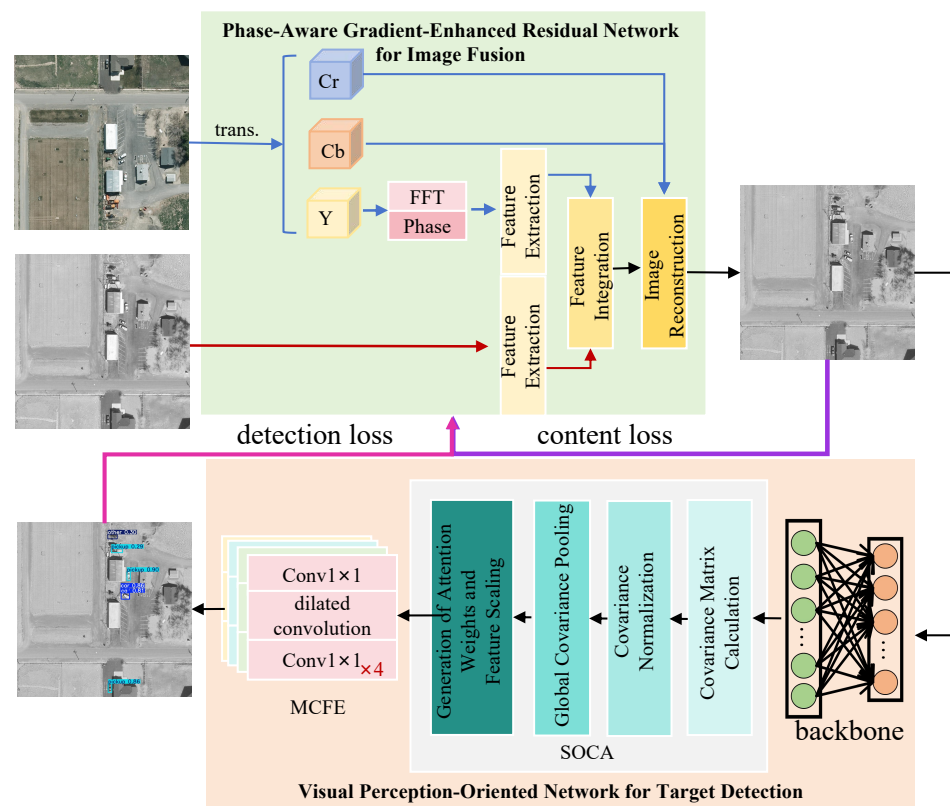


Figure 1. The overall architecture of JFDet.

3.2. Phase-Aware Gradient-Enhanced Residual Network for Image Fusion

The proposed joint adaptive optimization training network for image fusion and object detection employs a pixel alignment-based fusion network as its first stage. The overall architecture of the Phase-Aware Gradient-Enhanced Residual Image Fusion Network is shown in Figure 2.

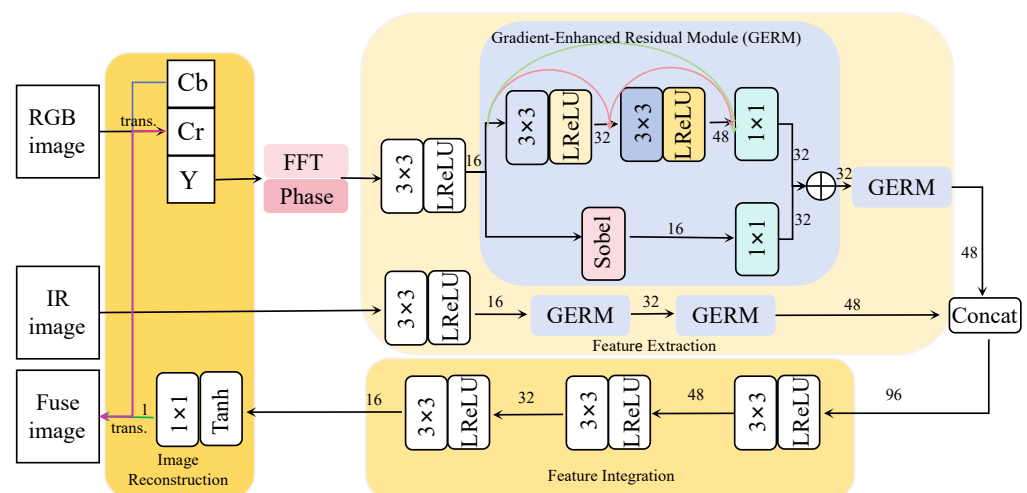


Figure 2. Phase-aware gradient-enhanced residual network for image fusion.

3.2.1. Phase-Aware Strategy Based on Fourier Transform

This network takes registered visible-light I_{vis} and infrared I_{inf} images as input. First, we convert the visible-light image to the YCbCr color space to decouple the brightness (Y-channel) from the color components (Cb and Cr channels). Traditional methods directly fuse the Y-channel, but they often fail to capture the high-frequency structural details and spatial features essential for contour recognition in object detection.

To address this issue, we propose a Y-channel optimization strategy based on Fourier phase enhancement. Instead of directly fusing the original Y-channel with the infrared image, our method first applies a Fourier transform to the Y-channel. This process decomposes the image into a amplitude and spectra. The high-frequency structural features, such as the spatial distribution of edge contours and texture gradients, are extracted from the phase spectrum. These features are then multiplied element-wise with the original Y-channel to create a phase-enhanced Y-channel (Y_{phase}). The multiplication operation leverages the brightness of the Y-channel as a prior weight, ensuring that structural information is effectively injected into the brightness representation. It enhances structural features in high-brightness regions and rapidly achieves precise alignment between structure and brightness.

This method offers two main advantages: First, it enhances spatial structure fidelity by fusing phase information with the original Y-channel. This preserves brightness levels while retaining edge contours and texture details, avoiding the blurring of object contours observed with traditional Y-channel fusion methods. Second, it improves the robustness of feature representation. Unlike the original Y-channel, which is sensitive to illumination changes and noise, the Fourier phase spectrum is more robust to variations in illumination, reflecting relative positional relationships rather than absolute brightness. As a result, the Y_{phase} channel stably preserves object structures in complex lighting conditions. When fused with infrared images, this enhances the resistance of the fused features to illumination disturbances, providing more stable input for subsequent object detection models.

After constructing the Y_{phase} channel, it is fused with the infrared image to create a single-channel image that combines intensity and structural features. The Cb and Cr channels preserve the chrominance information from the visible light image, maintaining color consistency. Finally, the fused single-channel image is combined with the original Cb and Cr channels and converted back to the RGB color space, ensuring a balanced fusion of structural, thermal, and color features.

3.2.2. Phase-Aware Gradient-Enhanced Residual Network for Image Fusion

To effectively fuse the complementary information from visible-light and infrared modalities, we propose a three-stage fusion network: feature extraction, feature aggregation, and image reconstruction. This modular architecture is designed to address three key challenges: extracting sufficient fine-grained information, achieving effective cross-modal integration, and ensuring perceptual consistency in the reconstructed image.

Feature Extraction: Feature extraction is a core component designed to capture deep, fine-grained features from both visible-light and infrared inputs. To achieve this, it first uses a 3×3 convolutional layer to extract low-level features, followed by two serially connected GERM. These GERM progressively improve the network's ability to express fine-grained textures and high-level semantic features.

