drones

Article

MFMG-Net: Multispectral Feature Mutual Guidance Network for Visible–Infrared Object Detection

Fei Zhao [1,2], Wenzhong Lou [1,2,*], Hengzhen Feng [1,2,*], Nanxi Ding [1,2] and Chenglong Li [1,2]

[1] School of Mechatronical Engineering, Beijing Institute of Technology, Beijing 100081, China; 3120205110@bit.edu.cn (F.Z.); 3120225105@bit.edu.cn (N.D.); 3220235232@bit.edu.cn (C.L.)
[2] Chongqing Innovation Center, Beijing Institute of Technology, Chongqing 401120, China
[*] Correspondence: louwz@bit.edu.cn (W.L.); fenglei@bit.edu.cn (H.F.)

Abstract: Drones equipped with visible and infrared sensors play a vital role in urban road supervision. However, conventional methods using RGB-IR image pairs often struggle to extract effective features. These methods treat these spectra independently, missing the potential benefits of their interaction and complementary information. To address these challenges, we designed the Multispectral Feature Mutual Guidance Network (MFMG-Net). To prevent learning bias between spectra, we have developed a Data Augmentation (DA) technique based on the mask strategy. The MFMG module is embedded between two backbone networks, promoting the exchange of feature information between spectra to enhance extraction. We also designed a Dual-Branch Feature Fusion (DBFF) module based on attention mechanisms, enabling deep feature fusion by emphasizing correlations between the two spectra in both the feature channel and space dimensions. Finally, the fused features feed into the neck network and detection head, yielding ultimate inference results. Our experiments, conducted on the Aerial Imagery (VEDAI) dataset and two other public datasets (M3FD and LLVIP), showcase the superior performance of our method and the effectiveness of MFMG in enhancing multispectral feature extraction for drone ground detection.

Keywords: RGB-IR image pair; multispectral feature; object detection; attention mechanism

check for updates

Citation: Zhao, F.; Lou, W.; Feng, H.; Ding, N.; Li, C. MFMG-Net: Multispectral Feature Mutual Guidance Network for Visible–Infrared Object Detection. *Drones* **2024**, *8*, 112. https://doi.org/10.3390/drones8030112

Academic Editor: Anastasios Dimou

Received: 26 January 2024
Revised: 13 March 2024
Accepted: 19 March 2024
Published: 21 March 2024

1. Introduction

Drones equipped with advanced imaging technology have become indispensable tools in aerial photography. Through the application of computer imaging techniques, these platforms empower ground personnel to gain valuable information into surface activities. One of the most prominent applications is drone ground detection, which plays a pivotal role in diverse fields including land monitoring, urban management, and mountain search and rescue. In recent years, the field of drone ground detection has witnessed significant advancements, largely driven by the evolution of artificial intelligence algorithms. Currently, deep learning methods [1,2] are the leading approach in object detection. A typical deep object detection system comprises three components: the backbone network, neck network, and detection head. The backbone network, using architectures like VGG16 [3], ResNet50 [4], or DarkNet53 [2], extracts image features. The neck network, typically employing techniques like Feature Pyramid Network (FPN) [5], enhances these features. The detection head uses these enhanced features for category and location determination, finalizing the detection.

In drone ground detection, single-modality data, such as RGB images [6–8], infrared images [9–11], and other spectral or radar data [12,13], are predominantly used. Multimodal data for target detection has received limited research [14,15]. Multi-modal methods are classified into traditional and deep learning approaches. Traditional methods rely on manually engineered features, like Histogram of Gradient (HOG) features, extracted from visible and infrared images, combined using a support vector machine (SVM) [14].

Hwang et al. [15] extracted three types of features, the channel feature (ACF), thermal (T), and thermal histogram of gradients (THOG), from visible light images for target detection. In target detection, deep learning has gained prominence for its exceptional feature representation capabilities, especially in multi-modal detection. Zhang et al. [16] found that image-level cascading outperforms feature-level cascading. However, merely stacking multi-spectral data does not enable precise feature learning from each spectrum. Fang et al. [17] introduced the Cross-Modal Attention Feature Fusion (CMAFF) module using the attention mechanism, selectively enhancing specific features and choosing shared ones across modalities. Konig et al. [18] achieved feature fusion across modalities with the Region Proposal Network (RPN). Moreover, several studies have taken a unique approach by emphasizing illumination-related information across modalities for joint detection. Li et al. [19] introduced confidence parameters associated with illumination information, which were tackled through the design of a dedicated light sensing network. Subsequently, a gate function predicated on the illumination value was employed to harmoniously fuse features from distinct modalities. Alongside these efforts, there has been a surge in similar research endeavors [20–22]. These include endeavors such as the integration of photophysical information into Convolutional Neural Networks (CNNs) to facilitate the learning of target features, drawing inspiration from neural networks that incorporate physical information.

Recent advancements in multi-modal drone ground detection have led to notable progress. However, current methods in this field often follow a conventional approach. They use established image feature extraction networks and then employ complex fusion strategies to combine features from different modalities (as shown in Figure 1a). While effective in straightforward scenarios, these methods struggle in complex environments and varying lighting conditions. Importantly, they do not consider the interaction between modalities during feature extraction.

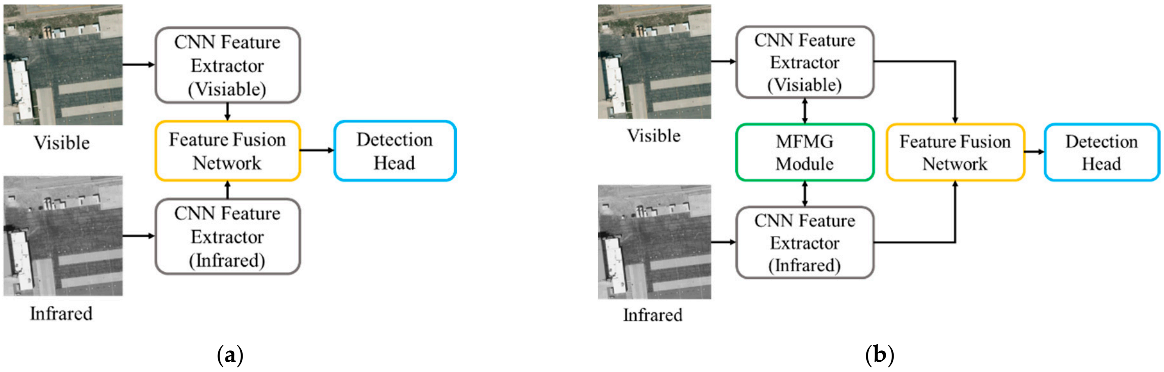

Visible
Infrared
CNN Feature Extractor (Visiable)
Feature Fusion Network
CNN Feature Extractor (Infrared)
Detection Head
(a)
Visible
Infrared
CNN Feature Extractor (Visiable)
MFMG Module
CNN Feature Extractor (Infrared)
Feature Fusion Network
Detection Head
(b)

Figure 1. Comparison of network structure: (**a**) represents the conventional RGB-IR joint target detection methods; (**b**) depicts the RGB-IR joint target detection method designed by our team.

To tackle these challenges, we have designed a network architecture shown in Figure 1b. Unlike traditional multi-modal drone ground detection methods, our approach integrates inter-modal information interaction at the feature extraction stage, followed by deep feature fusion. This is realized through our Multispectral Feature Mutual Guidance Network (MFMGF-Net). To address learning bias among modalities, we have created a mask-based data augmentation method, which employs a constraint network to improve multi-modal feature learning. Initially, we extract multi-modal features with two CSPDarkNet53 networks. We embed a Multispectral Feature Mutual Guidance Network between them, enhancing their feature extraction capabilities. Next, we fuse these features using an attention mechanism in both channel and spatial dimensions. These fused features are then fed into the network, resulting in improved multi-modal drone ground detection performance.

The contributions of this article can be succinctly summarized as follows:

- We present MFMG-Net, a novel architecture for multispectral drone ground detection. We combat potential feature bias across spectral data using a mask-based data enhancement method.
- We develop MFMG to boost feature extraction in spectral backbone networks. It enables cross-spectral information exchange during feature extraction, harnessing the power of complementary spectral data for enhanced feature fusion and detection.
- We propose efficient feature fusion using an attention mechanism. This technique discerns spectral feature correlations in feature channels and space, effectively fusing multispectral features, and improving multispectral drone ground detection.

The rest of this article is organized as follows. In Section 2, we review related work. In Section 3, we describe our proposed method in detail. In Section 4, we present the experimental setup and extensive experimental results. The conclusion of the article is drawn in Section 5.

2. Related Work

In this section, we will conduct a comprehensive review of contemporary object detection algorithms categorized by data type. Our specific emphasis will be on algorithms tailored for visible, infrared, and multispectral data.

2.1. Visible Object Detection

Object detection stands as a focal point in the realm of computer vision, captivating numerous researchers who dedicate their efforts to enhancing the precision, speed, and practicality of detectors. Over the course of the development of target detection algorithms, these detectors have evolved from two-stage detection to single-stage detection, and presently, they are advancing towards end-to-end detection grounded in Transformer architecture. The two-stage detector paradigm segregates the target detection process into two distinct steps. The first stage focuses on localizing the target, while the second stage engages in the classification and fine-tuning of the identified target's position. Prominent two-stage detectors encompass R-CNN [23], Faster R-CNN [1], and R-FCN [24]. Nevertheless, they exhibit noticeable drawbacks, particularly in terms of computational efficiency due to extended network pipelines, necessitating robust computational hardware for optimal performance. In response to these limitations, single-stage detectors emerged. They concurrently handle target positioning and classification, thereby enhancing inference speed. Well-known single-stage detectors include SSD [4], YOLO v4 [25], and others. Among them, the YOLO series of detectors have seen continuous development and refinement, culminating in the YOLO v4, which offers a clear network structure and strikes an excellent balance between accuracy and speed. This paper introduces a joint visible and infrared target detection network based on the YOLO v4 algorithm framework.

