remote sensing

MDPI

Article

A Method for Airborne Small-Target Detection with a Multimodal Fusion Framework Integrating Photometric Perception and Cross-Attention Mechanisms

Shufang Xu [1,2], Heng Li [3], Tianci Liu [1] and Hongmin Gao [1,*]

[1] College of Computer Science and Software Engineering, Hohai University, Nanjing 211100, China; xushufang@hhu.edu.cn (S.X.); 211607010051@hhu.edu.cn (T.L.)
[2] Key Laboratory of Optical Remote Sensing and Intelligent Information Processing, Xi'an 710119, China
[3] College of Information Science and Engineering, Hohai University, Changzhou 213200, China; 221307020008@hhu.edu.cn
* Correspondence: gaohongmin@hhu.edu.cn

Abstract: In recent years, the rapid advancement and pervasive deployment of unmanned aerial vehicle (UAV) technology have catalyzed transformative applications across the military, civilian, and scientific domains. While aerial imaging has emerged as a pivotal tool in modern remote sensing systems, persistent challenges remain in achieving robust small-target detection under complex all-weather conditions. This paper presents an innovative multimodal fusion framework incorporating photometric perception and cross-attention mechanisms to address the critical limitations of current single-modality detection systems, particularly their susceptibility to reduced accuracy and elevated false-negative rates in adverse environmental conditions. Our architecture introduces three novel components: (1) a bidirectional hierarchical feature extraction network that enables the synergistic processing of heterogeneous sensor data; (2) a cross-modality attention mechanism that dynamically establishes inter-modal feature correlations through learnable attention weights; (3) an adaptive photometric weighting fusion module that implements spectral characteristic-aware feature recalibration. The proposed system achieves multimodal complementarity through two-phase integration: first by establishing cross-modal feature correspondences through attention-guided feature alignment, then performing weighted fusion based on photometric reliability assessment. Comprehensive experiments demonstrate that our framework achieves an improvement of at least 3.6% in mAP compared to the other models on the challenging LLVIP dataset, and with particular improvements in detection reliability on the KAIST dataset. This research advances the state of the art in aerial target detection by providing a principled approach for multimodal sensor fusion,with significant implications for surveillance, disaster response, and precision agriculture applications.

Keywords: light sensing; cross-attention mechanism; multimodal; small-target detection; aerial photography with UAV

check for updates

Academic Editor: Shuying Li

Received: 12 February 2025
Revised: 13 March 2025
Accepted: 19 March 2025
Published: 21 March 2025

Citation: Xu, S.; Li, H.; Liu, T.; Gao, H. A Method for Airborne Small-Target Detection with a Multimodal Fusion Framework Integrating Photometric Perception and Cross-Attention Mechanisms. *Remote Sens.* **2025**, *17*, 1118. https://doi.org/10.3390/rs17071118

Copyright: © 2025 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) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, unmanned aerial vehicle (UAV) technology has rapidly evolved beyond military applications to civilian and research fields, finding widespread use in terrain mapping, logistics, precision agriculture, target tracking, and disaster relief. UAV imagery has contributed to datasets supporting research in hyperspectral anomaly detection [1], image detection [2], classification [3], remote sensing analysis [4], and visual data augmentation [5]. Despite its significance, small-target detection remains a major challenge.

Insufficient lighting reduces contrast and detail in visible-light images, hindering feature extraction. Moreover, small-target identification is influenced by lighting conditions, target size, and background interference, making traditional detection methods less effective.

Convolutional Neural Networks (CNNs) have achieved significant advancements in feature extraction [6], image classification [7], image denoising [8], object recognition [9], fault diagnosis [10], hyperspectral unmixing [11], super-resolution reconstruction [12], and change detection [13]. CNN-based target detection methods have garnered increasing attention, with the YOLO series being widely applied across various fields due to its superior performance [14]. The initial version of YOLO employed a grid-based approach with single-shot inference, enabling fast detection but exhibiting lower accuracy for small targets. YOLOv2 introduced anchor boxes and K-means clustering, enhancing small-target detection. YOLOv3 incorporated multi-scale detection and adopted Darknet-53 as its backbone to improve accuracy. YOLOv4 integrated advanced optimization strategies to achieve a better trade-off between speed and accuracy, while YOLOv5 further refined the model and introduced the Spatial Pyramid Pooling-Fast (SPPF) structure for broader applicability. In 2021, YOLOX adopted an anchor-free architecture to simplify the model, while YOLOR introduced a multi-task network for both detection and segmentation. In 2022, YOLOv6 integrated the Re-parameterizable Visual Geometry Group (RepVGG) structure, and YOLOv7 further enhanced performance through architectural optimization and dynamic label assignment.

Balancing detection accuracy and model size remains a significant challenge in UAV-based target detection. To address this issue, researchers have proposed various enhancements to YOLO models. For instance, Lou et al. [15] improved small-target detection by refining downsampling and feature fusion, while Guo et al. [16] introduced the Convolutional 3D Network (C3D) structure to better preserve spatial and temporal information. Wang et al. [17] developed the Spatial–Temporal Complementary Network (STC) and incorporated global attention (GAM) to mitigate feature loss. Additionally, Wang et al. [18] integrated the BiFormer attention mechanism and the Feed-Forward Neural Block (FFNB) module, expanding detection scales to reduce missed detections. However, the accuracy of detecting certain small objects, such as bicycles, still requires further improvement. Currently, most object recognition methods rely on single-modality visible-light images [19,20], which perform poorly under low-light or nighttime conditions. Infrared imagery, with its advantages in nighttime environments, serves as a valuable complement to visible-light images, improving detection accuracy and robustness. Compared to traditional pixel-level and decision-level fusion, feature-level fusion directly integrates image features, thereby avoiding registration errors while enhancing information utilization and recognition efficiency. As a result, feature-level fusion has emerged as a widely studied approach in multimodal image processing.

With the increasing role of deep learning in computer vision, image feature fusion has emerged as a key research focus. In 2018, Li et al. [21] proposed a feature fusion method that employs weighted averaging and L1-norm fusion to generate multi-level fused images. Hwang et al. [22] introduced the KAIST dataset and developed a multimodal fusion method—Aggregate Channel Features combined with Temporal Features and Temporal Histogram of Oriented Gradients (ACF+T+THOG). However, its recognition performance remained suboptimal. In 2016, Feichtenhofer et al. [23] demonstrated that multi-stage fusion in deep learning enhances feature extraction. That same year, Wagner [24] found that late-stage CNN fusion yielded superior performance, whereas Liu [25] identified mid-stage fusion as the optimal strategy. In 2018, Li et al. [26] introduced a Multi-Scale and Dual-Stream Faster R-CNN (MSDS-RCNN) with mid-layer fusion for pedestrian detection, significantly improving accuracy.

To address lighting variations in image fusion, researchers have developed adaptive fusion methods. In 2019, Guan et al. [27] designed a light-sensing network incorporating dedicated sub-networks for daytime and nighttime, thereby enhancing adaptability by dynamically predicting lighting conditions. Li et al. [28] leveraged light intensity as a fusion weight, training a network to estimate illumination levels and adjust detection results accordingly. Zhang et al. [29] introduced channel-level attention mechanisms for multimodal fusion, leading to improved detection accuracy. In 2020, Zhang et al. [30] proposed a spectral feature unification module, enhancing feature consistency and overall detection performance. Subsequently, in 2021, Zhang et al. [31] integrated intra- and inter-modal attention mechanisms to adaptively weight multispectral features, achieving higher accuracy while maintaining low computational cost. More recently, Xiongxin et al. [32] (2023) developed a YOLOv5-based UAV human detection framework incorporating visible-infrared fusion. In 2024, Sangin et al. [33] introduced the Infrared and Visible Image Saliency-Aware Network (INSANet), an attention-based fusion network designed to capture global spectral relationships. Meanwhile, Qian et al. [34] proposed a hallucination-based domain adaptation approach for thermal imaging pedestrian detection, effectively integrating virtual visible-light and thermal images to enhance accuracy.