Assuming the input feature to a GERM is F^i , its output F^{i+1} is defined as follows:

$$F^{i+1} = \text{GERM}(F^i) = \text{Conv}^n(F^i) \oplus \text{Conv}(\nabla F^i) \quad (1)$$

where $\text{Conv}^n(\cdot)$ corresponds to a 3-layer cascaded convolution structure ($n = 3$), specifically: the 1st layer is a 3×3 convolution (16 input channels, 32 output channels); the 2nd layer is a 3×3 convolution (32 input channels, 48 output channels); the 3rd layer is a 1×1 convolution (48 input channels, 48 output channels). $\text{Conv}(\cdot)$ refers to a 1×1 convolution (adjusting the 16-channel features output by the gradient operator to 48 channels), and ∇ is a gradient operator that extracts high-frequency details. The $\oplus$ symbol signifies an element-wise summation, which fuses the convolutional features and the gradient features. This specific structure is illustrated in Figure 2.

The GERM's design offers significant advantages. This module deeply fuses Sobel gradients with convolutional features. Leveraging the core characteristic of 3×3 neighborhood weighted summation, the Sobel operator can, on one hand, accurately capture high-frequency structural features in the visible-light Y_{phase} channel, and on the other hand, the gradients extracted by the Sobel operator can focus on large-scale continuous thermal edges formed by thermal radiation differences in infrared images. Meanwhile, its lightweight computation advantage does not introduce additional computational burden. Combined with the design of residual connections and dense connections, the module achieves sufficient transmission and deep fusion of cross-layer features. It not only avoids redundancy and computational loss caused by repeated feature calculations but also significantly enriches the dimension and effectiveness of feature representation through the key high-frequency details injected by Sobel gradients. This allows the feature extraction module to focus on two types of essential information for subsequent object detection: the texture details from the visible-light channel Y_{phase} and the thermal features from the infrared channel. This dual-feature focus helps the detection model quickly acquire key contour and thermal information, thereby reducing missed detections and improving overall performance.

Feature Aggregation: After being extracted, the visible-light and infrared features are fed into the feature aggregation. Here, the two feature streams are concatenated along the channel dimension before being processed by a cascade of three 3×3 convolutions, each followed by a Leaky ReLU (LReLU) activation function. This process gradually compresses the channels while enhancing spatial details.

The aggregation process can be expressed as follows:

$$\begin{aligned} F_0 &= [F_{vis}; F_{ir}], \\ F_1 &= \phi_1(\text{Conv}_{3 \times 3}(F_0)), \\ F_2 &= \phi_2(\text{Conv}_{3 \times 3}(F_1)), \\ F_{agg} &= \phi_3(\text{Conv}_{3 \times 3}(F_2)) \end{aligned} \quad (2)$$

where F_{vis} represents visible light features, F_{ir} represents infrared features, and $[\cdot; \cdot]$ represents channel concatenation. ϕ_1 , ϕ_2 and ϕ_3 represent the LReLU activation function. F_1 and F_2 represent intermediate features obtained through convolution and LReLU activation functions. $F_{agg} \in R^{C \times H \times W}$ is the fused multi-channel feature. The cascading of these three convolutions serves two key purposes: channel compression and detail enhancement. It reduces computational burden and improves inference speed by compressing redundant channels while simultaneously improving the alignment of infrared and visible-light features, which in turn reduces detection errors from feature misalignment and enhances recognition accuracy for small and blurry targets.

Image Reconstruction: The image reconstruction maps the aggregated feature F_{agg} to a single-channel luminance map $\hat{Y}$. This is achieved using a 1×1 convolution followed by an activation function that normalizes the output to the range $[0, 1]$.

$$\hat{Y} = \{\sigma(\text{Conv}_{1 \times 1}(F_{\text{agg}}))\} \quad (3)$$

where $\sigma(x) = \frac{1}{2}(\tanh(x) + 1)$.

Finally, the chrominance channels (Cb and Cr) from the original visible-light image are concatenated with the reconstructed luminance map to form a fused YCbCr image.

$$\hat{I}^{YCbCr} = [\hat{Y}, Cb, Cr] \quad (4)$$

This reconstructed YCbCr image $\hat{I}^{YCbCr}$ is then converted back to the RGB color space to produce the final fused image $\hat{I}_{rgb}$.

3.3. Visual Perception-Oriented Network for Object Detection

The second stage of our joint adaptive optimization training network is a visual perception-based detection network. This network takes the fused images from the first stage as input, effectively compensating for the limitations of infrared images in texture detail and visible-light images in low-illumination environments. The overall workflow is illustrated in Figure 3.

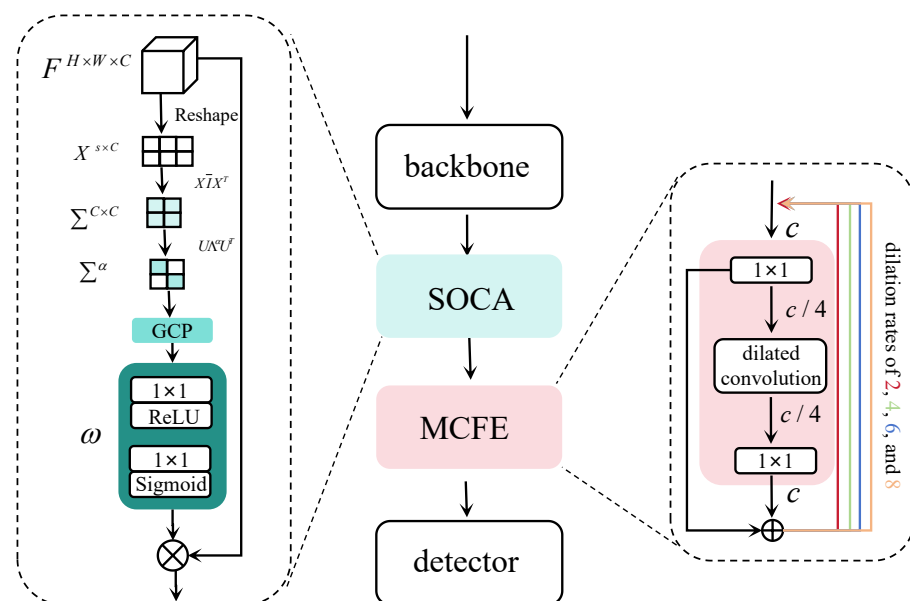


Figure 3. Visual perception-oriented network for object detection.

3.3.1. Second-Order Channel Attention (SOCA)

To address the unique challenges of remote sensing images—namely, small target sizes and complex morphologies—we introduce the Second-Order Channel Attention (SOCA) module proposed by Dai et al. [30]. Unlike most existing methods that construct channel attention based solely on first-order statistics (e.g., global average pooling), the SOCA module incorporates second-order feature statistics, enabling it to more comprehensively capture complex feature correlations. This core design was originally developed to enhance the feature discriminative ability in single image super-resolution (SISR) tasks; in this study, it is adapted to remote sensing scenarios. By adaptively highlighting key semantic regions, this module strengthens the model's ability to perceive and identify small-scale targets. The implementation of SOCA requires four steps, as described below.