Although research on visible light target detection has greatly promoted the development of the field of target detection, in practical applications, visible light has limited its application level due to its sensitivity to illumination.

2.2. Infrared Object Detection

The advancement of visible light target detection algorithms has significantly propelled the progress of infrared target detection algorithms. In contrast to visible light, infrared sensors find widespread use in various specialized environments, such as nighttime and foggy conditions, owing to their insensitivity to lighting conditions. Ghose et al. [26] pioneered the development of an infrared pedestrian target detector founded on Faster R-CNN. However, this model's complexity hampers its computational efficiency. Jhong et al. [22] introduced a lightweight infrared detector based on the single-stage YOLO detector, achieving the detection of both vehicles and pedestrians. Li et al. [27] implemented infrared image data detection using the YOLO v5 network, incorporating numerous techniques from visible light networks, including attention mechanisms and multi-size detection heads. Further details can be found in their research papers. Marnissi et al. [28] devised a domain-adaptive infrared detection algorithm based on Faster R-CNN, incorporating a multi-domain classifier that yielded significant performance enhancements. This domain adaptation method has since been widely adopted in subsequent research to bolster infrared detection capabilities and attain superior infrared detection performance.

Despite their resilience to lighting conditions, infrared sensors possess inherent limitations, including low resolution and the absence of color information. Furthermore, infrared imaging relies on thermal radiation, resulting in poor edge quality for objects and subsequently leading to elevated false detection and missed detection rates.

2.3. Visible–Infrared Fusion Object Detection

As previously mentioned, both visible light sensors and infrared sensors have inherent limitations. Therefore, the effective extraction and fusion of data from these two modalities for improved detection have become a focal point in current research. Deng et al. [29] utilized RGB and IR features to design a feature fusion network aimed at enabling target detection in low-light conditions. Konig et al. [18] achieved feature fusion through the integration of a region proposal network. The above-mentioned studies are centered on feature fusion, often involving pixel-level fusion followed by detection. Pixel-level fusion, which preserves both infrared and available light intensity, has garnered attention, as seen in the work by Chen et al. [30]. Attention mechanisms are also widely employed in joint target detection. For instance, Zhang et al. [31] accomplished valuable feature extraction by devising a modal-level attention module, facilitating the deep fusion of different spectral features. Fang et al. [17] introduced a lightweight fusion method using attention maps to strike a balance between accuracy and computational efficiency. The utilization of Transformer, a pivotal tool in computer vision [32], has extended to the realm of joint target detection. Zhu et al. [33] harnessed Transformer's capacity for learning contextual information to achieve feature integration and, consequently, enhance detection outcomes. However, due to the substantial computational demands of Transformers, these models exhibit high complexity, reduced computational efficiency, and challenging deployment.

3. Proposed Method

3.1. Overview Architecture

This section offers an in-depth elucidation of the architecture and processes of the proposed MFMG-Net. Figure 2 presents an overview of the network's architecture. In this design, we have developed a backbone network based on CSPDarknet53, a component of YOLOv4 [25]. Specifically, we employ two CSPDarknet53 networks for extracting features from RGB and IR images separately. Between these two networks, we have incorporated the MFMG module, facilitating intercommunication of multispectral feature information. This module harnesses the correlations and complementarities present in multispectral data, thereby enhancing the efficacy of feature extraction within the backbone network. Subsequently, the multispectral image features are directed to our feature fusion module, referred to as Fusion, for comprehensive feature fusion. Within this module, we fuse features from two dimensions: the channel dimension and the spatial dimension. Finally, we transmit the resulting three fused features (${F_3}^f$, ${F_4}^f$, and ${F_5}^f$) into the neck network to generate the feature pyramid. This feature pyramid is then subjected to inference through the detection head, ultimately yielding the final detection results. The specifics of each module are delineated in the subsequent sections.

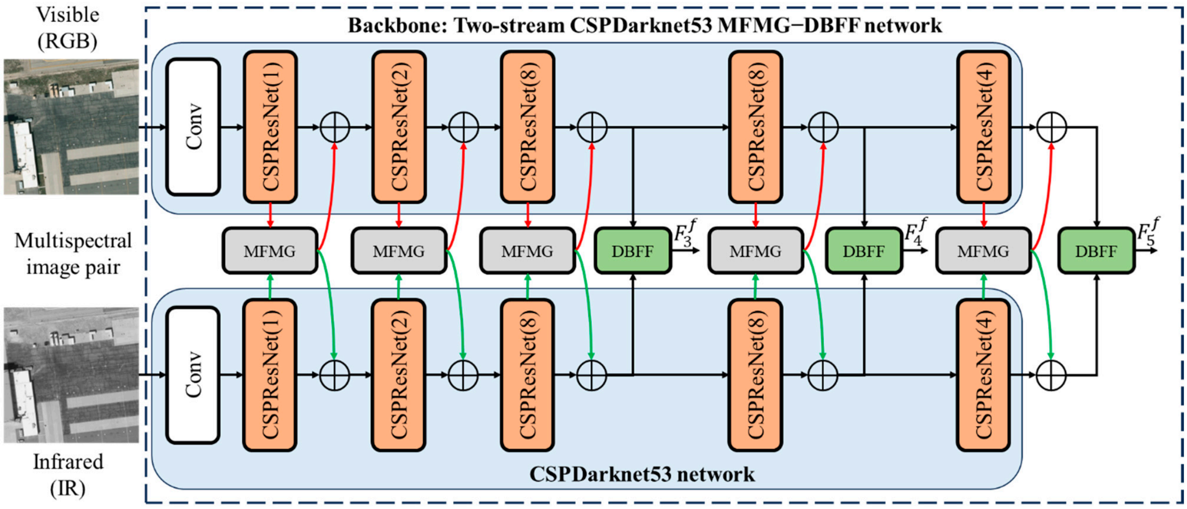
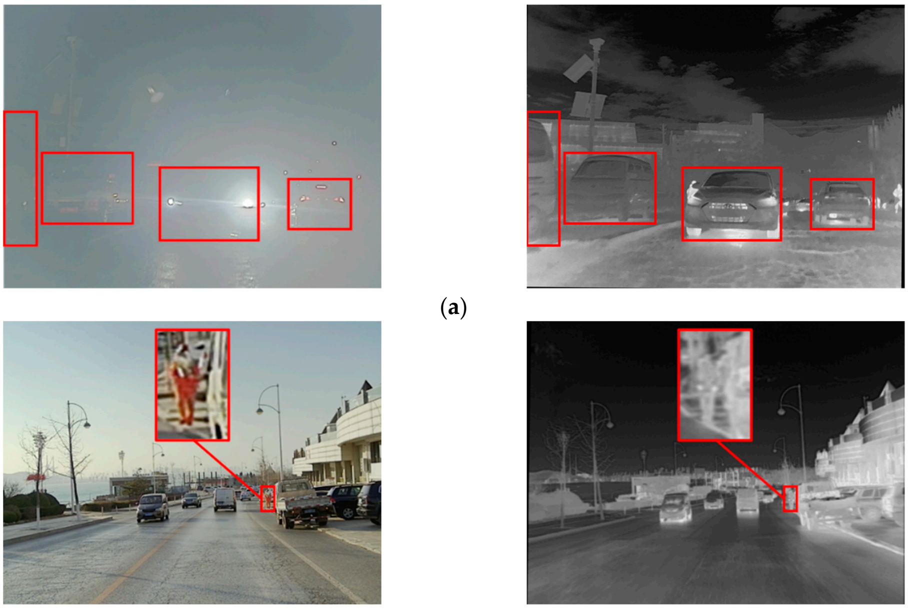


Figure 2. The network framework of the proposed MFMG-Net.

3.2. Data Augmentation

In general, RGB images tend to contain more information than IR images. This assertion is supported using performance comparisons of single-modal detectors, where RGB detection models typically outperform IR detection models. Consequently, when training models with RGB-IR image pairs, there is a risk of biasing the model towards learning RGB features and potentially neglecting IR image features. Conversely, when deploying RGB-IR image pairs for inference, not all modal features prove to be useful, as illustrated in Figure 3. Therefore, in cases where one mode fails to contribute effectively, the network can adapt by focusing on the more informative mode, mitigating potential shortcomings and enhancing overall detection effectiveness. This adaptability inherent in the network allows for the optimal utilization of effective modal features in target detection and can help reduce training-related losses.


(**a**)

(**b**)

Figure 3. Example of RGB-IR image pair. (**a**) In this scenario, the RGB image is affected by the intense light from car headlights, which can obscure details of traffic conditions. (**b**) In this scenario, the RGB image is advantageous due to their rich texture and color information. The red box indicates potential targets.