In summary, multimodal fusion in target detection presents several challenges, particularly in effectively determining optimal fusion stages and mitigating information loss. While numerous methods integrate visible and infrared images, they often fail to account for feature variations across different fusion stages, resulting in suboptimal performance—especially for small targets with sparse features that are susceptible to loss. Additionally, many studies do not adequately consider lighting conditions and modality differences. Lighting significantly impacts the quality of aerial imagery, yet numerous algorithms treat varying illumination conditions uniformly, overlooking their influence on detection performance. Furthermore, differences in color, texture, and brightness between visible and infrared images can lead to information omission or distortion, thereby reducing detection accuracy and overall robustness. To address these challenges, this paper proposes an improved UAV-based aerial target detection framework leveraging multimodal feature fusion. Specifically, we develop a YOLOv5-based fusion approach that integrates light sensing and cross-attention mechanisms to mitigate the limitations of visible-light-only detection, such as low accuracy and high false negative rates. The proposed method employs a bidirectional feature extraction network incorporating a cross-attention mechanism to enhance modality-specific feature representation. Additionally, a light-sensing weight module is designed to adaptively fuse features based on illumination conditions. The fused feature representations are subsequently fed into the detection network, enabling high-efficiency, all-weather target detection. The main contributions of this article are summarized as follows:

- To address the impact of illumination variations on feature fusion, we introduce a dedicated neural network within the model to construct a light-sensing weight module. This module leverages illumination information to dynamically adjust fusion weights, significantly enhancing target detection accuracy under all-weather conditions.
- To mitigate modality differences, we design a cross-attention module that processes feature maps generated by the dual-stream network. This module enables the model to capture correlations between different modalities and fully exploit cross-modal interactions, thereby improving the accuracy of target detection and recognition.
- The proposed method is evaluated on publicly available datasets, including DroneVehicle, KAIST, and LLVIP, and compared against several state-of-the-art algorithms. The experimental results demonstrate that our approach exhibits superior generalization

ability, accurately detecting targets even under conditions of indistinct features or occlusion, and achieving notable improvements in pedestrian target recognition.

The rest of this article is organized as follows. Section 2 introduces the datasets used in the experiments and the performance evaluation metrics and provides a detailed description of the proposed method. Section 3 includes the experimental details and results. Section 4 discusses the experimental results. Section 5 gives conclusions and some possible future works.

2. Materials and Methods

In this study, the experiments are conducted using publicly available multimodal datasets. During the training and testing phases, we utilize aligned visible-light and infrared image datasets, including KAIST, LLVIP, and DroneVehicle. Specifically, the DroneVehicle dataset is employed for various ablation studies, while the LLVIP and KAIST datasets are used to evaluate the generalization capability of the proposed algorithm. To assess the overall performance of different methods, we adopt two primary evaluation metrics: the logarithmic average missed detection rate (MR^{-2}) and average precision (AP). These metrics provide a comprehensive measure of detection accuracy and robustness. A detailed description of the datasets and evaluation metrics is provided in the subsequent sections.

2.1. Definition of Small Targets

In different scenarios, the definition of small targets varies based on specific contexts. However, in academic research, the definition is generally categorized into two types. The first category is based on a relative scale, often involving the ratio of the target to the image. Chen et al. [35] proposed a definition method, when defining small targets, the relative area of the target instance (the ratio of the bounding box area to the image area) is typically set within the range of 0.08% to 0.58% for instances in the same category. The second category is based on an absolute scale, where small targets are defined by their absolute pixel size. A widely accepted standard in the field of object detection comes from the Microsoft Common Objects in Context (MS COCO) dataset, where small targets are defined as those with a resolution smaller than 32×32 pixels. In this paper, we primarily define small targets as those smaller than 32×32 pixels, and all three datasets mentioned below contain images that meet this criterion. Furthermore, medium-sized targets are defined as those with an area ranging from 32×32 to 96×96 pixels, and large targets are defined as those with an area greater than 96×96 pixels.

2.2. Multimodal Datasets

The DroneVehicle dataset [36], developed and annotated by Tianjin University, is a large-scale UAV-based vehicle dataset designed for aerial target detection. It comprises 28,439 pairs of aligned RGB-IR images, all categorized as vehicles and captured using UAV-mounted cameras. The dataset covers diverse regional scenes, including urban roads, residential areas, and highways, under both daytime and nighttime lighting conditions. To ensure comprehensive data variation, images were captured from three different altitudes: 80 m, 100 m, and 120 m, with camera angles set at 15°, 35°, and 45°. During the dataset calibration phase, affine transformations and region clipping were applied to crop and align RGB-IR image pairs, ensuring precise cross-modal correspondence. Sample images from the DroneVehicle dataset are shown in Figure 1, where the first row presents visible-light images, and the second row displays infrared images.

The KAIST dataset, proposed by Hwang et al. [22], was collected using an onboard camera and includes images from two modalities that have undergone rigorous registration processing. The dataset features pedestrian images captured from diverse scenes, such as

campus and urban roads, with image dimensions of 640 × 512. The training set comprises 8963 pairs of visible and infrared images, while the test set contains 2252 pairs. A subset of the KAIST dataset images is shown in Figure 2, with the first row displaying visible-light images and the second row displaying infrared images.

The LLVIP dataset, proposed by the research team from Beijing University of Posts and Telecommunications [37], utilizes binocular cameras and contains 16,836 pairs of carefully aligned images. As most of the images in this dataset are captured in nighttime environments, it is particularly well-suited for testing the performance of algorithms in nighttime scenarios. A subset of the LLVIP dataset image pairs is shown in Figure 3, with the first row displaying visible-light images and the second row showing infrared images.



Figure 1. A subset of images from the DroneVehicle dataset.



Figure 2. A subset of images from the KAIST dataset.



Figure 3. A subset of images from the LLVIP dataset.

In this paper, for the sake of consistency, we will use the symbol RGB to represent visible-light images and the symbol IR to represent infrared images.

2.3. Evaluation Metrics

To comprehensively assess the performance of multimodal image fusion algorithms in object detection tasks, this study employs both subjective and objective analysis methods. For subjective analysis, human evaluators visually assess the detection results of selected image samples, providing an initial performance judgment. In contrast, objective analysis uses various evaluation metrics to quantitatively measure the algorithm's performance on the test set, ensuring more accurate and reliable results. Specifically, for binary classification tasks, the predicted results of image targets are divided into positive and negative classes, which can be clearly represented through a confusion matrix. The confusion matrix compares the model's predictions with the actual results, yielding four types of classification outcomes:

- True Positive (TP): The number of samples that are actually positive and correctly identified as positive by the model;
- True Negative (TN): The number of samples that are actually negative and correctly identified as negative by the model;
- False Positive (FP): The number of samples that are actually negative but incorrectly identified as positive by the model;
- False Negative (FN): The number of samples that are actually positive but incorrectly identified as negative by the model.

This combined approach of subjective visual assessment and objective quantitative analysis provides a comprehensive understanding of the performance of multimodal fusion techniques in object detection tasks. MR^{-2} and AP can be computed using these four classification outcomes. MR^{-2} is a metric introduced by Dollar et al. to measure the performance of multimodal pedestrian detection. This metric is derived from the MR-FPPI curve, where MR represents the miss rate, which is plotted on the vertical axis of the MR-FPPI curve. The formula for calculating MR is as follows:

$$MR = \frac{FN}{TP + FN}, \tag{1}$$

where TP + FN reflects the total number of positive samples in the dataset, while FN refers to the number of positive samples that are incorrectly classified as negative by the model,

$$FPPI = \frac{FP}{N_{images}}, \tag{2}$$

where FP refers to the number of false positive detections in all images, and N_{images} is the total number of images in the dataset.

The calculation of MR^{-2} involves selecting 9 evenly spaced FPPI (False Positives Per Image) points within the interval [0.01, 1.0] on a logarithmic scale, then averaging the corresponding miss rates (MRs) for these points. As a result, a decrease in the MR^{-2} value directly reflects a reduction in missed detections, indicating an improvement in the algorithm's performance in target detection and recognition. This evaluation method not only provides an accurate quantification of model performance but also enables comparison of performance differences between various algorithms.

The calculation of AP is similar to that of MR^{-2} and is also based on the P-R (precision–recall) curve. AP refers to the average of the precision values (P) at different recall levels (R), where precision (P) is the proportion of true positive samples among all samples predicted as positive. The equation for calculation is as follows:

$$P = \frac{TP}{TP + FP}, \tag{3}$$

where R represents recall, which indicates the number of true positive samples correctly predicted from the actual positive samples. The calculation equation is as follows:

$$R = \frac{TP}{TP + FN}, \tag{4}$$

$$AP = \int_0^1 P(R)dR. \tag{5}$$

The higher the AP value, the better the performance of the algorithm in target detection and recognition. mAP (mean average precision) is a commonly used metric for evaluating the performance of object detection models. It is the average of the AP values across all categories. The calculation equation is as follows:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i, \tag{6}$$

where N is the number of categories, and AP_i is the average precision for the ith category.

2.4. Overall Framework

The method primarily comprises a dual-stream backbone network, a light sensing module, a cross-attention module, and the Neck and Detection Head of the YOLOv5 network. The core function of the light sensing weight module is to extract light information from visible-light images by compressing their spatial features. It then calculates weight values for the two modalities using a normalization equation, dynamically adjusting their contribution ratios during the fusion process to avoid the early suppression of informative features under varying illumination conditions.