(1) *Covariance Matrix Calculation*: The input feature map $F^{H \times W \times C}$ is first reshaped into a feature matrix X of the size $H \times W \times C$. The sample covariance matrix Σ is then calculated to describe the correlation between channels:

$$\Sigma = X\bar{I}X^T, \quad (5)$$

where $\bar{I} = \frac{1}{s}(I - \frac{1}{s}\mathbf{1})$, I is an $s \times s$ identity matrix, $\mathbf{1}$ is an $s \times s$ all-ones matrix, and $s = H \times W$.

(2) *Covariance Normalization*: We perform eigenvalue decomposition on the covariance matrix Σ ,

$$\Sigma = U\Lambda U^T, \quad (6)$$

where U is an orthogonal matrix, and Λ is an eigenvalue diagonal matrix. Normalization is then achieved by adjusting the power of the eigenvalues:

$$\hat{Y} = \sum^{\alpha} = U\Lambda^{\alpha}U^T, \quad (7)$$

In this study, we set $\alpha = 1/2$ to enhance feature discriminability, reduce redundant information, and extract useful information.

(3) *Global Covariance Pooling*: The normalized matrix $\hat{Y}$ is converted into channel-wise statistics, where the statistic for the c -th channel, z_c , is calculated as follows:

$$z_c = \frac{1}{C} \sum_{i=1}^C y_c(i), \quad (8)$$

Compared to first-order pooling, global covariance pooling effectively captures higher-order feature distributions, significantly improving the network's feature representation capability.

(4) *Generation of Attention Weights and Feature Scaling*: The channel-wise statistics z are transformed into attention weights via a gating mechanism:

$$\omega = f(W_U \delta(W_D z)), \quad (9)$$

where W_U and W_D are 1×1 convolution weights, δ is the *ReLU* activation function, and f is the *Sigmoid* function. These weights are then used to scale the input features:

$$\hat{f}_c = \omega_c \cdot f_c. \quad (10)$$

This process effectively enhances key channels while suppressing redundant ones, thereby improving feature discriminability.

3.3.2. Multi-Scale Contextual Feature Encoding (MCFE)

Although the SOCA module significantly enhances the network's ability to perceive small-scale objects by focusing on channel-wise feature selection, it has limitations in integrating contextual information across different spatial scales. Single-scale feature representation struggles to capture both the fine-grained local textures of small objects and the global structural information of larger objects. This limitation can negatively impact the feature matching accuracy of subsequent detection heads for multi-scale objects.

To overcome this, we designed MCFE to address the limited receptive field of a single convolutional layer, which makes it difficult to capture both local details and global structure. The MCFE module achieves multi-scale feature encoding by cascading multiple bottleneck units. By incorporating grouped convolutions with different dilation rates, it expands the receptive field while keeping the parameter scale controllable. This allows it to model multi-scale contextual information hierarchically, from local textures to global

contours. Furthermore, the residual connection mechanism within the module effectively alleviates information attenuation in deep features, preserving fine-grained original structures.

In terms of overall design, the MCFE is configured with a sequence of units with increasing dilation rates (e.g., [2, 4, 6, 8]). This systematic approach allows the network to progressively enhance and integrate detailed features at different scales, striking a balance between computational cost and receptive field range. By capturing local to global multi-scale spatial contextual information, the MCFE module significantly enhances the network's ability to represent multi-scale semantic information and is a vital spatial dimension supplement to the SOCA module.

3.4. Loss Function and Training Mechanism

Most existing task-driven image fusion methods use one of two approaches: either they pre-train a separate object detection network to guide the fusion process or they use end-to-end joint training. However, these methods have limitations. The pre-training approach often lacks high-quality fused images, making it difficult to provide stable supervision signals for the detection network. Meanwhile, the end-to-end approach struggles to balance fusion quality with detection performance in a single training phase. To overcome these challenges, we propose a joint optimization strategy for image fusion and object detection to achieve effective synergy. This strategy is defined by its novel loss function and adaptive training mechanism.

3.4.1. Loss Function

In our approach, we use a loss function that balances the integration of complementary information from source images with the final performance of the detection task. The total loss, L_{total} , is a weighted sum of the content Loss $L_{content}$ and detection Loss $L_{detection}$.

$$L_{total} = L_{content} + \gamma L_{detection}, \quad (11)$$

where γ is a dynamically changing parameter that reflects the detection loss, striking a balance between visual quality and task-driven performance.

(1) Fusion Loss: The $L_{content}$ term ensures that the fused image retains key information like prominent targets in infrared images and fine-grained details in visible-light images. The $L_{detection}$ term provides feedback from the object detection network, ensuring that the fused image is optimized for the detection task.

The content loss $L_{content}$ is a combination of three components: intensity loss (L_{int}), texture loss ($L_{texture}$), and gradient consistency loss (L_{grad}).

$$L_{content} = \alpha L_{int} + \beta L_{texture} + \lambda L_{grad}, \quad (12)$$

where the intensity loss (L_{int}) term preserves the most important pixel-level information. It works by forcing the fused image's pixel values to match the element-wise maximum of the source images. This method effectively retains the salient targets from the infrared image while preventing local darkness in the visible-light image. The texture loss $L_{texture}$ ensures that fine-grained details are preserved. We use a Sobel operator to extract gradients from both source images and the fused image. The loss then constrains the fused image's gradient to match the element-wise maximum of the source images' gradients, ensuring a richer and more realistic texture in the final result. The gradient consistency loss (L_{grad}) is designed to prevent structural distortion and improve edge integrity. While other loss components help, they can sometimes lead to broken or blurry edges. Our approach addresses this issue by directly constraining the gradient field distribution between the source and fused images, ensuring that structural information is accurately preserved.

Through experimental verification, the optimal experimental performance is achieved when α , β , and λ are set to 0.2, 1, and 1 respectively. Unlike methods that only match the strength of gradients, our approach compares the horizontal and vertical gradient distributions. This prevents false edges or structural breaks that can be caused by shifted gradient patterns. As a result, the fused image contains more complete structures and sharper, cleaner edges.

(2) Detection Loss: To ensure the fused image not only has visual quality advantages but also better serves downstream object detection tasks, we introduced detection loss for network optimization. This loss is a composite of three main components: bounding box regression loss L_{box} , object confidence loss L_{obj} , and category classification loss L_{cls} . The combined detection loss is defined as follows:

$$L_{detection} = \lambda_{box}L_{box} + \lambda_{obj}L_{obj} + \lambda_{cls}L_{cls}, \quad (13)$$

where λ_{box} , λ_{obj} , and λ_{cls} are weight parameters used to balance the contribution of each loss term. Through experimental verification, the optimal experimental performance is achieved when λ_{box} , λ_{obj} , and λ_{cls} are set to 0.05, 0.7 and 0.5 respectively.

The bounding box regression loss utilizes Generalized Intersection over Union (GIoU) to measure the overlap between the predicted and ground-truth bounding boxes. GIoU provides an effective gradient even when the boxes do not overlap, which helps mitigate the bounding box offset issue commonly found in the HBB dataset, thereby improving detection accuracy.