To prevent the model from developing modality dependencies during training, which could introduce bias in feature learning and inference, we have implemented a dedicated data augmentation strategy. As depicted in Figure 4, this strategy is based on a masking method. Here is how it works: We divide the image into a grid of 3 × 3 squares, and then randomly select one square from each of the two rows to serve as the masking occlusion area for both the RGB and IR images, respectively. Within these selected areas, we set all pixel values to zero. The implementation process is as follows:

$$2\mathrm{Random}_{\mathrm{mask}} = \mathrm{RGB}_{\mathrm{mask}} + \mathrm{Ir}_{\mathrm{mask}} \tag{1}$$

$$\mathrm{RGB}_{\mathrm{mask}} \mid \mathrm{IR}_{\mathrm{mask}} = 0 \tag{2}$$

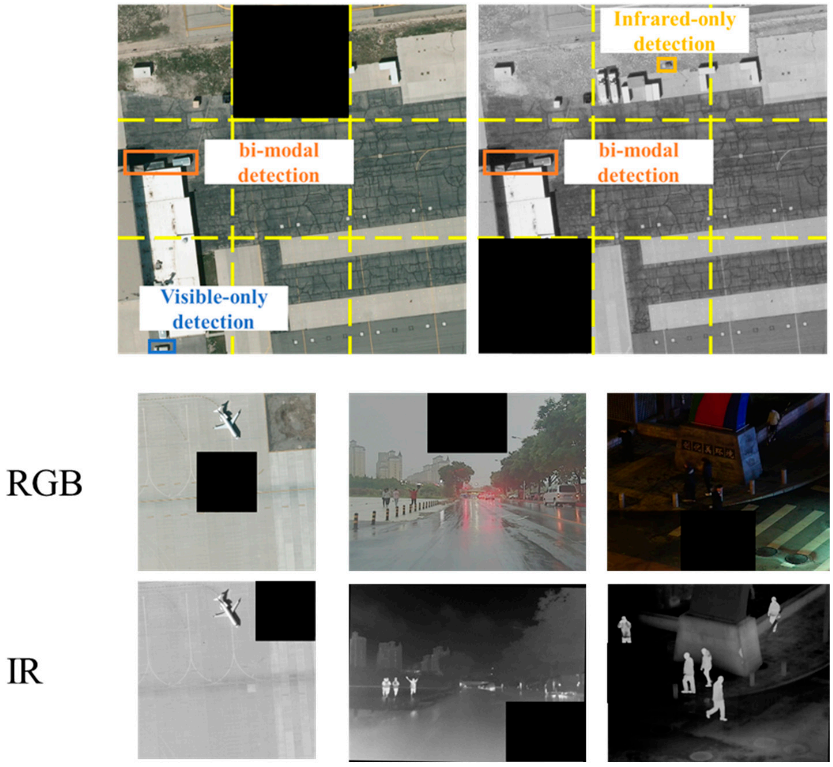


Figure 4. The data augmentation (DA) method proposed in this study. The two black squares represent randomly generated masks. One is used to block the RGB image and the other is used to block the IR image. The purpose of making the model learn only RGB detection, only IR detection and multi-modal detection is to prevent the model from biasing in multi-modal learning.

After applying the data augmentation process, the network is intentionally deprived of useful information within the masked occlusion areas during training. Instead, it is compelled to rely on information from the corresponding regions of the opposite modality, as illustrated in Figure 4. For instance, when detecting occluded regions in RGB images, the model must utilize the infrared features from the corresponding positions for detection. This enforced constraint ensures that the model fully learns and reasons about each modality's features during the training process, mitigating bias. A similar process occurs for occlusions in IR images. Additionally, the manually designed failure features enable the model to concentrate on learning features from other modalities in cases where one modality fails. Importantly, our data augmentation technique, compared to full-image occlusion of one modality in RGB-IR pairs, retains more RGB-IR feature pairs. Thus, our approach encourages the model to utilize fused features for detection effectively. It is worth noting that in our implementation, data augmentation is only introduced after 50 epochs to allow the network to converge quickly. Furthermore, the probability of using data augmentation is set to 30%. Importantly, data augmentation is not employed during the inference phase.

3.3. Multispectral Feature Mutual Guidance Module

In the preceding section, we conducted a comprehensive review of existing research on multi-modal joint target detection. A prevalent observation was that the current methods perform feature extraction independently for each modality. Specifically, they tend to extract features from different modalities separately and then engage in feature fusion, as depicted in Figure 1a. However, it is apparent that different modal information is interrelated and complementary. The primary goal of joint detection is to fully leverage the correlation and complementarity between modal information to enhance detector performance. Motivated by this analysis, we made a strategic decision to incorporate an information exchange mechanism within the feature extraction processes of different modalities. The objective is to guide feature extraction based on the characteristics of each modality, thereby enhancing the quality of feature extraction within the backbone network. We have coined this information exchange mechanism the Multispectral Feature Mutual Guidance (MFMG) module. This module is embedded multiple times throughout the feature extraction process, as illustrated in Figure 3.

As depicted in Figure 5, the MFMG module takes in feature data from the two modalities and concatenates them. This process can be expressed formally as follows:

$$F^c = \begin{bmatrix} F^v \\ F^i \end{bmatrix} \tag{3}$$

where F^v and F^i respectively represent the RGB image features and IR image features extracted by the backbone network, and F^c represents the features after the dual-mode features are cascaded.

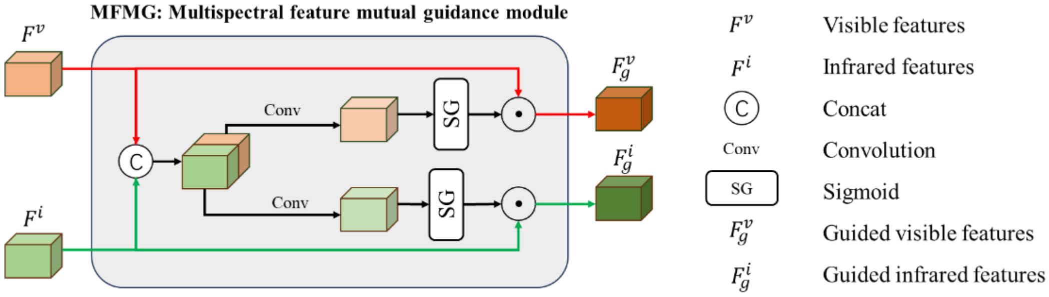


Figure 5. Proposed MFMG module.

Then, two parallel convolutional layers and sigmoid activation functions are connected to the cascaded feature data to obtain the guidance weights of the two modal features, which are formalized as follows:

$$\begin{aligned} W^v &= Sigmoid(F_1(F^c)) \\ W^i &= Sigmoid(F_2(F^c)) \end{aligned} \tag{4}$$

where F_1 and F_2 represent the 1×1 convolution operation, Sigmoid is the activation function, W^v and W^i are the RGB feature guidance weight and IR feature guidance weight, respectively.

Finally, the two calculated guidance weights are multiplied with the feature data received by the module to obtain the guided RGB features and IR features, which are formalized as follows:

$$\begin{aligned} F_g^v &= W^v \odot F^v \\ F_g^i &= W^i \odot F^i \end{aligned} \tag{5}$$

where F_g^v represents the RGB guidance feature after the RGB image features are guided, F_g^i represents the IR guidance feature after the IR image features are guided, and $\odot$ represents the element-wise multiplication operation between matrices.

After obtaining guided features based on multi-modal feature data, we feed back the guided RGB features and IR features to the original feature extraction backbone, respectively. This will prompt the key feature information that the feature extraction network should pay attention to, which is formalized as follows:

$$\begin{aligned} F_r^v &= F^v + F_g^v \\ F_r^i &= F^i + F_g^i \end{aligned} \tag{6}$$

where F_r^v represents the rectified RGB image features, and F_r^i represents the rectified IR image features.

3.4. Dual-Branch Feature Fusion Module

Currently, the attention mechanism finds widespread use in deep neural networks and has garnered significant attention and adoption in the domain of target detection. However, attention-based target detection methods often struggle to strike a proper balance between model complexity and inference efficiency. Typically, in a bid to reduce model complexity, feature data are heavily compressed, leading to substantial loss of image information. Nevertheless, this challenge is effectively addressed by the technique described in [34], building upon the technology presented in [35]. Drawing inspiration from these advancements, we have devised the attention-based Dual-Branch Feature Fusion (DBFF) module. In contrast to its intended purpose of feature enhancement [34], this study employs the DBFF module for deep feature fusion using the attention mechanism.

Notably, the DBFF module facilitates the fusion of multi-modal features both in the spatial and channel dimensions without substantial data compression. Instead, we ensure the network remains sufficiently lightweight through dimensionality reduction operations. Figure 6 illustrates the intricate network structure of the DBFF module, which accepts pairs of two-modal features (F_r^v and F_r^i). These feature pairs traverse through the channel-level fusion branch and the spatial-level fusion branch, resulting in fused features of two distinct dimensions. The fused features from the last two dimensions are concatenated and subsequently subjected to convolution to produce the ultimate fused feature, denoted as FF. In our specific implementation, we carried out feature fusion across three different feature scales, culminating in three fused features of varying scales: ${F_3}^f$, ${F_4}^f$, and ${F_5}^f$.

(1) Channel-level fusion branch. This branch receives the features F_r^v and F_r^i of the two modalities, and then performs the attention branch operations, respectively. One of the attention branches compresses the feature F_r^v in one direction, and the other attention branch maintains high-resolution features in the corresponding orthogonal direction. The operation for F_r^i is the same. The output in both directions (q and v) is formalized as follows:

$$\begin{aligned} W_q^{v|ch} &= \sigma_1(F_1(F_r^v)) \\ W_v^{v|ch} &= \sigma_2(F_2(F_r^v)) \end{aligned} \tag{7}$$

where σ_1 and σ_2 represent the reshape of the tensor, $F_1(\cdot)$ and $F_2(\cdot)$ are 1×1 convolutional layers, and $W_q^{v|ch} \in R^{HW\times 1\times 1}$ and $W_v^{v|ch} \in R^{C/2\times HW}$. The IR branch is the same.

Then, the obtained weights $W_q^{v|ch}$ and $W_q^{i|ch}$ are concatenated and sent to the softmax function to obtain the fused weight distribution. The calculation process is as follows:

$$\begin{bmatrix} W_k^{v|ch} \\ W_k^{i|ch} \end{bmatrix} = Softmax(\begin{bmatrix} W_q^{v|ch} \\ W_q^{i|ch} \end{bmatrix}) \tag{8}$$

We multiply the weights with the fused weight keys and then connect a 1×1 convolution, LayerNorm (LN) and sigmoid function. Among them, LN increases the number of channels to C, and the sigmoid function keeps the result in the range of 0–1, which is formalized as follows:

$$W_z^{v|ch} = Sigmoid(\sigma_3(F_3(W_v^{v|ch} \times W_k^{v|ch}))) \tag{9}$$

where × is the matrix dotproduct operation and $W_z^{v|ch} \in R^{C\times1\times1}$. And $W_z^{i|ch} \in R^{C\times1\times1}$.