To reduce the risk of information loss, especially for small targets with sparse and low-resolution features, a cross-attention module is introduced at multiple feature extraction stages. Unlike conventional late fusion methods that merge features after significant downsampling, our approach integrates cross-attention in earlier layers, enabling information exchange between infrared and visible-light modalities before spatial resolution is significantly reduced. This proactive fusion strategy enhances fine-grained feature retention and preserves critical small target details.

The specific architecture is shown in Figure 4. In the dual-stream backbone network, visible-light and infrared images serve as inputs. First, the visible-light image is processed by the light sensing weight module to obtain the light weight. Then, both images undergo four down sampling operations. After the first, second, and third down sampling steps, the feature maps from both modalities simultaneously enter the cross-attention module for feature enhancement. The first cross-attention module generates infrared feature Gir0 and visible-light feature Grgb0. Grgb0 and Frgb0 are added element-wise, and after convolution and down sampling, the result is Frgb1. Similar operations are performed for the infrared features. This process continues for subsequent layers. At the same time, features Frgb1 and Fir1, Frgb2 and Fir2, as well as the final Frgb3 and Fir3 features, are fused using the light weights calculated by the light sensing weight module, resulting in features G0, G1, and G2, which are then fed into the subsequent detection head for classification and localization.

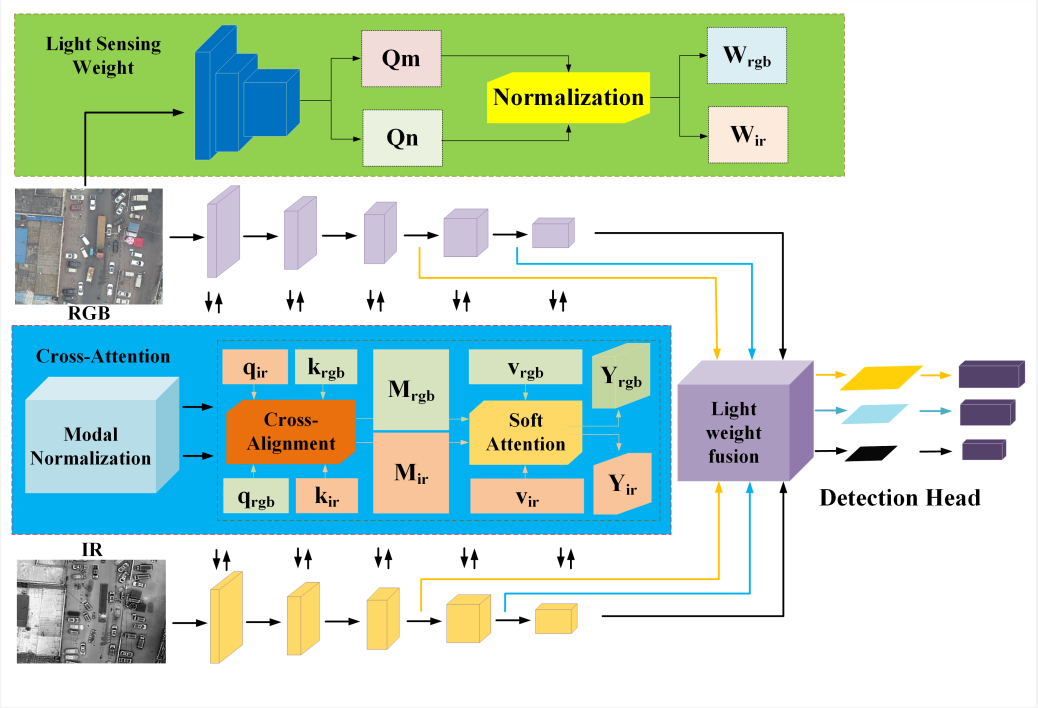


Figure 4. Overall framework of the model.

2.5. Light Sensing Weight Module

Many existing image fusion algorithms tend to overlook the impact of varying lighting conditions on the fusion process. This oversight can lead to the loss of valuable feature information from both infrared and visible-light images, thereby reducing the performance of all-weather image fusion. In such cases, critical details may be lost, and important features may not be fully explored, affecting the overall quality and effectiveness of the fused image. To address this issue, a light sensing weight module is designed. This module utilizes a miniaturized neural network to obtain the corresponding weights for the visible and infrared modalities. It preserves more effective information from both image sources, enhances detail retention, and ensures stronger robustness under various lighting conditions [38]. The module architecture is shown in Figure 5.

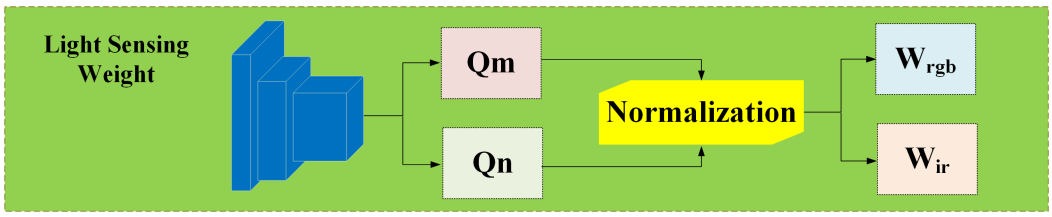


Figure 5. Light sensing weight module.

The specific content is as follows: Given a visible-light image I_{rgb}, the light sensing process is shown in Equation (7):

$$\{Q_m, Q_n\} = M_{lsm}(I_{rgb}), \tag{7}$$

where M_{lsm} represents the light sensing weight module, and Q_m and Q_N represent the light distribution values for daytime and nighttime input images, respectively, and are non-negative scalars. The design of the light sensing weight module aims to effectively address the issue of small-target detection under varying lighting conditions, based on the theoretical framework of light modeling and multimodal image fusion. In image processing,

variations in lighting conditions significantly affect the quality and informational content of RGB and IR, especially in low-light or no-light environments. In such conditions, infrared images provide more stable target information, while RGB images offer more detailed features under sufficient lighting. Therefore, designing a light sensing module that can adaptively adjust the contributions of RGB and IR images to target detection according to different lighting conditions is of great importance.

The specific steps of this module are as follows: First, CNN is used to process the input images and extract light-related features. The convolutional network effectively compresses spatial information and captures local lighting features, making it sensitive to changes in scene brightness. Next, a global max pooling layer integrates these lighting features, capturing the global lighting information of the entire image, which ensures that the model can make effective judgments based on overall lighting conditions. Then, two fully connected layers are used to compute the lighting distribution values, with a ReLU activation function applied to ensure non-negative outputs, avoiding negative weights and ensuring the validity of the light weights. Finally, a normalization function is used to process the calculated lighting weights, dynamically adjusting the weights of RGB and IR images based on different environmental conditions. The weight allocation mechanism uses a simple normalization function, as shown in Equation (8):

$$\begin{cases} W_{ir} = \frac{Q_n}{Q_m + Q_n} \\ \\ W_{rgb} = \frac{Q_m}{Q_m + Q_n} \end{cases}, \tag{8}$$

where W_{ir} and W_{rgb} represent the respective weights of the infrared image and the visible-light image in the feature fusion process.

In well-lit environments, RGB images are given a higher weight, as they provide more visual information. In low-light or no-light conditions, the IR image is given a higher weight, enhancing the stability and reliability of target detection. This weight distribution mechanism ensures that the model can adaptively optimize the fusion ratio of different modalities under various lighting conditions, thereby improving detection accuracy.

The theoretical basis of this module comes from research on light perception and multimodal image fusion, particularly in low-light environments where IR images can supplement RGB images by compensating for potential information loss. Meanwhile, the light sensing mechanism dynamically adjusts the fusion ratio of features through an effective weighting process, improving the model's robustness and accuracy in complex lighting conditions. Additionally, the normalization function used for weight distribution is a simple yet effective approach to dynamically adjust the influence of different modalities during the feature fusion process, enhancing the model's adaptability to lighting changes. Through this light perception-based feature fusion, the model not only achieves significant improvements in accuracy but also handles small-target detection tasks under varying lighting conditions.

2.6. Cross-Attention Module

To address the differences between modalities and effectively extract complementary features, this section introduces two key modules.

2.6.1. Modal Normalization

Since visible-light images and infrared images belong to different spectral bands, there is an inherent modal difference between them. Directly using features from one modality to calculate similarity scores with features from another modality is not an optimal solution.

Therefore, inspired by [39], this section proposes the use of modal normalization for transformation, as illustrated in Figure 6. The figure demonstrates the modal normalization process, where the mean and variance of one modality are used to reduce the modal differences when transforming the features of the other modality.