Object confidence loss uses the Binary Cross Entropy (BCE) to determine if a candidate box contains a target. Quantifying the difference between the predicted confidence and the ground-truth label guides the model to learn the distinction between targets and backgrounds in the fused image.

The category classification loss employs multi-class cross-entropy loss to constrain the prediction distribution of target categories. By measuring the difference between the predicted probability distribution and the ground-truth label, it helps the model capture unique features for different object categories in the fused image.

By incorporating this detection loss, our fusion network creates a closed-loop system for “visual optimization–detection adaptation.” This allows the network to not only maintain visual fidelity but also significantly improve the detection accuracy of targets, achieving dual optimization of perceptual quality and task performance.

3.4.2. Adaptive Training Strategy

We designed an adaptive training strategy that combines image fusion and object detection to optimize the network. The fusion and detection networks are trained iteratively for a total of M iterations.

First, the Adam optimizer is used to update all parameters in the fusion network, guided by a joint loss function. We dynamically adjust the coefficient of the detection loss in each iteration:

$$\gamma = \mu \times (m - 1), \quad (14)$$

where m represents the current iteration, and the constant coefficient μ is used to balance the detection loss and content loss. Experimental verification shows that when $\mu = 1$, the model can achieve the optimal balance between fusion quality and detection performance. In the first iteration ($m = 1$), the weight of the detection loss is set to 0, since the detection network has not yet been trained. For subsequent iterations ($m > 1$), we use the detection network from the previous iteration to calculate the detection loss, which is then incorporated into

the fusion network's training. Adopting the linearly increasing weight strategy can avoid interference from unreliable detection loss on the fusion network in the early training stage. As the detection network gradually converges, it enables the fusion network to adapt to the requirements of detection tasks incrementally, and ensures the stability of joint optimization for the two tasks without additional computational complexity.

As the number of iterations increases, the detection network fits the fusion network more closely. This allows the detection loss to provide more accurate guidance for training the fusion network. Therefore, the weight of the detection loss increases with each iteration. Here, μ is a constant used to balance the detection and content losses. Finally, we use the current fusion network to generate fused images and then update the parameters of the detection model by optimizing its detection loss. In each iteration, the number of training epochs is p for the fusion model, q for the detection model and b denotes the training batch size, respectively. The detailed process of the joint adaptive training strategy for the image fusion and object detection tasks is shown in the Algorithm 1.

Algorithm 1: Adaptive training strategy for joint image fusion and object detection tasks.

Input: Infrared Image I_{ir} and Visible Light Image I_{vi} ;
Output: Fused Image I_f

```

1 for  $m$  epochs do do
2   for  $p$  epochs do do
3     Select  $b$  infrared images  $\{I_{ir}^1, I_{ir}^2, \dots, I_{ir}^b\}$ ;
4     Select  $b$  visible light images  $\{I_{vi}^1, I_{vi}^2, \dots, I_{vi}^b\}$ ;
5     Update the detection loss weight  $\gamma$  according to Equation (14);
6     Update the fusion network parameters using the Adam optimizer:
        $\nabla_{NF}(L(N_F))$ ;
7   Fuse the images in the training dataset to generate fused images;
8   for  $q$  epochs do do
9     Select  $b$  fused images  $\{I_f^1, I_f^2, \dots, I_f^b\}$ ;
10    Update the detection network parameters using the SGD optimizer:
        $\nabla_{ND}(L(N_D))$ ;

```

4. Experiments and Discussion

In this section, we present quantitative and qualitative comparisons conducted using two public datasets: the Vehicle Detection in Aerial Imagery (VEDAI) dataset, which consists of optical remote sensing images, and the Teledyne FLIR ADAS dataset, which contains natural scene images.

4.1. Experimental Dataset

In this section, we describe the use of the public optical remote Vehicle Detection in Aerial Imagery (VEDAI) dataset (<https://downloads.greyc.fr/vedai/>, accessed on 29 December 2025) and natural-scene Teledyne FLIR ADAS dataset (<https://oem.flir.com/solutions/automotive/adas-dataset-form/>, accessed on 29 December 2025) to quantitatively evaluate the performance of the proposed method.

(1) VEDAI dataset: This dataset provides a benchmark for small object detection in aerial imagery. It originates from orthophoto aerial images (HRO 2012) supplied by Utah AGRC (Utah Geospatial Resource Center) in Salt Lake City, Utah, USA, with a spatial resolution of 12.5 cm/pixel and four channels (RGB and near-infrared). To support detection research, the large-format images are cropped into sub-images, among which

the VEDAI1024 subset contains 1210 images of the size 1024×1024 , which are available in both RGB and NIR modalities. Downsampled versions (512×512) are also provided, but this work used the full 1024×1024 resolution.

The dataset annotates nine vehicle categories, ranging from small cars to large trucks and special vehicles, with an average of 5.5 targets per image occupying only 0.7% of the pixels, highlighting the challenge in detecting small objects. Following standard practice, categories with fewer than 50 instances (e.g., bus, motorcycle, plane) are excluded, leaving eight categories (car, pick-up, truck, tractor, camping car, van, boat, and other) for experiments. Examples of images in the VEDAI dataset are shown in Figure 4.

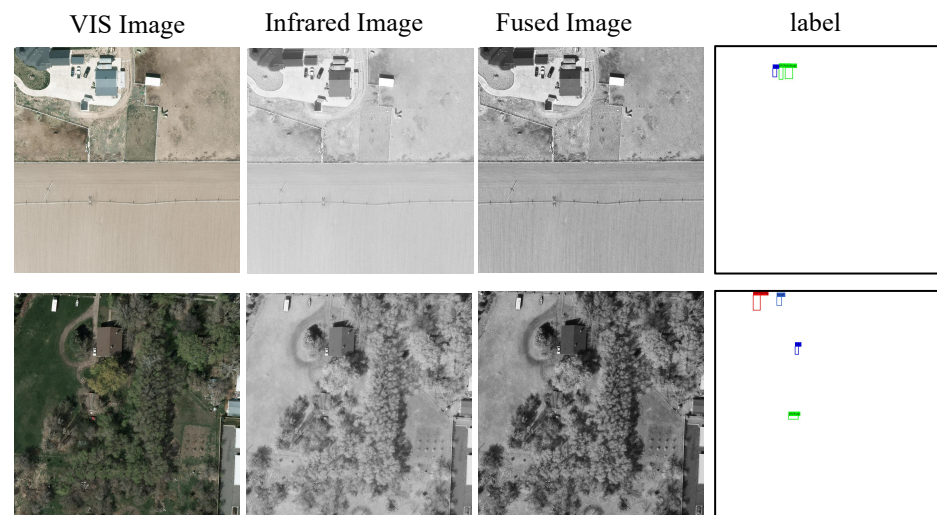


Figure 4. Example images from the VEDAI dataset.

(2) FLIR ADAS dataset: This dataset is a thermal imaging benchmark for Advanced Driver Assistance Systems (ADASs), which focuses on object detection in challenging environments such as low-exposure and low-light conditions. It consists of 10,228 images captured with FLIR Tau2 cameras at a uniform resolution of 640×512 , ensuring data consistency for algorithm evaluation.