Then, we perform channel-level multiplication of F_r^v and $W_z^{v|ch}$ to obtain low-noise feature representation, as follows:

$$W^{v|ch} = F_r^v * W_z^{v|ch} \tag{10}$$

where $W^{v|ch} \in R^{C\times H\times W}$. Similarly, $W^{i|ch} \in R^{C\times H\times W}$.

Finally, the two modal features are cross-added, and then the addition results are cascaded again and convolved to obtain the final channel-level fusion feature $F^{f|ch}$, which is formalized as follows:

$$F^{v|ch} = W^{i|ch} + F_r^v \tag{11}$$

$$F^{i|ch} = W^{v|ch} + F_r^i \tag{12}$$

$$F^{f|ch} = F_4(\begin{bmatrix} F^{v|ch} \\ F^{i|ch} \end{bmatrix}) \tag{13}$$

(2) Spatial level fusion branch. In order to make full use of the complementarity of information between modalities, similar to the channel-level fusion branch, this branch also performs calculations from two directions, and is formalized as follows:

$$W_q^{v|sp} = \sigma_4(F_{GP}(F_5(F_r^v))) \tag{14}$$

$$W_v^{v|sp} = \sigma_5(F_{GP}(F_6(F_r^v))) \tag{15}$$

where F_5 and F_6 are 1 × 1 convolutional layers, $F_{GP}(\cdot)$ is a global pooling operator, σ_4 and σ_5 represent the reshape of the tensor. $W_q^{v|sp} \in R^{1\times2/C}$ and $W_v^{v|sp} \in R^{C/2\times HW}$. The IR branch is the same.

Then, similarly to the channel-level fusion branch, the concatenated features are fed into the softmax function. Therefore, we obtain $W_k^{v|sp}$ and $W_k^{i|sp}$. Then, we perform multiplication, reshape, and sigmoid in sequence, formalized as follows:

$$W_z^{v|sp} = Sigmoid(\sigma_6(W_v^{v|sp} \times W_k^{v|sp})) \tag{16}$$

where $W_z^{v|sp} \in R^{1\times HW}$. Similarly, $W_z^{i|sp} \in R^{1\times HW}$.

The spatially fused feature map can be obtained by multiplying F_r^v and $W_z^{v|sp}$:

$$W^{v|sp} = F_r^v * W_z^{v|sp} \tag{17}$$

where $W^{v|sp} \in R^{C\times H\times W}$. Similarly, $W^{i|sp} \in R^{C\times H\times W}$

We cross-add the features of the two modalities again, and then cascade the addition results again and perform a convolution operation to obtain the final spatial-level fusion feature $F^{f|sp}$, which is formalized as follows:

$$F^{v|sp} = W^{i|sp} + F_r^v \tag{18}$$

$$F^{i|sp} = W^{v|sp} + F_r^i \tag{19}$$

$$F^{f|sp} = F_7(\begin{bmatrix} F^{v|sp} \\ F^{i|sp} \end{bmatrix}) \tag{20}$$

Finally, we cascade the channel-level fusion feature $F^{f|ch}$ obtained in (1) and the spatial-level fusion feature $F^{f|sp}$ obtained in (2) and perform a convolution operation to obtain the deep fusion feature F^f, which is formalized as follows:

$$F^f = F_8 \begin{bmatrix} F^{f|ch} \\ F^{f|sp} \end{bmatrix} \tag{21}$$

After the above operations, we obtained the deep fusion feature F^f. More specifically, we obtain three fused features of different scales, denoted as $F_3{}^f$, $F_4{}^f$, and $F_5{}^f$, and then we send them into the neck network and detection head for detection. In the implementation, we directly used the neck network and detection head of YOLO v4. For details, please refer to the literature [25].

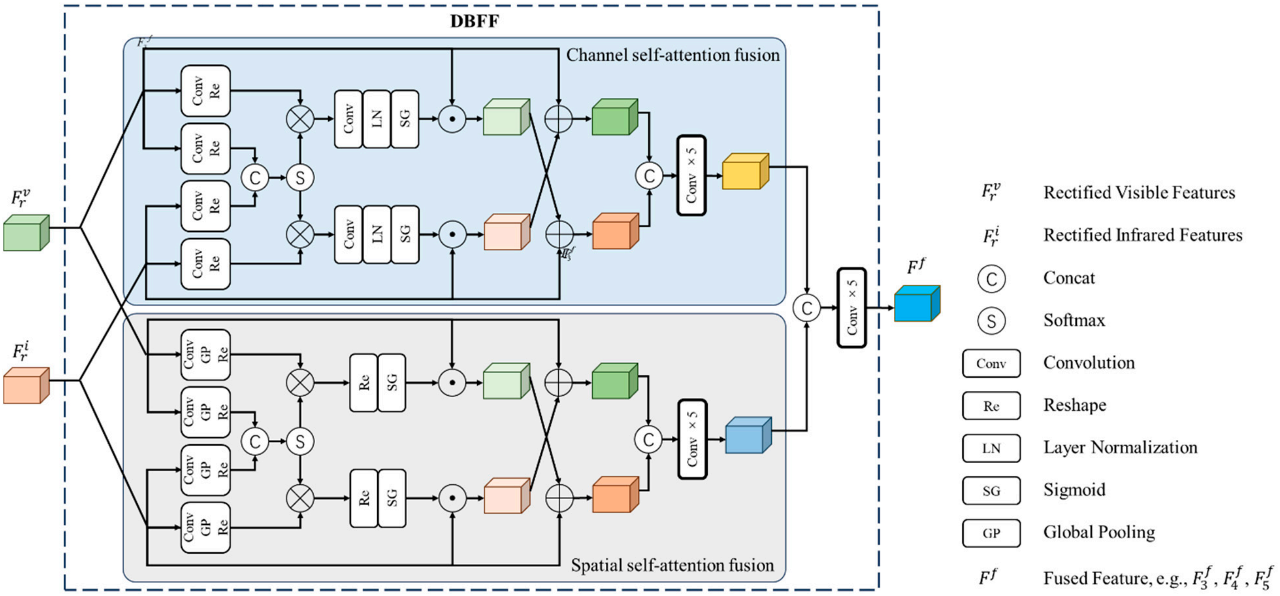

DBFF
Channel self-attention fusion
Spatial self-attention fusion
F_r^v
F_r^i
F^f
Conv Re
Conv GP Re
Conv LN SG
Re SG
Conv × 5
C
S
F_r^v Rectified Visible Features
F_r^i Rectified Infrared Features
C Concat
S Softmax
Conv Convolution
Re Reshape
LN Layer Normalization
SG Sigmoid
GP Global Pooling
F^f Fused Feature, e.g., F_3^f, F_4^f, F_5^f

Figure 6. Dual-branch feature fusion module.

4. Experiments

This section encompasses five comprehensive summaries that elucidate the experimental design and implementation. Initially, we introduce the datasets employed in our experiments. Subsequently, we delve into crucial aspects of algorithm implementation. Following this, we undertake ablation experiments and benchmark our performance against state-of-the-art related methods. Finally, we execute real-world tests of the algorithm using laboratory shooting equipment to substantiate the efficacy of our approach.

4.1. Datasets

To evaluate the proposed multispectral combined target detection methods, we utilized four public datasets: VEDAI [36], M3FD [37], LLVIP [38], and FLIR [39], for conducting our experiments. These datasets consist of RGB-IR image pairs, with specific details outlined in Table 1.

Table 1. Dataset overview.

Item	**VEDAI**	**M3FD**	**LLVIP**	**FLIR**
Classes	9	6	1	3
Data	RGB-IR	RGB-IR	RGB-IR	RGB-IR
Size	1024 × 1024	1024 × 768	1280 × 1024	640 × 512
Format	png	png	jpg	jpg
Amount	1250 pairs	4200 pairs	15,488 pairs	5142 pairs

(1) VEDAI: The VEDAI dataset comprises aerial imagery data with pixel-level annotations. It categorizes targets into nine classes, presenting challenges like small target scales, single perspectives, and varying lighting conditions. This dataset primarily focuses on the detection and study of vehicle targets, encompassing various vehicle types such as cars, RVs, and pickup trucks. It provides two image resolutions, 515 × 512 and 1024 × 1024. In our experiments, we selected the higher resolution version, 1024 × 1024.

(2) M3FD: The M3FD dataset covers a range of challenging scenarios, including daytime, evening, and nighttime conditions, as well as scenarios with smoke obscuration. This diverse dataset offers six target categories and comprises 4200 RGB-IR image pairs with an image resolution of 1024 × 768. The dataset is variable in lighting conditions, which provides an excellent platform for testing algorithm performance.

(3) LLVIP: The LLVIP dataset is a comprehensive RGB-IR dataset specifically designed for visible light and infrared joint inspection research under low-light conditions, with most data collected at night. This dataset is characterized by a high resolution of 1024 × 1024 and focuses on a single detection category, namely pedestrians.

(4) FLIR: The FLIR dataset captures data from traffic road scenarios and includes three object categories for detection: people, bicycles, and cars. It encompasses both day and night scenes and is frequently employed for testing multi-modal combined target detection algorithms. While this dataset is publicly accessible, official RGB-IR images were not provided. To ensure consistent comparison data, we used the data provided in [40], using 4129 images for training and reserving the remainder for testing.

4.2. Experiment Details

We conducted a comprehensive array of experiments, including comparative studies and external tests, to thoroughly evaluate the performance of our algorithms. Notably, we designed a baseline network rooted in the YOLOv4 algorithm framework. Specifically, we augmented YOLOv4 with an infrared branch dedicated to the extraction of infrared image features. These features were then combined with visible light image features using a simple fusion approach, as illustrated in Figure 7. To assess the individual contributions of each component within our method, we carried out ablation experiments using the LLVIP dataset. For performance comparisons with state-of-the-art methods, we conducted tests on the VEDAI, M3FD, and LLVIP datasets. To ensure equitable comparisons, all detectors employed the same data splits. Additionally, we evaluated the inference speed of our model through tests conducted on the FLIR dataset. All training and testing processes were executed on an NVIDIA RTX 3090 desktop. The training process spanned 300 epochs, with an initial learning rate of 0.001. At the fiftieth epoch, the learning rate was reduced by a factor of 0.1. The training batch size was set to 4, and we employed the ADAM optimizer.