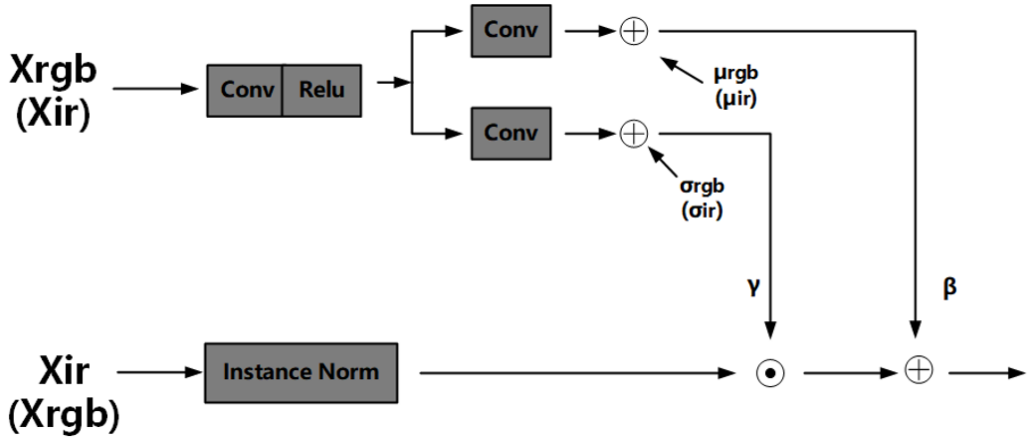


Figure 6. Modal normalization.

The transformation process for the two modalities is the same, so let us take the modality normalization of visible-light features as an example. First, X_{rgb} represents the input feature. The following is the Z-score normalization equation [40], where μ_{rgb} and σ_{rgb} represent the standard deviation tensor and mean tensor of the visible-light image features, respectively. The distribution of X_{rgb} is remapped using Equation (9). W_b, W_d and W_g represent 3×3 convolutions, which are used to predict two learnable parameter tensors β_{ir} and γ_{ir}, as shown in Equation (10). Finally, using Equation (11), the previously obtained parameters are recalculated to obtain the normalized infrared feature $\mathbf{x_{ir}}$ and visible-light feature x_{rgb}.

$$\tilde{X}_{rgb} = \frac{X_{rgb} - \mu_{rgb}}{\sigma_{rgb}}, \tag{9}$$

$$\begin{cases} \beta_{rgb} = W_b * \mathrm{relu}(W_d * X_{ir}) + \mu_{ir} \\ \gamma_{rgb} = W_g * \mathrm{relu}(W_d * X_{ir}) + \sigma_{ir} \end{cases}, \tag{10}$$

$$\begin{cases} x_{ir} = x_{ir} \cdot \gamma_{ir} + \beta_{ir} \\ x_{rgb} = x_{rgb} \cdot \gamma_{rgb} + \beta_{rgb} \end{cases}. \tag{11}$$

The association alignment phase following modality normalization allows for complementary refinement of feature representations. By aligning the features from both modalities, this phase improves the model's ability to capture more precise and coherent information, which, in turn, enhances the overall detection performance of the model.

2.6.2. Attention Mechanism Module

The cross-attention mechanism has a broad theoretical foundation in multimodal learning and object detection tasks [41]. Multimodal learning aims to improve a model's detection capabilities by fusing data from different sources, such as RGB and infrared images. In complex environments, information from a single modality often fails to capture the full range of target features. The cross-attention mechanism establishes interactions between different modalities, enabling the model to extract complementary feature information from each modality, thereby enhancing small-object detection performance.

The theoretical foundation of the cross-attention mechanism originates from the self-attention mechanism and the attention mechanism in Transformer architectures. These techniques were initially applied in natural language processing tasks and later successfully adapted to image processing and visual tasks. In traditional attention mechanisms, the model automatically assigns weights to input features to learn important areas or information. In self-attention, the weights of features are dynamically calculated based on the correlations between them, while the cross-attention mechanism extends this idea by enabling information interaction and weighting between two modalities, ensuring that each modality's features effectively complement the other.

Specifically, the cross-attention mechanism first calculates the similarity between different modalities, capturing the relationship between key features in one modality and related features in another. By weighting these relationships, the model learns the contribution of each modality to object detection, thus improving the fusion of features from both modalities. For example, in the fusion of RGB and infrared images, the cross-attention mechanism guides the model to focus on the target areas in the infrared image while benefiting from the rich visual information provided by the RGB image, thereby enhancing detection accuracy. Unlike traditional attention mechanisms, the cross-attention mechanism not only focuses on features within a single modality but also achieves feature fusion through interactive weighting between the two modalities, maximizing the utility of each modality's information.

In multimodal object detection tasks, the cross-attention mechanism enhances the complementarity of information through interactions between different modalities. For instance, when features in one modality are limited or unclear, the other modality can compensate for these deficiencies, improving the model's robustness and accuracy. The cross-attention mechanism adaptively adjusts the information fusion process based on the actual conditions of each modality, ensuring that the model maintains high detection accuracy under various environments and conditions.

The application of the cross-attention mechanism not only improves detection accuracy but also effectively addresses issues of information redundancy and loss in multimodal fusion. Through this mechanism, the model can focus on complementary features from different modalities, maximizing each modality's advantages during fusion and ultimately improving overall performance.

Theoretically, the cross-attention mechanism is based on the idea of attention mechanisms, leveraging the complementarity between modalities to enhance object detection capabilities. Particularly in small-object detection tasks, the cross-attention mechanism can effectively capture subtle features, improving object recognition and localization. This mechanism has been widely applied in multimodal object detection and has proven effective, especially in complex scenes with significant lighting and viewpoint variations.

As shown in Figure 7, in the cross-attention module of this section, suppose the input features are the visible-light features x_{rgb} and infrared features x_{ir}. First, the input features are used to generate query vectors q, key vectors k, and value vectors v descriptors. For each modality, the q, k, and v descriptors are computed with themselves to obtain the attention weights. In the bimodal model used in this section, the cross-attention module, as shown in Figure 7, computes the dot product between the query vector q derived from the visible-light features x_{rgb} and the key vector k derived from the infrared features x_{ir}, resulting in a similarity matrix M_{ir}. Then, this similarity matrix is used for soft attention fusion with the value vector v derived from the infrared features x_{ir}. This module can dynamically supplement the features based on the attention weights and inject contextual feature

information, making the complementary features more robust.The similarity matrices M_{ir} and M_{rgb} are calculated by the cross-correlation alignment step, as shown in Equation (12),

$$\begin{cases} M_{rgb} = \text{Softmax}\left(\text{MatMul}\left(q_{ir}, k_{rgb}^{T}\right)/\sqrt{d}\right) \\ M_{ir} = \text{Softmax}\left(\text{MatMul}\left(q_{rgb}, k_{ir}^{T}\right)/\sqrt{d}\right) \end{cases}, \tag{12}$$

where MatMul represents matrix multiplication, T denotes the transpose of a matrix, and $\sqrt{d}$ is the scaling factor. Then, a softmax layer is applied to obtain the normalized cross-correlation matrices M_{ir} and M_{rgb}.

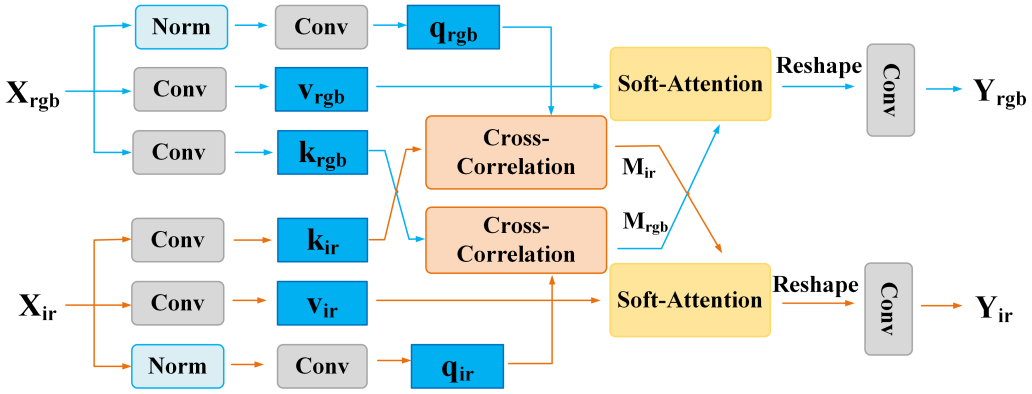


Figure 7. Cross-attention module.