The dataset covers both 6136 daytime images and 4092 nighttime images, enabling the study of perception challenges under varying levels of illumination. It provides both RGB and thermal modality images, although only the thermal images are annotated. The dataset was originally annotated with four categories (car, person, bicycle, and dog). However, due to the limited number of ‘dog’ samples, most studies, including this work, adopt only three categories: car, person, and bicycle. The dataset is partitioned into training and validation sets following the official protocol. The training set comprises 8862 images used for model training, whereas the validation set consists of 1366 images employed to evaluate model performance. Examples of the images in the FLIR ADAS dataset are shown in Figure 5.

4.2. Experimental Setup

In this experiment, the total number of iterations M is set to 8. In each iteration, the image fusion model is trained for 20 epochs, and the object detection model is trained for 50 epochs, with CSPDarknet53 adopted as the backbone. All experiments are implemented in PyTorch and conducted on a server with two NVIDIA A100-PCIE 40GB GPUs under Ubuntu 18.04, CUDA 12.4.



Figure 5. Example images from the FLIR ADAS dataset.

4.3. Evaluation Criteria

To quantitatively assess the performance of our object detection framework, we employ standard metrics widely used in the field: the Precision–Recall Curve (PRC) and the mean Average Precision at an IoU threshold of 0.5 (mAP50). The mAP 50 is the primary indicator of overall detection accuracy and is calculated as the average of the Average Precision (AP) across all k object categories:

$$\text{mAP} = \frac{\sum_{i=1}^k \text{AP}_i}{k}, \quad (15)$$

where AP_i is the area under the Precision–Recall curve (PRC) for category i , defined as: $\text{AP} = \int \text{P(R)} d\text{R}$. A higher mAP value (approaching 1) signifies better detection performance, indicating both high precision and recall across all target classes.

4.4. Ablation Experiments

To verify the effectiveness of the proposed image fusion network, object detection network, and the adaptive training strategy for the joint image fusion and object detection task, a series of ablation experiments were conducted on the VEDAI dataset and are described in this subsection.

Ablation of Fusion and Detection Methods: Table 1 presents the quantitative results of different combinations of fusion and detection methods on the VEDAI dataset. The quantitative results show that when the traditional fusion method GTF is combined with Yolov5s and the self-developed detection network (Ours), the mAP values are 73.7% and 74.2% respectively. Among deep learning-based fusion methods, DenseFuse and IFCNN achieve mAP values of 75.9% and 74.8% when combined with Yolov5s; when paired with the self-developed detection network, their mAP values increase to 76.7% and 75.0% respectively. In contrast, the fusion method proposed in this paper exhibits stronger framework adaptability. When combined with mainstream detectors such as Yolov5s, FCOS, Faster RCNN, and Mask RCNN, the mAP values reach 76.3%, 63.0%, 76.0%, and 73.8% in sequence. When this fusion method is deeply integrated with the self-developed detection network (i.e., JFDet), the mAP further breaks through to 79.6%, significantly outperforming all comparative combinations. This result fully proves that the proposed fusion method can effectively tap into the complementary value of multimodal information, and the collaborative design with the self-developed detection network can maximize the advantages of joint optimization, verifying the superiority and robustness of JFDet in complex scenarios of remote sensing image object detection.

Table 1. Ablation Analysis of Fusion and Detection Methods on the VEDAI Dataset.

Image Fusion Methods	Object Detection Methods	mAP@0.5
GTF	Yolov5s	73.7
DenseFuse	Yolov5s	75.9
IFCNN	Yolov5s	74.8
Ours	Yolov5s	76.3
GTF	Ours	74.2
DenseFuse	Ours	76.7
IFCNN	Ours	75.0
Ours	FCOS	63.0
Ours	Faster RCNN	76.0
Ours	Mask RCNN	73.8
Ours	Ours	79.6

Module Ablation: Table 2 presents the results of module ablation experiments on the VEDAI dataset. To ensure the fairness and comparability of the experiments, the infrared single-modality baseline (mAP = 75.4%) adopts the identical preprocessing and feature optimization strategies as the fusion network, except that multi-modal fusion and subsequent enhancement modules (GERM, SOCA, MCFE) are not incorporated. This value serves as the baseline for verifying the effectiveness of the experiments. It can be observed from the results that when the Gradient-Enhanced Residual Module (GERM, without Fourier transform) is used alone, the mAP reaches 75.8%. After introducing the Fourier transform (FFT) into the module, the accuracy increases to 76.3%. This indicates that the combination of Fourier-domain features and gradient information can better preserve image details.

Table 2. Ablation analysis of modules on the VEDAI dataset.

GERM Without FFT	GERM with FFT	SOCA	MCFE	mAP@0.5
				75.4
✓				75.8
✓		✓		76.3
✓			✓	75.9
✓		✓	✓	77.8
	✓			76.2
	✓	✓		77.5
	✓		✓	77.4
	✓	✓	✓	79.6

When the detection module and the GERM are jointly introduced, it can be observed that the performance gain of the combination of the detection module and the GERM with FFT is more significant. For example, the combination of the GERM (with FFT) and SOCA achieves an mAP of 77.8%, and the combination of the GERM (with FFT) and the feature encoding module reaches 77.4%. When all three modules are introduced simultaneously, the mAP reaches 79.6%, achieving optimal performance. These results indicate that each module can not only improve detection accuracy when used independently but also exhibit obvious complementarity when used in combination. Notably, compared with the infrared single-modality baseline (mAP = 75.4%), JFDet achieves a 4.2% mAP improvement to 79.6% when all modules are integrated. This gain precisely stems from the multimodal fusion's compensation for inherent single-modality defects—supplementing the lack of

texture details in infrared images and enhancing the anti-interference capability of visible light images—thus directly verifying the necessity of the proposed fusion strategy and the rationality and advancement of the JFDet structural design.

Table 3 presents the results of the loss function ablation study on the VEDAI dataset. It can be observed that when the three types of losses—intensity loss, texture loss, and gradient loss—are combined with the detection loss in pairwise combinations, the mean Average Precision (mAP) values are 78.5%, 77.6%, and 77.2% respectively, all demonstrating a certain degree of performance improvement. However, when all three types of losses are simultaneously combined with the detection loss, the mAP increases to 79.6%, reaching the optimal level. This indicates that multi-dimensional losses exhibit significant complementarity in both, ensuring the fidelity of fused images and enhancing downstream detection performance, thereby verifying the rationality and effectiveness of the joint loss design in JFDet.

Table 3. Ablation analysis of content loss function on the VEDAI dataset.

Intensity Loss	Texture Loss	Gradient Loss	Detection Loss	mAP@0.5
✓	✓		✓	78.5
✓		✓	✓	77.6
	✓	✓	✓	77.2
✓	✓	✓	✓	79.6

4.5. Quantitative Experiments

This section compares the proposed method with other state-of-the-art methods. To verify the robustness of the proposed method, experiments are conducted on two datasets, namely VEDAI and FLIR-ADAS.