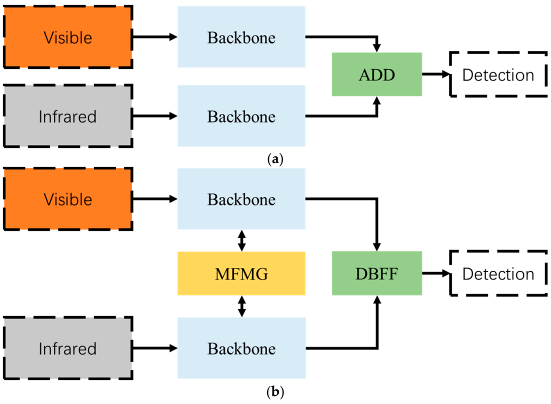


Figure 7. The baseline detector and the detector network implementation plan. (**a**) Base-line detector, the feature extraction process is independent, and the characteristics of the features are added with elements. (**b**) The proposed MFMG-Net, the feature extraction under mutual guidance and the characteristic fusion of the characteristic fusion module of the two-point branches.

4.3. Experiment Results

4.3.1. Ablation Experiment

This section is dedicated to ablation experiments aimed at assessing the impact of the various proposed modules on the network and evaluating the overall performance of our network solution. These experiments are conducted using the LLVIP dataset. We initiated the testing by evaluating the effect of the proposed Data Augmentation (DA) method on detection performance, thereby confirming the effectiveness of this augmentation technique. Subsequently, we delved into the contribution of the Multispectral Feature Mutual Guidance (MFMG) module and the Dual-Branch Feature Fusion (DBFF) module to network performance. To carry this out, we conducted experiments using real-world images, both with and without DA. Moreover, we assessed the individual contributions of MFMG and DBFF by enabling or disabling these modules. In the following sections, we provide a detailed analysis of the results obtained from these experiments, shedding light on the significance of each module and its impact on the overall performance of our multispectral target detection network.

(1) Effectiveness of DA: To assess the impact of the proposed Data Augmentation (DA) method on detector training, we conducted a comparative analysis by training detectors both with and without DA. This assessment was performed on both the baseline detector and our proposed detector. Initially, we evaluated the influence of DA on the training performance of the baseline detector. We introduced a probability parameter for applying DA during algorithm implementation. This probability parameter was adjusted during training to produce detection models with varying DA probabilities, and we observed their respective performance differences. The results are presented in Table 2. As shown in the table, we applied DA with probability parameters of 0.1, 0.3, 0.5, 0.7, and 0.9 to the baseline detector. Subsequent testing on the LLVIP dataset revealed that the proposed DA method improved the mean Average Precision (mAP) evaluation metric for the baseline detector by 0.2%, 0.9%, 0.8%, 0.6%, and 0.3%, respectively. Furthermore, other evaluation metrics for the baseline detector also demonstrated improvements when influenced by the DA method. This comprehensive improvement underscores the effectiveness of the proposed DA method. The results indicate that the baseline detector performed optimally when DA was applied with a probability of 0.3. The underlying principle of the DA method is to introduce noise while retaining the original data. When the noise level becomes excessive and overshadows the original data, it can lead to a degradation in training results. Consequently, we selected a DA probability of 0.3 for testing the effects of the MFMG and DBFF modules.

Table 2. Comparison of AP among the proposed method, the base-line detector, and ablation experiment on the LLVIP dataset. √ indicates that data augmentation has been performed.

Method	DA	MFMG	DBFF	PR	RE	mAP$_{50}$	mAP
Baseline	-	-	-	0.958	0.883	0.921	0.615
Baseline	√ (0.1)	-	-	0.960	0.887	0.927	0.617
Baseline	√ (0.3)	-	-	0.970	0.908	0.948	0.624
Baseline	√ (0.5)	-	-	0.967	0.901	0.942	0.623
Baseline	√ (0.7)	-	-	0.963	0.897	0.937	0.621
Baseline	√ (0.9)	-	-	0.961	0.891	0.933	0.618
Baseline	√ (0.3)	√	-	0.982	0.913	0.960	0.652
Baseline	√ (0.3)	-	√	0.981	0.912	0.957	0.647
MFMGF-Net	-	√	√	0.983	0.925	0.962	0.659
MFMGF-Net	√ (0.3)	√	√	**0.985**	**0.941**	**0.981**	**0.665**

(2) Effectiveness of the MFMG module: In Figure 7, we integrated an MFMG module between the two backbone networks to facilitate the exchange of information during the feature extraction process. This module's bidirectional design ensures seamless information interchange without affecting the output dimensions of the two backbone networks. The

primary function of the MFMG module is to enhance the feature extraction capabilities of both backbone networks. It achieves this by facilitating the exchange of information between modalities, thereby promoting superior feature learning through the exploitation of associated complementary information. To validate the efficacy of this module, we conducted an ablation experiment. Specifically, we set the DA probability to 0.3 and evaluated the impact of the presence or absence of the MFMG module on the baseline detector's performance. As shown in Table 2, the MFMG module significantly improved the baseline detector's mAP performance by 2.8%. This result provides empirical evidence of the MFMG module's effectiveness.

(3) Effectiveness of the DBFF module: In contrast to the MFMG module, the DBFF module is designed to facilitate the fusion of information from two modalities. This module combines features from both modalities in the channel and spatial dimensions using the attention mechanism. Subsequently, it performs a deep fusion of the two fused features through data-driven learning. The fusion in the channel dimension primarily focuses on merging different semantic layers, which is advantageous for classification tasks. On the other hand, spatial dimension fusion emphasizes combining location information and is particularly relevant for localization tasks. Similar to the MFMG ablation experiment, we set the DA probability to 0.3 and tested the real effectiveness of the DBFF module by enabling or disabling it on the baseline detector. As shown in Table 2, this module improved the mAP performance of the baseline detector by 2.3%, unequivocally confirming its effectiveness.

In summary, we conducted ablation experiments on the three designed components, and the test results demonstrated the independent effectiveness of the designed components. Additionally, the test results of the joint use of these components also highlighted their compatibility. As indicated in Table 2, using MFMG and DBFF together with the baseline detector improved the mAP performance by 4.4%. When all three components were combined, the performance of the baseline detector improved by 5%.

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

In order to further validate the performance of the proposed method, this section will compare it with current state-of-the-art methods on four datasets, demonstrating the advancement of our proposed method from both qualitative and quantitative perspectives. Specifically, we will compare test accuracy on three public datasets: VEDAI, M3FD, and LLVIP, and finally, we will compare algorithm inference efficiency on the FLIR dataset.

(1) Comparative experiments on the VEDAI dataset: In this comparison, we selected advanced single-modal detection methods and multi-modal detection methods to assess the performance of the proposed method. The single-modal methods considered included YOLOv5, YOLOv8, EfficientDet [41], SSSDET [42], among others, while the multi-modal methods encompassed LRAF-Net, YOLO Fusion, CFT, and similar approaches. The detection test results of these advanced methods, along with those of the proposed method on the VEDAI dataset, are summarized in Table 3. The findings clearly demonstrate the superiority of the proposed method. Specifically, in comparison to the best single-modal method, the proposed method improved the mAP evaluation metric by an impressive 13.2%. When compared with the best-performing multi-modal method, the proposed method enhanced the mAP evaluation index by 0.3%.

The qualitative detection results are depicted in Figure 8. In comparison to the baseline detectors, our proposed method significantly mitigates both missed detections and false detections. In Figure 8, missed detections are indicated by red arrows, while false detections are denoted by green arrows. As evident in Figure 8a,d, the baseline detector employs a simple addition for feature fusion, leading to a higher occurrence of missed detections and false detections. Conversely, all methods, including the proposed one, have been augmented in their feature extraction capabilities and depth of feature fusion, resulting in a substantial improvement in detection and classification accuracy. As seen in Figure 8b,c, although the baseline detector can locate the target, it faces challenges in accurately classifying it. In contrast, the proposed detection method capitalizes on its robust feature learning capability and feature channel-level fusion, enhancing its ability to discern small objects. Since this dataset encompasses complex backgrounds and poses difficulties in detecting small targets, the qualitative comparison reaffirms the remarkable detection capabilities of the proposed method.

Table 3. Comparison of AP among the proposed method, the state-of-the-art methods, and comparison experiment on the VEDAI dataset.