2.6.3. Loss Function and Optimization Method

The loss function in the YOLOv5 network mainly consists of three components: bounding box regression loss (L_{box}), objectness loss (L_{obj}), and classification loss (L_{cls}) [14]. The specific loss function can be represented by the following Equation (13):

$$\begin{aligned} L_{total} &= \sum_{i}^{N}(\lambda_1 L_{box} + \lambda_2 L_{obj} + \lambda_3 L_{cls}) \\ &= \sum_{i}^{N}(\lambda_1 \sum_{j}^{B_i} L_{CIoU_j} + \lambda_2 \sum_{j}^{S_i \times S_i} L_{obj_j}, \\ &+ \lambda_3 \sum_{j}^{B_i} L_{cls_j}) \end{aligned} \tag{13}$$

where N represents the number of detection layers, and B refers to the number of targets assigned to each anchor box. Additionally, S × S represents the number of grid cells the network divides the input into, i and j represent the grid indices in the feature map. In this network model, the light sensing weight module is used to enhance the rationality of the fusion, which adds a fusion loss function L_1, as shown in the following Equation (14):

$$L = L_1 + L_{box} + L_{obj} + L_{cls}. \tag{14}$$

The fusion loss function L_1 is also composed of several parts, where $\mathbf{L_f}$ represents the light sensing loss, defined as follows:

$$L_f = L_i^{ir} \cdot W_{ir} + L_i^{rgb} \cdot W_{rgb}, \tag{15}$$

where L_i^{ir} and L_i^{rgb} represent the intensity loss for infrared and visible-light images, respectively, and W_{ir} and W_{rgb} represent the light sensing weights for infrared and visible-light

images, respectively. These weights are specifically defined in Equation (8). The intensity loss measures the difference between the fused image and the source images at the pixel level. Therefore, the intensity loss for the infrared and visible-light images can be defined as follows:

$$\begin{cases} L_i^{ir} = \| I_f - I_{ir} \|_1 \frac{1}{HW} \\ L_i^{rgb} = \| I_f - I_{rgb} \|_1 \frac{1}{HW} \end{cases}, \tag{16}$$

where H and W represent the height and width of the input image, respectively, and $\|\cdot\|_1$ denotes the L_1 norm. I_f is the image after feature fusion. To ensure that the fused image maintains consistent light intensity with the original images, the light sensing weights W_{ir} and W_{rgb} are used to control the light distribution in the fused image.

Although the light loss function can automatically adjust according to light conditions to preserve the original image's light information, this alone is not sufficient to ensure the optimal state of the intensity distribution in the fused image. Therefore, additional auxiliary intensity loss needs to be introduced for further optimization:

$$L_{inten} = \| I_f - \max(I_{ir}, I_{rgb}) \|_1 \frac{1}{HW}. \tag{17}$$

In extensive experiments, it has been demonstrated that the best texture performance for infrared and visible-light images is achieved when the image textures are concentrated. Therefore, to maintain both optimal brightness distribution and rich texture details in the fused image, the concept of texture loss is adopted, which aims to enrich the texture information in the fused image. The definition of texture loss is as follows:

$$L_{texture} = \| |\nabla I_f| - \max(|\nabla I_{ir}|, |\nabla I_{rgb}|) \|_1 \frac{1}{HW}, \tag{18}$$

where ∇ represents the Sobel gradient operator used to measure the image texture information by calculating the gradients. Finally, the light sensing loss L_f, auxiliary intensity loss L_{inten}, and texture loss $L_{texture}$ are combined through weighted coefficients λ to obtain the fusion loss functionL_1, which is expressed as follows:

$$L_1 = \lambda_1 \cdot L_f + \lambda_2 \cdot L_{inten} + \lambda_3 \cdot L_{texture}. \tag{19}$$

The network, under the constraints of light sensing loss and auxiliary intensity loss, dynamically adjusts the light intensity distribution of the generated image based on the input light scene, enabling it to better reflect the complex lighting conditions found in the real world. Additionally, the texture loss constraint helps preserve the texture details of the original image in the generated output. Consequently, the generated image not only excels in terms of light adjustment but also maintains high texture quality. As a result, the light sensing weight module enhances both the quality of image fusion and the network's adaptability and stability in varying lighting conditions, providing a solid foundation for multimodal detection tasks.

3. Results

This section presents both quantitative and qualitative evaluations of the proposed detection method, with experiments conducted on three datasets: DroneVehicle, LLVIP, and KAIST.

3.1. Configuration and Parameter Settings

All experiments in this paper were conducted under the experimental configurations listed in Table 1. In consideration of the memory limitations of the experimental equipment

and the duration of the experiment, the epoch is set to 16, the batch size is chosen to be 4, and the initial learning rate is set to 0.0001.

Table 1. Experimental configuration.

Environment Configuration	Information
CPU	Intel i5-8400
GPU	GeForce RTX 3090
Operating System	Windows10
Development Language	Python 3.8.3
Cuda	Cuda 11.1

3.2. Ablation Experiment Analysis

This section conducts ablation experiments on the DroneVehicle dataset to demonstrate the effectiveness of each module. As shown in Table 2, LSM represents the light sensing weight module, MNM represents the modal normalization module, and AMM represents the attention mechanism module. The experiments compared networks using different combinations of modules, with the evaluation metrics being MR^{-2} and mAP50.

Table 2. Comparison of multimodal network recognition performance.

RGB	IR	LSM	MNM	AMM	MR^{-2}/%	mAP0.5/%
✓					25.1	78.5
	✓				25.4	78.9
✓	✓				26.1	79.7
✓	✓		✓		25.4	80.3
✓	✓			✓	23.6	81.2
✓	✓	✓			23.3	81.6
✓	✓	✓	✓		23.1	82.4
✓	✓		✓	✓	22.7	83.1
✓	✓	✓		✓	21.6	83.9
✓	✓	✓	✓	✓	**20.8**	**84.8**

Bolded values in the table represent the optimal value, and the same notation applies to subsequent tables.

The specific data are shown in the Table 2. When using the simple concat method to stack RGB and IR data, the miss detection rate is 26.1% and the average precision is 79.7%. After adding the light sensing weight module, the miss detection rate improves by 2.8%, and precision increases by 1.9%. The specific fusion changes are shown in Figure 8. A simple pixel-wise addition leads to overexposure in daylight, which causes a loss of fine texture details. However, after balancing through the light sensing weight module, the advantages of good lighting conditions are better reflected. In low light at night, the visible-light image still retains some color and texture features of the target. Without light sensing, the infrared features become more prominent. However, after balancing through the light sensing weight module, the exposure of the visible-light image is increased, and while more features of the visible-light image are preserved, some targets that are not obvious in darkness are also enhanced. Under dark conditions, the visible-light image shows almost no target features, and the changes in the infrared features after fusion become less noticeable. Therefore, the presence or absence of the light sensing weight module plays a critical role in multimodal fusion tasks.

When both the light sensing weight module and the cross-attention module are used together, the miss detection rate improves by 1.7% and precision increases by 2.3% compared to the network with only the light module. This further confirms the importance of the cross-attention mechanism in the method. As shown in Figure 9, whether it is

targets at the edges in daylight or subtle features that are not noticeable at night, the cross-attention module allows the network to focus more on target information and enhances the features. When the modality normalization is added at the end, the miss detection rate improves by 0.8% and precision increases by 0.7%, proving that modality normalization can effectively reduce the differences between the two modalities, thereby enhancing the detection performance.

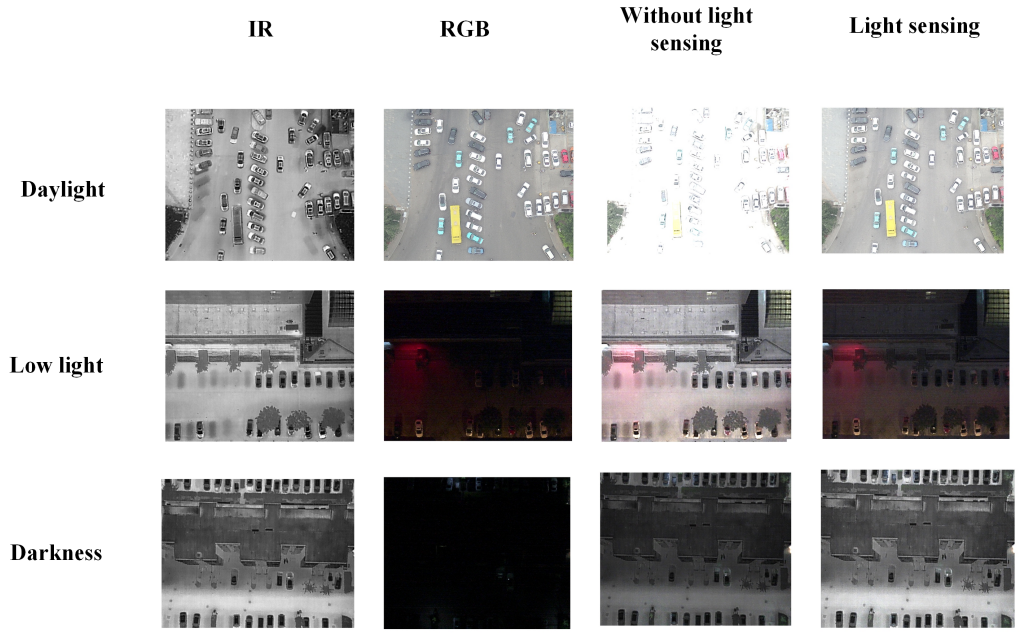


Figure 8. Light sensing fusion comparison.