Table 4 presents the comparison results on the VEDAI dataset. Compared with existing methods, JFDet achieves advantages in both overall accuracy and category performance: its mAP reaches 79.6%, which is 1.7% higher than that of MMFDet (77.9%) and 4.8% higher than that of FFCA-YOLO (74.8%). In key categories, the detection accuracy of Van and Truck reaches 90.7% and 86.6%, respectively, the highest among the compared methods, and the Camping category is also improved to 81.0%. These improvements are attributed to the improvements in the structure and optimization strategy of JFDet: the gradient residual dense module strengthens the fidelity of edges and details, the second-order channel attention and multi-scale feature encoding enhance the modeling ability for semantic regions and scale changes, and the introduction of the detection loss feedback mechanism ensures the adaptability of the fused image to the detection task. In contrast, existing methods still have certain limitations in the utilization of cross-modal features. Although CMAFF is competitive in terms of overall accuracy, its performance in complex categories such as “Other” is weak; SuperYOLO and FFCA-YOLO mainly rely on shallow feature fusion, and the mining of cross-modal complementary information is insufficient, resulting in limited detection accuracy in categories such as Boat and Other. ICAFusion improves the global feature interaction ability through the interactive attention mechanism, but its detection balance in key categories is still inferior to that of JFDet. MMFDet is outstanding in overall performance, but there is a deficiency in detection accuracy for categories such as Van. Overall, the improvements in the structural design and optimization strategy of JFDet effectively enhance the cross-modal feature modeling ability, achieving higher overall performance and better category balance, and demonstrating stronger robustness and generalization ability. Some detection results for the VEDAI dataset are shown in Figure 6.

Table 4. Quantitative performance comparisons on the VEDAI dataset. The optimal one is shown in bold.

Methods	Car	Pickup	Camping	Truck	Tractor	Boat	Van	Other	mAP@0.5
CMAFF [31]	91.7	85.9	78.9	78.1	71.9	71.7	75.2	54.7	76.01
SuperYOLO [32]	91.13	85.66	79.3	70.18	80.41	60.24	76.5	57.33	75.09
ICAFusion [33]	-	-	-	-	-	-	-	-	76.62
FFCAYOLO [34]	89.6	85.7	78.7	85.7	81.8	61.5	67.0	48.6	74.8
MMFDet [35]	88.3	78.5	81.6	59.8	86.2	76.0	88.3	63.5	77.9
JFDet	89.9	81.8	81.0	86.6	76.0	71.1	90.7	59.6	79.6

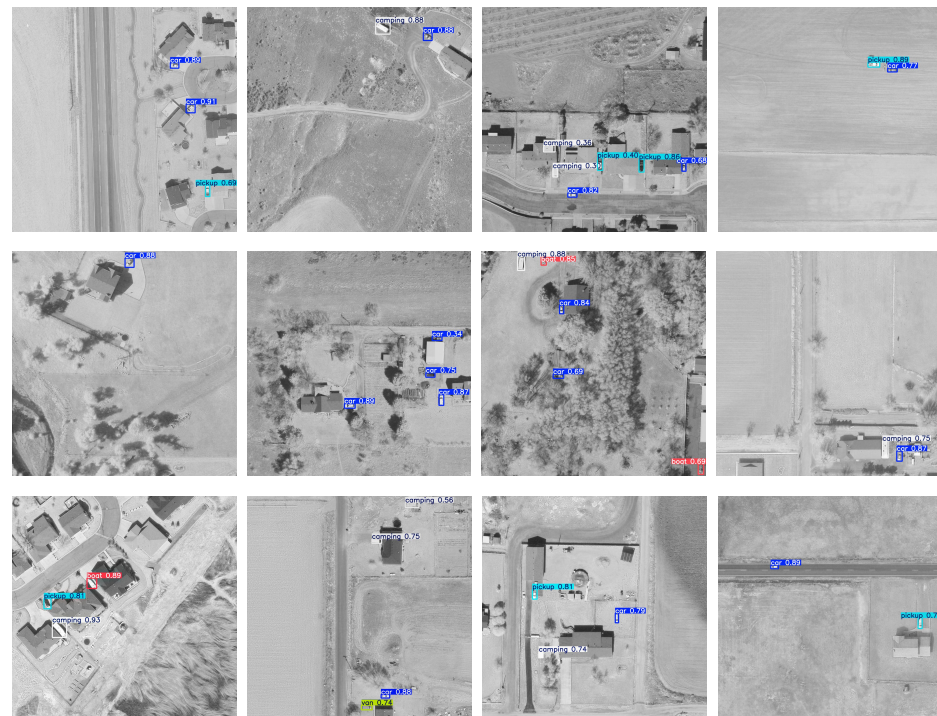
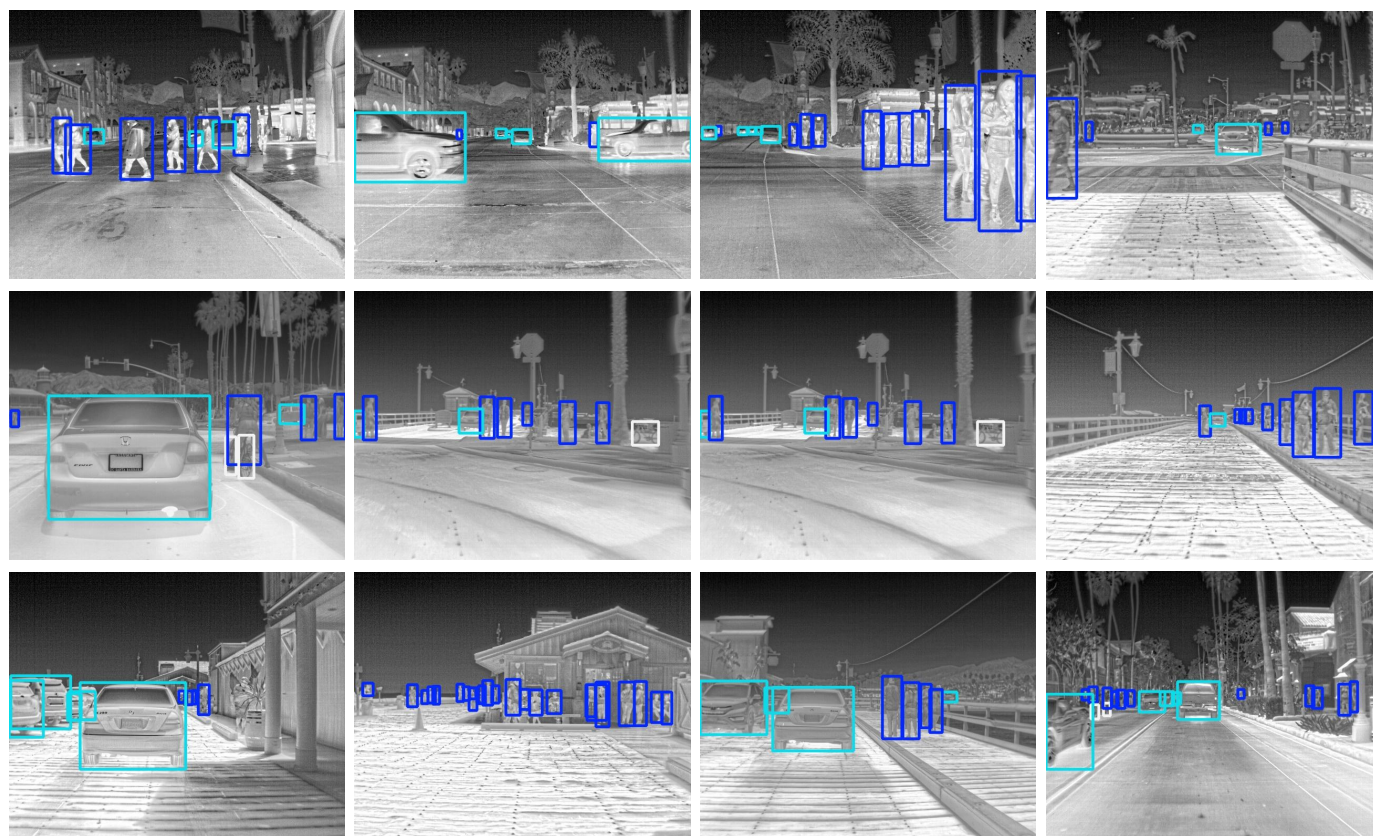
**Figure 6.** Visual detection results for the VEDAI dataset.