Model	Dataset Type	Backbone	mAP_{50}	mAP
Retina [43]	RGB	ResNet-50	-	0.435
Faster R-CNN	RGB	ResNet-101	-	0.348
SSSDET	RGB	shallow network	-	0.460
EfficientDet(D1)	RGB	EfficientNet(B1)	0.740	-
EfficientDet(D1)	IR	EfficientNet(B1)	0.712	-
YOLO-fine	RGB	Darknet53	0.760	-
YOLO-fine	IR	Darknet53	0.752	-
YOLO v5	RGB	CSPDarknet53	0.743	0.462
YOLO v5	IR	CSPDarknet53	0.740	0.462
YOLOv3 e fusion [44]	RGB + IR	Darknet53	-	0.440
YOLOv3 m fusion [44]	RGB + IR	two-stream Darknet53	-	0.446
YOLO Fusion	RGB + IR	two-stream CSPDarknet53	0.786	0.491
CFT	RGB + IR	CFB	0.853	0.560
LRAF-Net	RGB + IR	two-stream CSPDarknet53	0.859	0.591
Baseline	RGB + IR	two-stream CSPDarknet53	0.792	0.453
MFMGF-Net	RGB + IR	two-stream CSPDarknet53	**0.868**	**0.594**

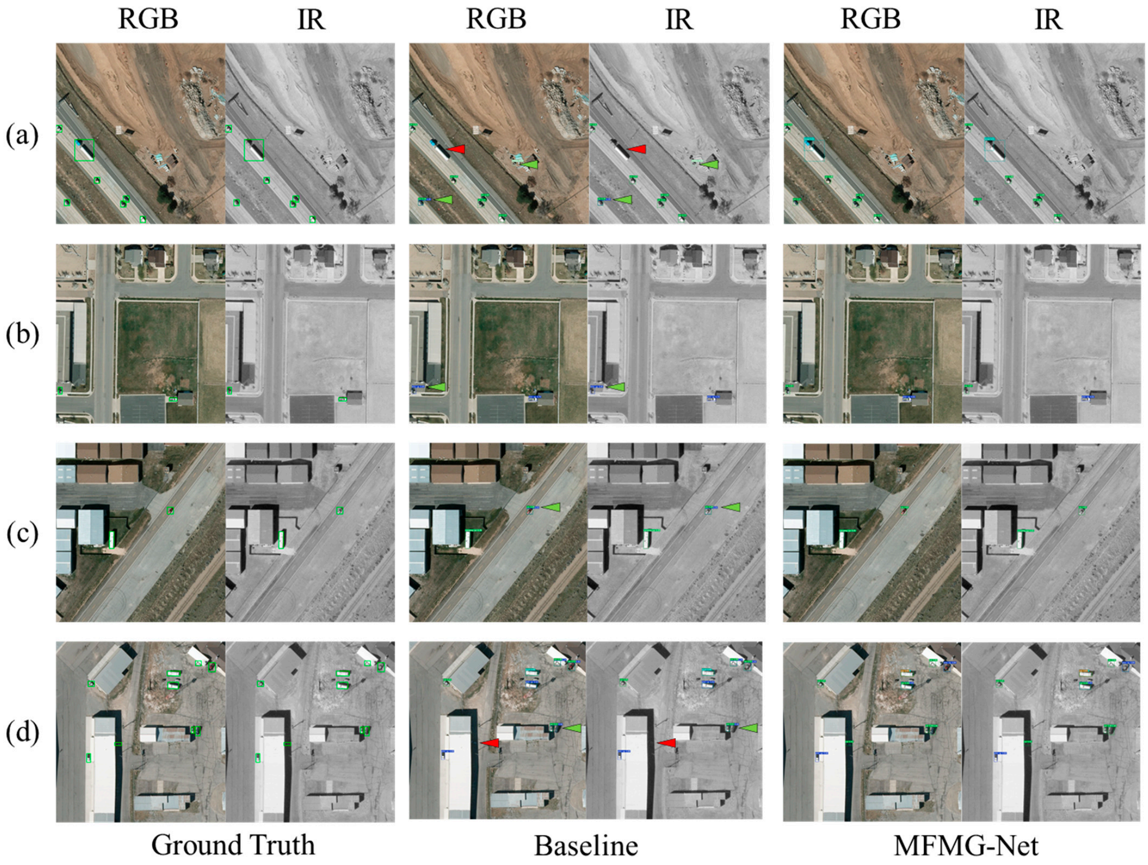


Figure 8. Detection results for four representative scenarios in the VEDAI dataset. Note that red inverted triangles indicate FNs, and green inverted triangles show FPs. Zoomed in to see details.

(2) Comparative experiments on the M3FD dataset: In a manner akin to the comparison experiment conducted on the VEDAI dataset, we selected advanced single-modal target detection methods and multi-modal target detection methods for comparison with the proposed method on the M3FD dataset. The quantitative comparison results are presented in Table 4. As observed in the table, the proposed method enhances the mAP evaluation metric by 2.7% compared to the best single-modal method and improves it by 1.7% compared to the best multi-modal method.

Table 4. Comparison of AP among the proposed method, the state-of-the-art methods, and comparison experiment on the M3FD dataset.

Model	Dataset Type	Backbone	mAP$_{50}$	mAP
Faster R-CNN	RGB	ResNet-50	0.871	0.562
Faster R-CNN	IR	ResNet-101	0.803	0.558
YOLOv7 [45]	RGB	ELAN-Net	0.916	0.631
YOLOv7 [45]	IR	ELAN-Net	0.891	0.573
YOLO Fusion	RGB + IR	two-stream CSPDarknet53	0.928	0.641
GAFF	RGB + IR	ResNet18	0.891	0.576
CFT	RGB + IR	CFB	0.765	0.492
Baseline	RGB + IR	two-stream CSPDarknet53	0.927	0.635
MFMGF-Net	RGB + IR	two-stream CSPDarknet53	**0.930**	**0.658**

In addition, the qualitative comparison results are shown in Figure 9. As shown in Figure 9a,b, in low-light challenging environments, the baseline detector cannot detect small or blurred targets. The proposed method successfully detects them with its powerful representation learning and feature fusion capabilities. Furthermore, as shown in Figure 9c,d, both the baseline detector and the proposed method can effectively detect objects; however, the classification performance of the baseline method still lags slightly behind the proposed method, that is, misidentifications occur.

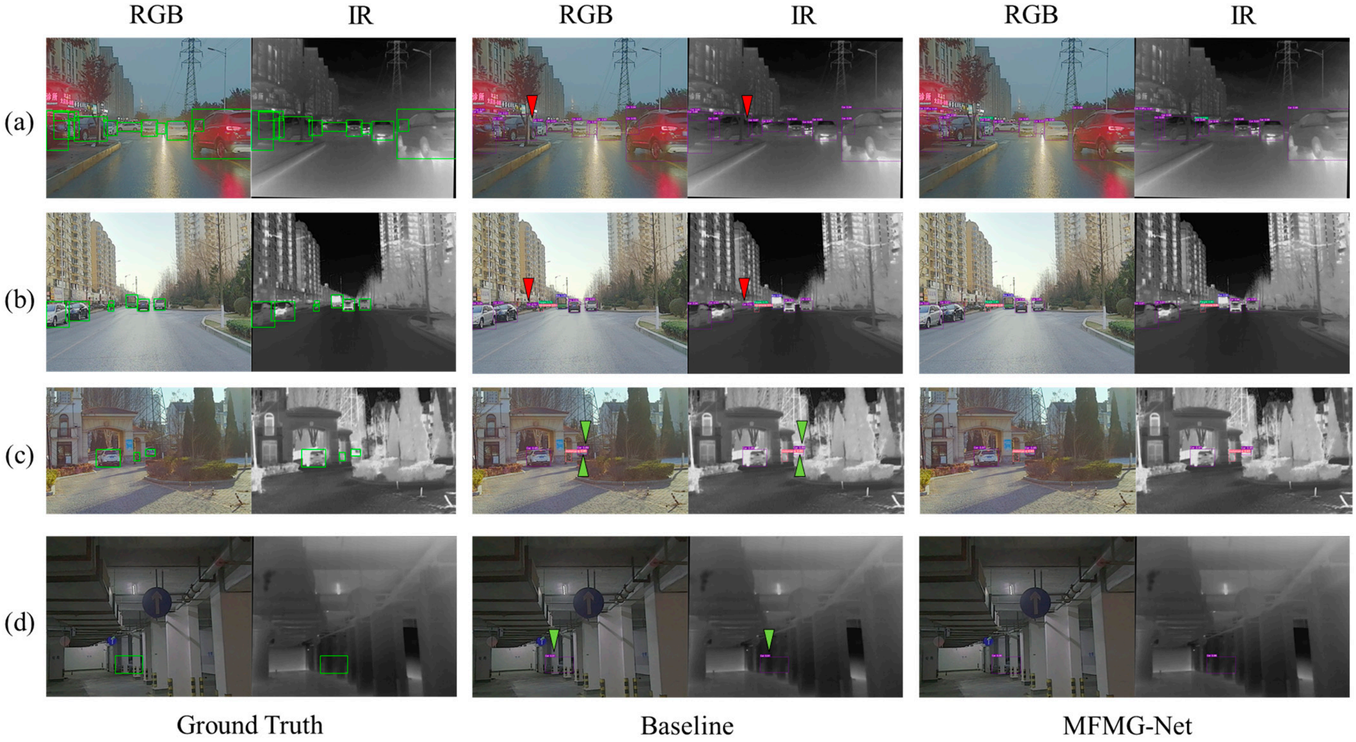


Figure 9. Detection results for four representative scenarios in the M3FD dataset. Note that red inverted triangles indicate FNs, and green inverted triangles show FPs. Zoomed in to see details.

(3) Comparative experiments on the LLVIP dataset: The quantitative comparison results between the proposed method and advanced target detection methods on the LLVIP dataset are presented in Table 5. As evident from the table, the proposed method once again excelled in target detection performance. Specifically, when compared to the best single-modal method, the proposed method enhances the mAP evaluation metric by 4.6%, and when compared to the best multi-modal method, it improves the mAP evaluation metric by 0.2%. Additionally, the qualitative comparison results are depicted in Figure 10. As observed in Figure 10, the detection scenes in LLVIP are exclusively nocturnal street scenes where visible light information is limited. Therefore, treating visible light and infrared information equally would unlikely yield better detection results. As shown in Figure 10a–c, the baseline detector struggles to weight the information from the two modalities effectively, due to independent feature extraction and simple feature addition, resulting in numerous false positives. In Figure 10d, the baseline detector struggles to accurately identify targets with overlap and occlusion, while the proposed method achieves precise detection by effectively fusing the complementary information from the two modalities.

Table 5. Comparison of AP among the proposed method, the state-of-the-art methods, and comparison experiment on the LLVIP dataset.