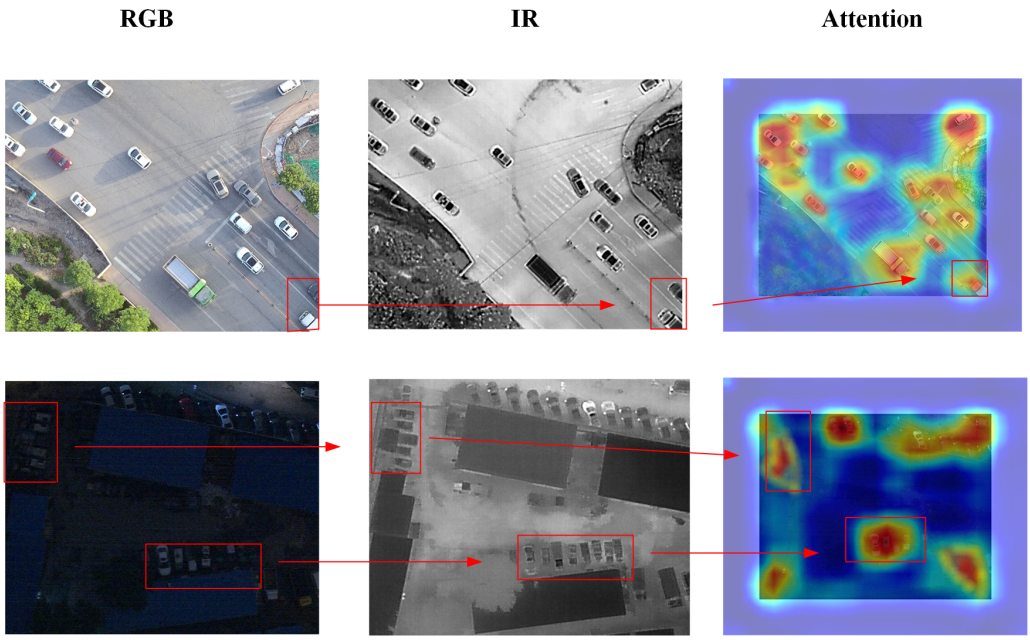


Figure 9. Comparison of feature map visualization.

Figure 10 presents the validation results of the proposed method on the DroneVehicle dataset, with the upper and lower images representing scenes in daylight and at night, respectively, to demonstrate the feasibility of the method. From the first row of images, it is clear that, under good lighting conditions with no obstructions, the detection results are accurate. Even the half-visible car in the top-left corner is correctly detected. Although the

small portion of the car in the bottom-right corner is not detected, this is due to incomplete features and low confidence, which is not a significant issue as it does not affect the overall performance of the method. The second row of images focuses on the nighttime scene, where only part of the car's front is visibly distinct in the visible-light image. However, with the proposed method, the advantages of combining infrared images allow for accurate detection of trucks hidden in the shadows, further demonstrating the superiority of the method.

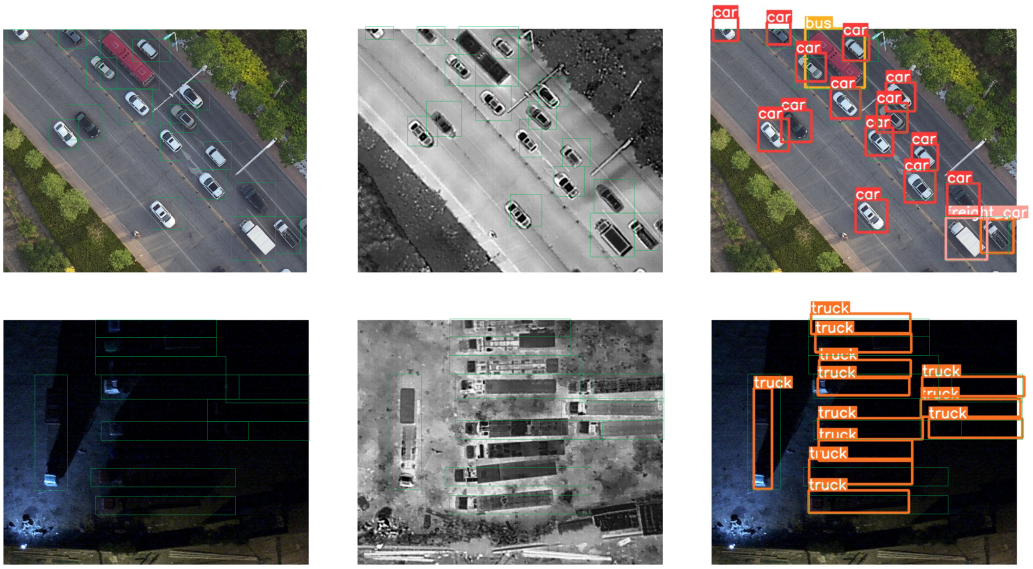


Figure 10. Day and night detection results on the DroneVehicle dataset.

The introduction of the light sensing weight module, cross-attention module, and their complementary modal normalization components demonstrates excellent performance in multimodal tasks. These modules not only significantly enhance the algorithm's performance but also have important practical significance in real-world applications.

3.3. Comparison with Other Algorithms

This section demonstrates the effectiveness of the proposed light sensing weight module in enhancing UAV small-object detection and recognition performance, further improved by the cross-attention module. To evaluate the generalization capability of the proposed method, experiments were conducted on different multimodal datasets, comparing the results with existing algorithms. The selected benchmark datasets are the publicly available KAIST and LLVIP datasets, both primarily designed for pedestrian detection but also containing small objects, making them suitable for testing the algorithm's generalization ability and performance in small-object detection. In the comparative experiments, we introduced the IAF-RCNN algorithm, which performs well with lightweight fusion strategies, the IA-YOLO algorithm, which achieves optimal performance under low-light conditions, and the AR-CNN algorithm, which demonstrates superior overall performance. The MR^{-2} missed detection rate comparison of different algorithms across the KAIST dataset for all-day (AD), daytime (DT), and nighttime (NT) scenarios is presented in Table 3. Through the performance analysis of MR^{-2} in Table 3 and Figure 11, the multimodal fusion method based on light sensing and cross-attention mechanisms proposed achieved favorable MR^{-2} values under all-day, daytime, and nighttime lighting conditions.

Additionally, this section evaluates the detection and recognition performance of other algorithms on the LLVIP dataset under the AP metric, as shown in Table 4. On the LLVIP dataset, the proposed method outperforms other advanced YOLO models and

multimodal algorithms in terms of detection performance, achieving the highest values in the mAP50 metric. Although this method incurs an increase in parameter count and inference time, the trade-off remains within an acceptable range while demonstrating significant advantages in detection accuracy. This further indicates that the proposed algorithm exhibits strong generalization capability.

Table 3. Comparison of missed detection rates of different algorithms on the KAIST dataset.

Algorithm	AD MR^{-2}/%	DT MR^{-2}/%	NT MR^{-2}/%
ACF [22]	47.32	42.57	56.15
Halfway Fusion [25]	25.75	24.88	26.59
SSD-Vis	18.33	19.06	17.78
Fusion-RPN [42]	18.29	19.57	16.27
IAF-RCNN [28]	15.73	14.55	18.26
SSD-Multiscale	15.66	17.85	11.39
MSDS-RCNN [26]	11.63	10.60	13.73
IPWF+ECA	11.16	12.40	8.46
IA-YOLO	9.52	11.28	**7.76**
AR-CNN [43]	9.34	9.94	8.38
ours	**9.3**	**9.83**	8.54

Table 4. Comparison of AP of different algorithms on the LLVIP dataset.