Table 5 presents the comparison results on the FLIR-ADAS dataset. Compared with existing methods, JFDet demonstrates advantages in both overall accuracy and category-specific performance: its mAP reaches 76.3%, which is 3.9% higher than that of CFR_3 (72.39%) and 3.2% higher than that of CAPTM_3 (73.15%), and it achieves performance that is on par with or slightly better than MSANet (76.2%) and CAMDet (T2V) (76.3%). In key categories, the detection accuracy for Person, Car, and Bicycle reaches 81.6%, 88.5%, and 58.8%, respectively, with all three categories achieving the best performance among the compared methods. By contrast, existing methods still have shortcomings in cross-modal information utilization and category balance: CFR_3 and CAPTM_3 show relatively low accuracy in the Bicycle category; GAFF and CAMDet (V2T) have limited overall performance; and although MSANet possesses strong feature modeling capabilities, its performance in the Car category is still lower than that of JFDet. Some detection results for the FLIR-ADAS dataset are shown in Figure 7.

Table 5. Quantitative performance comparisons on the FLIR-ADAS dataset. The optimal one is shown in bold.

Methods	Person	Car	Bicycle	mAP@0.5
CFR_3 [36]	74.49	84.91	57.77	72.39
GAFF [37]	-	-	-	72.9
MSANet [38]	-	-	-	76.2
CAPTM_3 [39]	77.04	84.61	57.79	73.15
CAMDet(V2T) [40]	-	-	-	75.4
CAMDet(T2V) [40]	-	-	-	76.3
JFDet	81.6	88.5	58.8	76.3

**Figure 7.** Visual detection results for the FLIR-ADAS dataset.

5. Conclusions

In this study, we address the critical task mismatch between low-level image fusion and high-level object detection in remote sensing imagery by proposing JFDet. Unlike conventional approaches that optimize fusion independently, JFDet achieves an effective detection-driven synergy by jointly optimizing both processes. Unlike existing methods in which fusion quality is often poorly aligned with detection performance, our framework jointly optimizes low-level image fusion and high-level object detection to create effective synergy. This network extracts high-frequency structural features from the phase spectrum of the visible light Y-channel and fuses them with the infrared channel, ensuring a complementary fusion of both structural and color features. For enhanced structural representation, we developed the Gradient-Enhanced Residual Module (GERM), which provides a lightweight solution to preserving crucial high-frequency details while balancing accuracy and computational cost. For robust object detection, particularly for small and multi-scale targets, we integrate a Second-Order Channel Attention (SOCA) mechanism to

capture higher-order semantic dependencies and a Multi-scale Contextual Feature Encoding (MCFE) module to enrich contextual information. Experimental results on the VEDAI and FLIR-ADAS datasets demonstrate that JFDet significantly improves both fusion quality and detection accuracy compared to state-of-the-art methods. Future work will explore the application of this framework to more diverse remote sensing tasks, such as semantic segmentation and change detection, as well as its performance with other multimodal sensor data.

Although the proposed JFDet has demonstrated excellent performance in infrared-visible multimodal remote sensing object detection tasks, it still has certain limitations: first, the scope of modal adaptation is limited, with validation only conducted on the infrared-visible dual-modality, and its generalization ability in other multimodal combinations (such as SAR-visible and hyperspectral-visible) needs further verification; second, the robustness in complex scenarios and for special targets is insufficient, as the experiments do not fully cover scenarios like strong interference and extreme weather, leaving room for improvement in the detection performance of ultra-small targets; third, the computational efficiency and practical deployment adaptability require optimization. The model has not been specially optimized for real-time requirements, and quantitative analysis of computational complexity and inference speed as well as in-depth lightweight improvements have not been carried out, which may restrict its practical application in time-sensitive scenarios. In the future, we will continue to explore the potential of extending this framework to more remote sensing tasks and multimodal data, and conduct targeted optimizations for the existing limitations to further enhance the model's generalization ability, scene adaptability, and practical deployment value.

Author Contributions: Conceptualization: W.X., Y.Y.; Methodology: W.X., Y.Y.; Code writing: W.X.; Experiment validation and data processing: W.X., Y.Y.; Result analysis and discussion: W.X., Y.Y.; Original draft preparation: W.X.; Review and editing: Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original data presented in the study are openly available in VEDAI Dataset Repository at <https://downloads.greyc.fr/vedai/>, accessed on 29 December 2025, and Teledyne FLIR ADAS Dataset Repository at <https://oem.flir.com/solutions/automotive/adas-dataset-form/>, accessed on 29 December 2025, which are publicly available open-source datasets used for model training, rather than datasets constructed by our research team.

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

References

1. Tang, D.; Cao, X.; Wu, X. Li, J.; Yao, J.; Bai, X.R.; Jiang, D.S.; Li, Y.; Meng, D.Y. AeroGen: Enhancing remote sensing object detection with diffusion-driven data generation. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR2025), Nashville, TN, USA, 11–15 June 2025; pp. 3614–3624.
2. Li, H.; Zhang, R.; Pan, Y.; Ren, J.C.; Shen, F. Lr-fpn: Enhancing remote sensing object detection with location refined feature pyramid network. In Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN2024), Piscataway, NJ, USA, 30 June–5 July 2024; pp. 1–8.
3. Wang, J.; Ma, L.; Zhao, B.; Gou, Z.; Yin, Y.; Sun, G. MRLF: Multi-Resolution Layered Fusion Network for Optical and SAR Images. *Remote Sens.* **2025**, *17*, 3740. [[CrossRef](#)]
4. Ding, X.; Fang, J.; Wang, Z. Liu, Q.; Yang, Y.; Shu, Z.Y. Unsupervised learning non-uniform face enhancement under physics-guided model of illumination decoupling. *Pattern Recognit.* **2025**, *162*, 111354. [[CrossRef](#)]
5. Khan, R.; Yang, Y.; Liu, Q.; Shen, J.L.; Li, B. Deep image enhancement for ill light imaging. *J. Opt. Soc. Am. A* **2021**, *38*, 827–839. [[CrossRef](#)] [[PubMed](#)]
6. Liu, Z.; Yang, Y.; Wu, K.; Liu, Q.; Xu, X.H.; Ma, X.X.; Tang, J. ASIFusion: An adaptive saliency injection-based infrared and visible image fusion network. *ACM Trans. Multim. Comput. Commun. Appl.* **2024**, *20*, 1–23. [[CrossRef](#)]