Model	Dataset Type	Backbone	mAP_{50}	mAP
YOLO v3	RGB	DarkNet53	0.859	0.433
YOLO v3	IR	DarkNet53	0.897	0.534
YOLO v5	RGB	CSPDarkNet53	0.908	0.500
YOLO v5	IR	CSPDarkNet53	0.946	0.619
YOLO v8	RGB	-	0.925	0.541
YOLO v8	IR	-	0.966	0.632
CFT	RGB + IR	CFB	0.975	0.636
LRAF-Net	RGB + IR	two-stream CSPDarkNet53	0.979	0.663
Baseline	RGB + IR	two-stream CSPDarkNet53	0.943	0.638
MFMGF-Net	RGB + IR	two-stream CSPDarkNet53	**0.981**	**0.665**

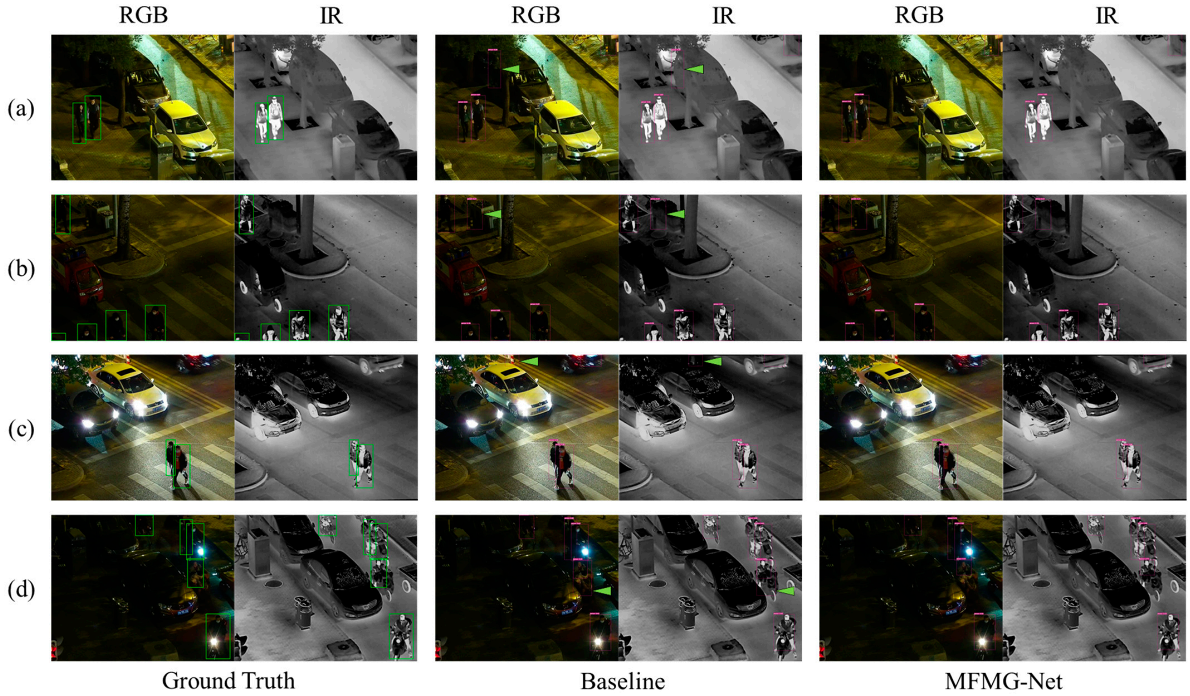


Figure 10. Detection results for four representative scenarios in the LLVIP dataset. Note that green inverted triangles show FPs. Zoomed in to see details.

In summary, the proposed method does not extract features independently and treat the two modalities equally, as the baseline detector does. Instead, it interactively extracts features and deeply fuses the features from both modalities. Consequently, the proposed method is exceptionally well-suited to various challenging detection scenarios.

(4) Computational efficiency analysis: To compare the computational efficiency of the proposed method with other advanced object detection methods, we conducted inference efficiency comparison experiments. Specifically, we used the same computing platform (1080Ti) and dataset (FLIR) as in previous research for a fair comparison. The results, including network parameters, floating-point operations (FLOPs), and inference time for the proposed method and related comparison methods, are presented in Table 6. As observed in the table, the proposed method achieves real-time inference speed, with an inference time of 23.4 ms. While the computational efficiency of this method is slightly lower than that of several comparison methods, it maintains higher accuracy.

Table 6. Comparison of parameters, FLOPs, and runtime on the FLIR dataset.

Model	Data Type	Param.	FLOPs	Runtime/ms
YOLOv5s	RGB	7.1 M	15.9	10.7
YOLOv5s	IR	7.1 M	15.9	10.7
GAFF R	RGB + IR	23.8 M	-	10.9
GAFF V	RGB + IR	31.4 M	-	9.3
CFT	RGB + IR	73.7 M	154.7	91.2
LRAF-Net	RGB + IR	18.8	40.5	21.4
Baseline	RGB + IR	11.6 M	26.4	17.2
MFMGF-Net	RGB + IR	21.8 M	45.5	23.4

In summary, our proposed method has demonstrated superior detection performance on three public datasets: VEDAI, M3FD, and LLVIP, while also showcasing its adaptability to different environmental conditions. Furthermore, inference experiments on the FLIR dataset have confirmed its real-time inference capabilities.

4.4. Algorithm Testing in Real Scenarios

The previously mentioned experiments were all conducted on public datasets. To validate the effectiveness of the proposed algorithm using custom hardware, we conducted field experiments using in-house camera hardware in the laboratory.

(1) Experimental setup: The camera equipment used is homemade laboratory equipment, mounted on a tripod. This custom device is equipped with both visible light and infrared sensors capable of capturing visible light images and infrared pictures. The camera device is connected to the computer via a USB cable. Prior to detection, the dual-modal images have been registered to ensure that they have the same resolution and are pixel-aligned, as depicted in Figure 11. The image resolution after registration is 1024 × 768. The shooting locations were chosen from various spots and times on the campus, encompassing teaching buildings, roads, and squares during daytime, dusk, and nighttime, as presented in Figure 12. After framing, we utilize the model on the server to detect the framed image.

(2) Experimental results: Some qualitative test results from this field location experiment are displayed in Figure 13. The figures demonstrate that the detection method we have designed can achieve excellent detection results under various lighting conditions and in various scenes, showcasing the remarkable detection robustness of the proposed algorithm. Furthermore, it is worth noting that there are significant disparities in background and spatial resolution between the experimental scene and the public dataset. Additionally, imaging is compromised due to high sensor noise in non-commercial devices. However, the detector's final performance remains unaffected. This underscores the enhancement of target features through the attention mechanism, and despite poor imaging quality from non-commercial equipment, our proposed algorithm can still effectively

detect targets. This indirectly illustrates the strong hardware adaptability of the algorithm presented in this paper.

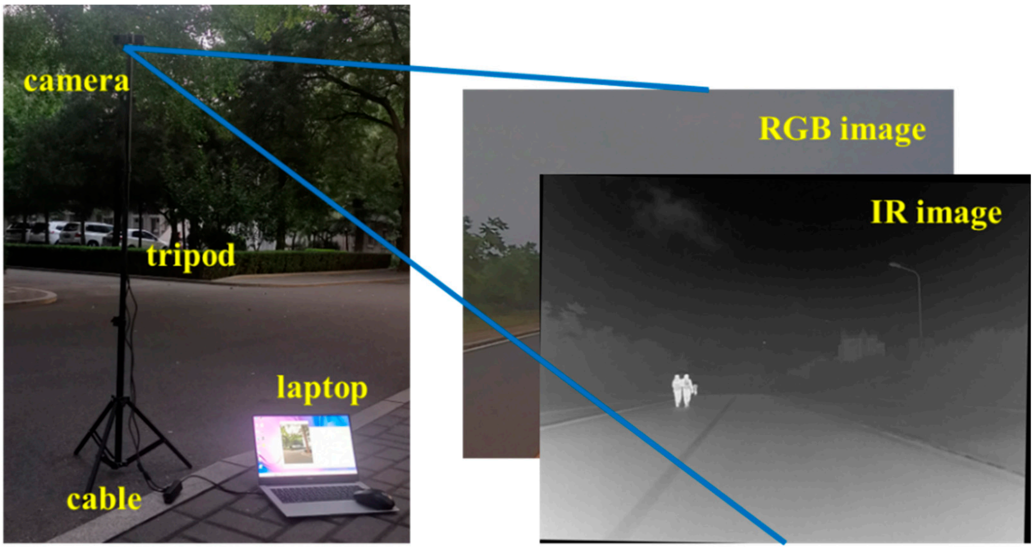


Figure 11. Experimental equipment and scenarios, as well as sample data.

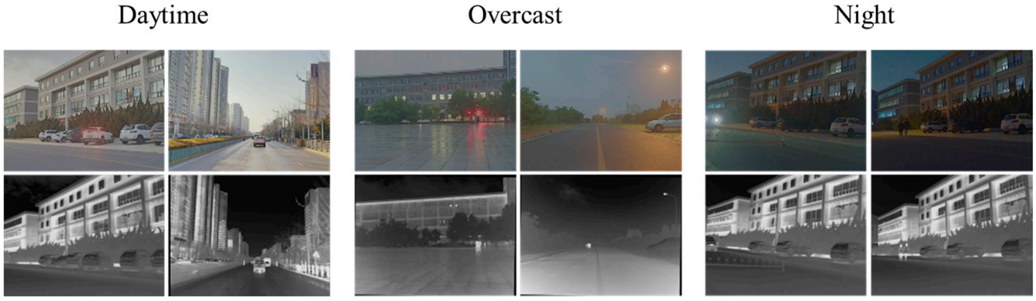


Figure 12. Some typical RGB-IR data pairs taken at different time periods and different sampling places.