Algorithm	Dataset	mAP50/%	Param./M	Runtime/ms
YOLOv5s	RGB	90.8	7.2	10.7
YOLOv8s	RGB	92.7	11.2	12.6
YOLOv9s	RGB	94.2	**7.1**	14.6
YOLOv10s	RGB	93.5	7.2	14.1
YOLOv11s	RGB	95.7	9.4	13.8
CFT [44]	RGB + IR	97.5	11.2	16.7
Infusion-Net [45]	RGB + IR	98.6	23.8	19.9
LRAF-Net [46]	RGB + IR	97.9	18.8	21.4
CAMF [47]	RGB + IR	89.01	31.4	**9.3**
ours	RGB + IR	**98.8**	19.7	17.2

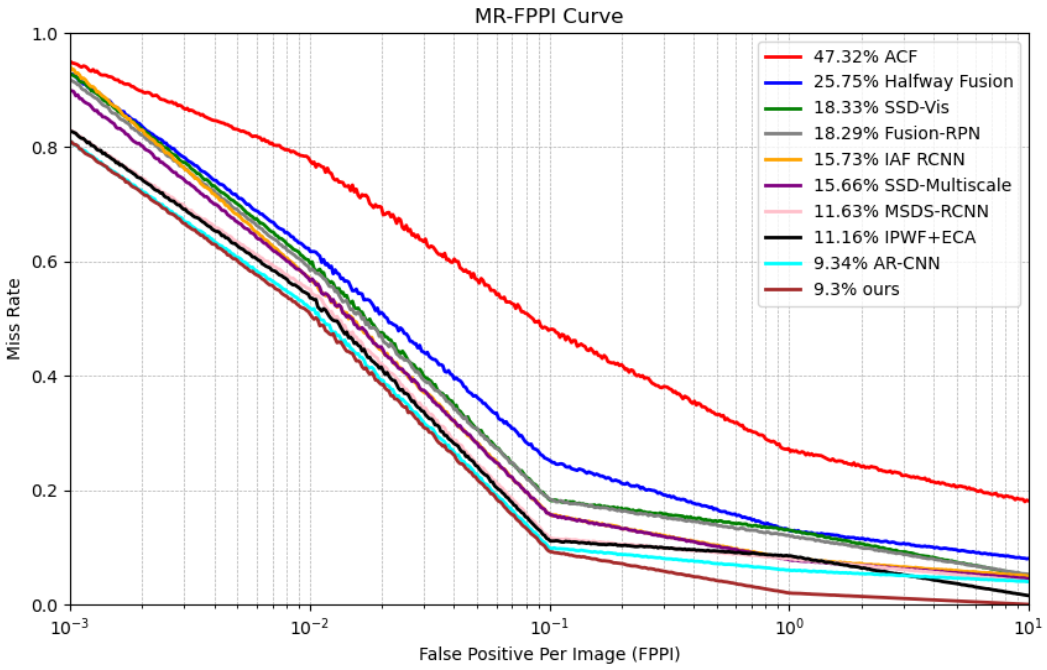


Figure 11. MR-FPPI curves of different algorithms.

3.4. Detection Results Visualization

To visually verify the recognition performance of the algorithm presented in this paper, Figures 12 and 13 compare the detection results of the proposed algorithm with the existing IAF-RCNN algorithm, which performs well and incorporates illumination perception, under both daytime and nighttime scenarios. The first row shows the original annotations, the second row shows the detection results of the IAF-RCNN algorithm, and the third row shows the results of the method we proposed. The green boxes represent the original bounding boxes, while the red boxes represent the detected bounding boxes.

Figure 12 presents the detection results in a daytime scene. Although the IAF-RCNN algorithm can accurately detect pedestrian targets, it still fails to detect small targets, as shown in the two images on the far right. The pedestrians at the edges and those walking in pairs at a distance are not detected. This may be due to insufficient target features and the image not being sufficiently fused, causing the network to lose information about these small targets, leading to missed detections. In contrast, the method proposed utilizes cross-attention to further enhance the fusion of information, helping reduce missed detections. Additionally, the light sensing weight module balances the light intensity, enhancing the lighting conditions in nighttime scenes, which is more conducive to target detection.



Figure 12. Comparison of detection results from different algorithms in daytime scenario.

Figure 13 shows detection results in a nighttime scene where pedestrians often blend into the background. The IAF-RCNN algorithm, combined with the light sensing weight module, performs well but struggles with edge cases, overlapping pedestrians, and occasional misidentification of roadside trees due to complex lighting conditions. In contrast, the proposed method leverages the cross-attention mechanism to better integrate infrared information, focusing on target features even when obstacles are present. This improves detection in low-light and challenging environments.

Furthermore, experiments were conducted on the LLVIP dataset to verify the generalization ability of the proposed method, with results shown in Figure 14. In the first image, with good lighting, the cyclist is correctly identified. However, a child riding ahead of the adult was not detected, likely due to the extreme overlap and similar features between the two. Despite this, the method's overall generalization ability remains strong. Notably, in the third image, where the target is heavily obstructed by trees and the lighting is dim,

the proposed method still detects the target, demonstrating the excellent generalization ability of the method we proposed.



Figure 13. Comparison of detection results from different algorithms in nighttime scenario.



Figure 14. Pedestrian detection results on the LLVIP dataset.

4. Discussion

In response to the existing models that neglect the impact of lighting intensity on target recognition performance, this study introduces a lighting-aware weight module to address this issue. This module improves image fusion by considering the lighting intensity

contribution of the image. Additionally, to ensure that each modality compensates for the weaknesses of the other before fusion, a cross-modal attention mechanism is introduced to extract and enhance complementary features.

First, through ablation experiments on the individual modules, we validate the key role of the lighting-aware weight module and the cross-attention mechanism in the algorithm. The experimental results show that, on the DroneVehicle dataset, the baseline fusion network has a miss detection rate of 26.1% and an accuracy of 79.7%. After adding the lighting-aware weight module, the miss detection rate is reduced by 2.8%, and the average accuracy increases by 1.9%. When only modality normalization and the cross-attention module are added, the miss detection rate drops by 3.4%, and the average accuracy improves by 3.4%. Finally, after incorporating all modules, the miss detection rate decreases by 5.3%, and the accuracy improves by 5.1%. These results demonstrate that these two modules not only effectively balance lighting intensity but also enhance target information by utilizing complementary features from both modalities, significantly improving detection accuracy and reducing the miss detection rate.

To further validate the performance of the algorithm, we applied it to the publicly available KAIST and LLVIP datasets for pedestrian detection and compared it with several current algorithms. The experimental results show that the algorithm exhibits strong generalization ability, accurately detecting targets even when the target features are unclear or occluded, and demonstrates superior pedestrian recognition performance.

5. Conclusions

This paper presents a method for small-target detection on aircraft based on multi-modal feature fusion. The method utilizes the cross-attention mechanism to enable mutual learning of complementary features between modalities, while continuously enhancing target detection. The light fusion module learns the image's light intensity to generate light weights, ensuring that the fused image effectively accounts for the lighting impact and achieves optimal feature fusion. The method is analyzed and validated on various datasets, demonstrating its effectiveness and good generalization ability.

Although the proposed method demonstrates excellent performance in detection, significantly improving detection accuracy and reducing the miss detection rate, the introduction of various attention modules and the lighting-aware weight module has increased the model's complexity, leading to a reduction in detection speed. Additionally, while the early fusion strategy benefits small-object detection, it may further impact inference efficiency, potentially becoming a bottleneck in real-time applications that require fast responses. This issue could become a bottleneck in real-time applications that require fast responses. To balance accuracy and speed, future research will focus on optimizing these modules and exploring methods to reduce the model's parameters and computational cost while maintaining high accuracy. On one hand, lightweight neural network architectures can be designed to reduce redundant computations; on the other hand, more efficient attention mechanisms, such as adaptive attention modules, may help maintain low computational complexity while improving detection performance. Additionally, model distillation techniques could serve as a potential optimization approach, helping to compress the model and accelerate inference. Ultimately, the goal is to further enhance the detection speed and real-time response capabilities of the network to meet the requirements of various application scenarios, including autonomous driving, security surveillance, and other fields.

Author Contributions: Conceptualization, S.X. and H.L.; methodology, S.X., H.L. and T.L.; software, S.X., H.L. and T.L.; investigation, S.X. and H.L.; formal analysis, S.X., H.L., T.L. and H.G.; writing—original draft preparation, S.X. and H.L.; writing—review and editing, S.X., H.L. and H.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Open Research Fund of Shaanxi Key Laboratory of Optical Remote Sensing and Intelligent Information Processing under Grant KF20230301, in part by the National Natural Science Foundation of China under Grant 62071168.