7. Khan, R.; Zhang, J.; Wang, Z.; Chen, J.; Wang, F.; Gao, L. An Automated Framework for Abnormal Target Segmentation in Levee Scenarios Using Fusion of UAV-Based Infrared and Visible Imagery. *Remote Sens.* **2025**, *17*, 3398.
8. Atrey, P.K.; Hossain, M.A.; El Saddik, A.; Kankanhalli, M.S. Multimodal fusion for multimedia analysis: A survey. *Multimed. Syst.* **2010**, *16*, 345–379. [\[CrossRef\]](#)
9. Li, H.; Wu, X.J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2018**, *28*, 2614–2623. [\[CrossRef\]](#)
10. Wang, Z.; Wang, J.; Wu, Y.; Xu, J.W.; Zhang, X.Q. UNFusion: A unified multi-scale densely connected network for infrared and visible image fusion. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 3360–3374. [\[CrossRef\]](#)
11. Long, Y.; Jia, H.; Zhong, Y.; Jiang, Y.D.; Jia, Y.M. RXDNFuse: A aggregated residual dense network for infrared and visible image fusion. *Inf. Fusion* **2021**, *69*, 128–141. [\[CrossRef\]](#)
12. Ma, J.; Yu, W.; Liang, P.; Li, C.; Jiang, J.J. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2019**, *48*, 11–26. [\[CrossRef\]](#)
13. Li, H.; Wu, X.J.; Kittler, J. RFN-Nest: An end-to-end residual fusion network for infrared and visible images. *Inf. Fusion* **2021**, *73*, 72–86. [\[CrossRef\]](#)
14. Xu, H.; Ma, J.; Jiang, J.; Guo, X.J.; Ling, H.B. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 502–518. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Zhang, H.; Xu, H.; Xiao, Y.; Guo, X.J.; Ma, J.Y. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI2020), New York, NY, USA, 7–12 February 2020; pp. 12797–12804.
16. Li, J.; Zhu, J.; Li, C.; Chen, X.; Yang, B. CGTF: Convolution-guided transformer for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–14. [\[CrossRef\]](#)
17. Zhou, H.; Hou, J.; Zhang, Y.; Ma, J.Y.; Ling, H.B. Unified gradient-and intensity-discriminator generative adversarial network for image fusion. *Inf. Fusion* **2022**, *88*, 184–201. [\[CrossRef\]](#)
18. Wang, J.; Xi, X.; Li, D.; Li, F. FusionGRAM: An infrared and visible image fusion framework based on gradient residual and attention mechanism. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–12. [\[CrossRef\]](#)
19. Xiang, X.; Zhou, G.; Niu, B.; Pan, Z.; Huang, L.; Li, W.; Wen, Z.; Qi, J.; Gao, W. Infrared-Visible Image Fusion Meets Object Detection: Towards Unified Optimization for Multimodal Perception. *Remote Sens.* **2025**, *17*, 3637. [\[CrossRef\]](#)
20. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
21. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
22. Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.G.; Ma, Y. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE-CAA J. Autom. Sin.* **2022**, *9*, 1200–1217. [\[CrossRef\]](#)
23. Li, S.; Kang, X.; Fang, L.; Hu, J.W.; Yin, H.T. Pixel-level image fusion: A survey of the state of the art. *Inf. Fusion* **2017**, *33*, 100–112. [\[CrossRef\]](#)
24. Li, H.; Manjunath, B.S.; Mitra, S.K. Multisensor image fusion using the wavelet transform. *CVGIP Graph. Model. Image Process.* **1995**, *57*, 235–245. [\[CrossRef\]](#)
25. Rockinger, O. Image sequence fusion using a shift-invariant wavelet transform. In Proceedings of International Conference on Image Processing (ICIP1997), Washington, DC, USA, 26–29 October 1997; pp. 288–291.
26. Haghighat, M.B.A.; Aghagolzadeh, A.; Seyedarabi, H. Multi-focus image fusion for visual sensor networks in DCT domain. *Comput. Electr. Eng.* **2011**, *37*, 789–797. [\[CrossRef\]](#)
27. Yao, Z.; Fan, G.; Fan, J.; Gan, M.; Chen, C. L. P. Spatial-frequency dual-domain feature fusion network for low-light remote sensing image enhancement. *IEEE Trans. Geosci. Remote Sens.* **2024**, *in press*. [\[CrossRef\]](#)
28. Cao, L.; Jin, L.; Tao, H.; Li, G.N.; Zhuang, Z.; Zhang, Y. F. Multi-focus image fusion based on spatial frequency in discrete cosine transform domain. *IEEE Signal Process. Lett.* **2014**, *22*, 220–224. [\[CrossRef\]](#)
29. Liu, Y.; Wang, L.; Cheng, J.; Li, C.; Chen, X. Multi-focus image fusion: A survey of the state of the art. *Inf. Fusion* **2020**, *64*, 71–91. [\[CrossRef\]](#)
30. Dai, T.; Cai, J.; Zhang, Y.; Xia, S.T.; Zhang, L. Second-order attention network for single image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2019), Long Beach, CA, USA, 16–20 June 2019; pp. 11065–11074.
31. Qingyun, F.; Zhaokui, W. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognit.* **2022**, *130*, 108786. [\[CrossRef\]](#)
32. Zhang, J.; Lei, J.; Xie, W.; Fang, Z.M.; Li, Y.S.; Du, Q. SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery. *TGRS IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15 [\[CrossRef\]](#)

33. Shen, J.; Chen, Y.; Liu, Y.; Zuo, X.; Fan, H.; Yang, W.K. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognit.* **2024**, *145*, 109913. [[CrossRef](#)]
34. Zhang, Y.; Ye, M.; Zhu, G.; Liu, Y.; Guo, P.Y.; Yan, J.H. FFCA-YOLO for Small Object Detection in Remote Sensing Images. *TGRS IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5611215. [[CrossRef](#)]
35. Zhao, W.; Zhao, Z.; Avignon, B. Differential multimodal fusion algorithm for remote sensing object detection through multi-branch feature extraction. *Expert Syst. Appl.* **2025**, *265*, 125826. [[CrossRef](#)]
36. Zhang, H.; Fromont, E.; Lefevre, S.; Wu, X.J.; Kittler, J. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP2020), Virtual, 25–28 October 2020; pp. 276–280.
37. Zhang, H.; Fromont, E.; Lefèvre, S.; Avignon, B. Guided attentive feature fusion for multispectral pedestrian detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV2021), Virtual, 5–9 January 2021; pp. 72–80.
38. You, S.; Xie, X.; Feng, Y.; Mei, C.J.; Ji, Y.M. Multi-scale aggregation transformers for multispectral object detection. *IEEE Signal Process. Lett.* **2023**, *30*, 1172–1176. [[CrossRef](#)]
39. Zhou, H.; Sun, M.; Ren, X.; Wang, X. Visible-thermal image object detection via the combination of illumination conditions and temperature information. *Remote Sens.* **2021**, *13*, 3656. [[CrossRef](#)]
40. Jang, J.; Lee, J.; Paik, J. CAMDet: Condition-adaptive multispectral object detection using a visible-thermal translation model. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2025), Hyderabad, India, 6–11 April 2025; pp. 1–5.

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.