Data Availability Statement: The original contributions presented in the study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Tu, B.; Wang, Z.; Ouyang, H.; Yang, X.; Li, J.; Plaza, A. Hyperspectral anomaly detection using the spectral–spatial graph. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5542814.
2. Zhang, Y.; Xu, S.; Hong, D.; Gao, H.; Zhang, C.; Bi, M.; Li, C. Multimodal Transformer Network for Hyperspectral and LiDAR Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3283508. [CrossRef]
3. Su, Y.; Chen, J.; Gao, L.; Plaza, A.; Jiang, M.; Xu, X.; Sun, X.; Li, P. ACGT-Net: Adaptive cuckoo refinement-based graph transfer network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5521314.
4. Zhang, L.; Zhang, L. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 270–294.
5. Zhang, L.; Song, L.; Du, B.; Zhang, Y. Nonlocal low-rank tensor completion for visual data. *IEEE Trans. Cybern.* **2019**, *51*, 673–685.
6. Chen, Z.; Hong, D.; Gao, H. Grid network: Feature extraction in anisotropic perspective for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 5507105.
7. Su, Y.; Gao, L.; Jiang, M.; Plaza, A.; Sun, X.; Zhang, B. NSCKL: Normalized spectral clustering with kernel-based learning for semisupervised hyperspectral image classification. *IEEE Trans. Cybern.* **2022**, *53*, 6649–6662.
8. Zhuang, L.; Ng, M.K.; Gao, L.; Michalski, J.; Wang, Z. Eigenimage2Eigenimage (E2E): A self-supervised deep learning network for hyperspectral image denoising. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 16262–16276.
9. Wang, J.; Zhang, X.; Gao, G.; Lv, Y. OP mask R-CNN: An advanced mask R-CNN network for cattle individual recognition on large farms. In Proceedings of the 2023 International Conference on Networking and Network Applications (NaNA), Qingdao, China, 18–21 August 2023; pp. 601–606.
10. Wang, Z.; Cao, Y.; Li, J. A Detection Algorithm Based on Improved Faster R-CNN for Spacecraft Components. In Proceedings of the 2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA), Changchun, China, 11–13 August 2023; pp. 1–5.
11. Han, Z.; Hong, D.; Gao, L.; Yao, J.; Zhang, B.; Chanussot, J. Multimodal hyperspectral unmixing: Insights from attention networks. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5524913.
12. Gao, L.; Li, J.; Zheng, K.; Jia, X. Enhanced autoencoders with attention-embedded degradation learning for unsupervised hyperspectral image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5509417.
13. Chen, Z.; Wang, Y.; Gao, H.; Ding, Y.; Zhong, Q.; Hong, D.; Zhang, B. Temporal difference-guided network for hyperspectral image change detection. *Int. J. Remote Sens.* **2023**, *44*, 6033–6059.
14. Xu, S.; Chen, X.; Li, H.; Liu, T.; Chen, Z.; Gao, H.; Zhang, Y. Airborne small target detection method based on multi-modal and adaptive feature fusion. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5637215.
15. Lou, H.; Duan, X.; Guo, J.; Liu, H.; Gu, J.; Bi, L.; Chen, H. DC-YOLOv8: Small-size object detection algorithm based on camera sensor. *Electronics* **2023**, *12*, 2323. [CrossRef]
16. Guo, J.; Lou, H.; Chen, H.; Liu, H.; Gu, J.; Bi, L.; Duan, X. A new detection algorithm for alien intrusion on highway. *Sci. Rep.* **2023**, *13*, 10667.
17. Wang, F.; Wang, H.; Qin, Z.; Tang, J. UAV target detection algorithm based on improved YOLOv8. *IEEE Access* **2023**, *11*, 116534–116544.
18. Wang, G.; Chen, Y.; An, P.; Hong, H.; Hu, J.; Huang, T. UAV-YOLOv8: A small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios. *Sensors* **2023**, *23*, 7190. [CrossRef]
19. Zhao, F.; Wei, R.; Chao, Y.; Shao, S.; Jing, C. Infrared bird target detection based on temporal variation filtering and a gaussian heat-map perception network. *Appl. Sci.* **2022**, *12*, 5679. [CrossRef]
20. Zhu, K.; Xu, C.; Wei, Y.; Cai, G. Fast-PLDN: Fast power line detection network. *J. Real-Time Image Process.* **2022**, *19*, 3–13.

21. Li, H.; Wu, X.J.; Kittler, J. Infrared and visible image fusion using a deep learning framework. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 2705–2710.
22. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; So Kweon, I. Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1037–1045.
23. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.
24. Wagner, J.; Fischer, V.; Herman, M.; Behnke, S.; et al. Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks. In Proceedings of the ESANN, Bruges, Belgium, 27–29 April 2016; Volume 587, pp. 509–514.
25. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral deep neural networks for pedestrian detection. *arXiv* **2016**, arXiv:1611.02644.
26. Li, C.; Song, D.; Tong, R.; Tang, M. Multispectral pedestrian detection via simultaneous detection and segmentation. *arXiv* **2018**, arXiv:1808.04818.
27. Guan, D.; Cao, Y.; Yang, J.; Cao, Y.; Yang, M.Y. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Inf. Fusion* **2019**, *50*, 148–157.
28. Li, C.; Song, D.; Tong, R.; Tang, M. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognit.* **2019**, *85*, 161–171.
29. Zhang, L.; Liu, Z.; Zhang, S.; Yang, X.; Qiao, H.; Huang, K.; Hussain, A. Cross-modality interactive attention network for multispectral pedestrian detection. *Inf. Fusion* **2019**, *50*, 20–29.
30. Zhang, H.; Fromont, E.; Lefevre, S.; Avignon, B. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 276–280.
31. 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, Virtual, 5–9 January 2021; pp. 72–80.
32. Zou, X.; Peng, T.; Zhou, Y. UAV-Based Human Detection With Visible-Thermal Fused YOLOv5 Network. *IEEE Trans. Ind. Inform.* **2023**, *20*, 3814–3823.
33. Lee, S.; Kim, T.; Shin, J.; Kim, N.; Choi, Y. INSANet: INtra-INter spectral attention network for effective feature fusion of multispectral pedestrian detection. *Sensors* **2024**, *24*, 1168. [CrossRef]
34. Xie, Q.; Cheng, T.Y.; Dai, Z.; Tran, V.; Trigoni, N.; Markham, A. Illumination-Aware Hallucination-Based Domain Adaptation for Thermal Pedestrian Detection. *IEEE Trans. Intell. Transp. Syst.* **2023**, *25*, 315–326.
35. Chen, C.; Liu, M.Y.; Tuzel, O.; Xiao, J. R-CNN for small object detection. In Proceedings of the Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; Revised Selected Papers, Part V 13; Springer: Berlin/Heidelberg, Germany, 2017; pp. 214–230.
36. Sun, Y.; Cao, B.; Zhu, P.; Hu, Q. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6700–6713.
37. Jia, X.; Zhu, C.; Li, M.; Tang, W.; Zhou, W. LLVIP: A visible-infrared paired dataset for low-light vision. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3496–3504.
38. Zhou, K.; Chen, L.; Cao, X. Improving multispectral pedestrian detection by addressing modality imbalance problems. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XVIII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 787–803.
39. Kim, J.; Lee, J.K.; Lee, K.M. Deeply-recursive convolutional network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1637–1645.
40. Ulyanov, D. Instance normalization: The missing ingredient for fast stylization. *arXiv* **2016**, arXiv:1607.08022.
41. Zhang, Y.; Gao, H.; Zhou, J.; Zhang, C.; Ghamisi, P.; Xu, S.; Li, C.; Zhang, B. A cross-modal feature aggregation and enhancement network for hyperspectral and LiDAR joint classification. *Expert Syst. Appl.* **2024**, *258*, 125145.
42. 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 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 June 2017; pp. 49–56.
43. Zhang, L.; Zhu, X.; Chen, X.; Yang, X.; Lei, Z.; Liu, Z. Weakly aligned cross-modal learning for multispectral pedestrian detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5127–5137.
44. Qingyun, F.; Dapeng, H.; Zhaokui, W. Cross-modality fusion transformer for multispectral object detection. *arXiv* **2021**, arXiv:2111.00273.
45. Yun, J.S.; Park, S.H.; Yoo, S.B. Infusion-net: Inter-and intra-weighted cross-fusion network for multispectral object detection. *Mathematics* **2022**, *10*, 3966. [CrossRef]

46. Fu, H.; Wang, S.; Duan, P.; Xiao, C.; Dian, R.; Li, S.; Li, Z. Lraf-net: Long-range attention fusion network for visible–infrared object detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 13232–13245.
47. Tang, L.; Chen, Z.; Huang, J.; Ma, J. Camf: An interpretable infrared and visible image fusion network based on class activation mapping. *IEEE Trans. Multimed.* **2023**, *26*, 4776–4791.

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.