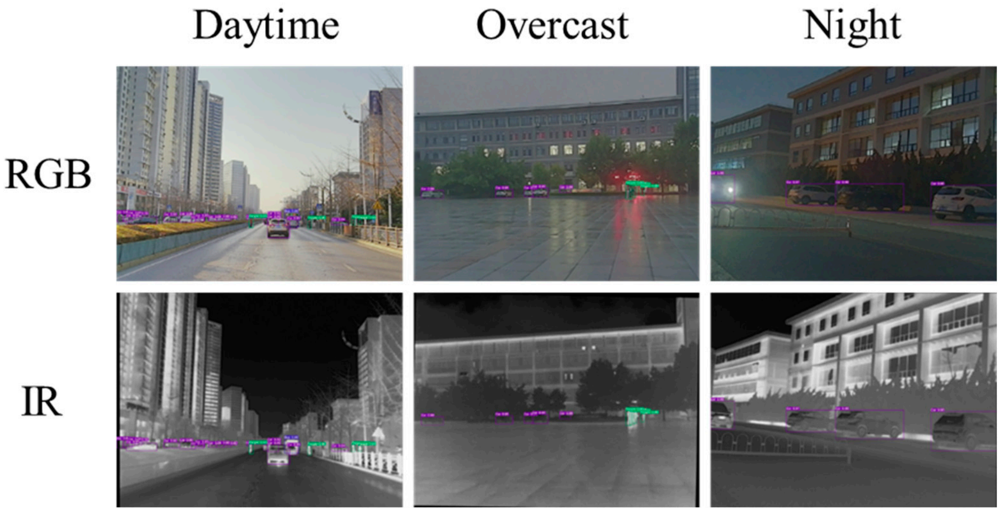


Figure 13. Detection results of typical RGB-IR data pairs captured in the field.

5. Conclusions

In this research, we introduced the MFMG-Net, an algorithm for joint target detection using visible and infrared data. To mitigate learning bias between modalities during training, we developed an effective data augmentation method based on the mask technique. To enhance feature extraction from different modalities, we introduced the MFMG module, facilitating information exchange during the feature extraction process. Additionally, we designed the DBFF module for deep feature fusion, considering feature channel and spatial dimensions, making the algorithm adaptable to complex scenarios. Our experiments, conducted on four public datasets and real-world data, consistently demonstrate the algorithm's high performance in terms of detection accuracy and computational efficiency. Although this study has made some progress in dual-modal fusion target detection, the limitations of the model during deployment were not fully considered. That is to say, due to the limited computing power of UAV airborne equipment, we need to focus on model deployment methods in subsequent research.

At the same time, it should be pointed out that due to the public availability of dual-modal datasets, except for the VEDAI data set, the other data are not aerial image data, which is also a regret of this study. Therefore, we hope that more dual-modal aerial image data will be open sourced in the future.

Author Contributions: Conception and design of study, acquisition of data, analysis and interpretation of data, writing—original draft preparation, software, F.Z.; conception and design of study, writing—original draft preparation, W.L.; conception and design of study, visualization, investigation, software, H.F.; conception and design of study, Visualization, software, N.D. and C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (No. 62304022), Young Elite Scientists Sponsorship Program by CAST (No. YESS20220385), Science and Technology on Electromechanical Dynamic Control Laboratory (China, No. 6142601012304).

Data Availability Statement: All the datasets presented in this study can be found through the referenced papers.

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

References

1. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
2. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
3. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
4. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
5. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 2117–2125.
6. Qin, C.; Wang, X.; Li, G.; He, Y. An Improved Attention-Guided Network for Arbitrary-Oriented Ship Detection in Optical Remote Sensing Images. *IEEE Geosci. Remote. Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
7. Pham, M.-T.; Courtrai, L.; Friguet, C.; Lefèvre, S.; Baussard, A. YOLO-Fine: One-Stage Detector of Small Objects Under Various Backgrounds in Remote Sensing Images. *Remote. Sens.* **2020**, *12*, 2501. [CrossRef]
8. Dong, X.; Qin, Y.; Fu, R.; Gao, Y.; Liu, S.; Ye, Y.; Li, B. Multiscale Deformable Attention and Multilevel Features Aggregation for Remote Sensing Object Detection. *IEEE Geosci. Remote. Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
9. Du, J.; Lu, H.; Zhang, L.; Hu, M.; Chen, S.; Deng, Y.; Shen, X.; Zhang, Y. A Spatial-Temporal Feature-Based Detection Framework for Infrared Dim Small Target. *IEEE Trans. Geosci. Remote. Sens.* **2022**, *60*, 3000412. [CrossRef]
10. Wang, H.; Liu, C.; Ma, C.; Ma, S. A Novel and High-Speed Local Contrast Method for Infrared Small-Target Detection. *IEEE Geosci. Remote. Sens. Lett.* **2020**, *17*, 1812–1816. [CrossRef]
11. Yi, H.; Yang, C.; Qie, R.; Liao, J.; Wu, F.; Pu, T.; Peng, Z. Spatial-Temporal Tensor Ring Norm Regularization for Infrared Small Target Detection. *IEEE Geosci. Remote. Sens. Lett.* **2023**, *20*, 7000205. [CrossRef]
12. Su, N.; Chen, X.; Guan, J.; Huang, Y. Maritime Target Detection Based on Radar Graph Data and Graph Convolutional Network. *IEEE Geosci. Remote. Sens. Lett.* **2022**, *19*, 4019705. [CrossRef]

13. Qin, F.; Bu, X.; Zeng, Z.; Dang, X.; Liang, X. Small Target Detection for FOD Millimeter-Wave Radar Based on Compressed Imaging. *IEEE Geosci. Remote. Sens. Lett.* **2022**, *19*, 4020705. [CrossRef]
14. Krotosky, S.J.; Trivedi, M.M. On color-, infrared-, and multimodalstereo approaches to pedestrian detection. *IEEE Trans. Intell. Transp. Syst.* **2007**, *8*, 619–629. [CrossRef]
15. 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.
16. Zhang, J.; Lei, J.; Xie, W.; Fang, Z.; Li, Y.; Du, Q. SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery. *arXiv* **2022**, arXiv:2209.13351. [CrossRef]
17. Fang, Q.; Wang, Z. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognit.* **2022**, *130*, 108786.
18. Konig, D.; Adam, M.; Jarvers, C.; Layher, G.; Neumann, H.; Teutsch, M. Fully convolutional region proposal networks for multispectral person detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 49–56.
19. Li, C.; Song, D.; Tong, R.; Tang, M. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognit.* **2019**, *85*, 161–171. [CrossRef]
20. 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. [CrossRef]
21. Yang, X.; Qiang, Y.; Zhu, H.; Wang, C.; Yang, M. BAANet: Learning bi-directional adaptive attention gates for multispectral pedestrian detection. *arXiv* **2021**, arXiv:2112.02277.
22. Zhuang, Y.; Pu, Z.; Hu, J.; Wang, Y. Illumination and Temperature-Aware Multispectral Networks for Edge-Computing-Enabled Pedestrian Detection. *IEEE Trans. Netw. Sci. Eng.* **2021**, *9*, 1282–1295. [CrossRef]
23. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
24. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object detection via regionbased fully convolutional networks. *Proc. Adv. Neural Inf. Process. Syst.* **2016**, *29*, 1–22.
25. Bochkovskiy, A.; Wang, C.Y.; Liao HY, M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
26. Ghose, D.; Desai, S.M.; Bhattacharya, S.; Chakraborty, D.; Fiterau, M.; Rahman, T. Pedestrian detection in thermal images using saliency maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 1–10.
27. Li, S.; Li, Y.; Li, Y.; Li, M.; Xu, X. YOLO-FIRI: Improved YOLOv5 for Infrared Image Object Detection. *IEEE Access* **2021**, *9*, 141861–141875. [CrossRef]
28. Marnissi, M.A.; Fradi, H.; Sahbani, A.; Ben Amara, N.E. Feature distribution alignments for object detection in the thermal domain. *Vis. Comput.* **2022**, *39*, 1081–1093. [CrossRef]
29. Deng, Q.; Tian, W.; Huang, Y.; Xiong, L.; Bi, X. Pedestrian detection by fusion of RGB and infrared images in low-light environment. In Proceedings of the 2021 IEEE 24th International Conference on Information Fusion (FUSION), Sun City, South Africa, 1–4 November 2021; pp. 1–8.
30. Chen, X.; Liu, L.; Tan, X. Robust Pedestrian Detection Based on Multi-Spectral Image Fusion and Convolutional Neural Networks. *Electronics* **2021**, *11*, 1. [CrossRef]
31. Zhang, H.; Fromont, E.; Lefevre, S.; Avignon, B. Guided attentive feature fusion for multispectral pedestrian detection. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 72–80.
32. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Online, 23–28 August 2020; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
33. Zhu, J.; Chen, X.; Zhang, H.; Tan, Z.; Wang, S.; Ma, H. Transformer Based Remote Sensing Object Detection with Enhanced Multispectral Feature Extraction. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
34. 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**. [CrossRef] [PubMed]
35. Liu, H.; Liu, F.; Fan, X.; Huang, D. Polarized self-attention: Towards high-quality pixel-wise regression. *arXiv* **2021**, arXiv:2107.00782.
36. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery: A small target detection benchmark. *J. Vis. Commun. Image Represent.* **2016**, *34*, 187–203. [CrossRef]
37. Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; Luo, Z. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5802–5811.

38. Jia, X.; Zhu, C.; Li, M.; Tang, W.; Zhou, W. LLVIP: A visibleinfrared paired dataset for low-light vision. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 3496–3504.
39. FLIR. FLIR Thermal Dataset for Algorithm Training. 2018. Available online: https://www.flir.in/oem/adas/adas-dataset-form (accessed on 19 January 2022).
40. Zhang, H.; Fromont, E.; Lefevre, S.; Avignon, B. Multispectral fusion for object detection with cyclic Fuse-and-Refine blocks. In Proceedings of the IEEE International Conference on Image Processing, Virtual, 25–28 October 2020; pp. 276–280.
41. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
42. Mandal, M.; Shah, M.; Meena, P.; Vipparthi, S.K. SSSDET: Simple short and shallow network for resource efficient vehicle detection in aerial scenes. In Proceedings of the IEEE International Conference on Image Processing, Taiwan, China, 22–25 September 2019; pp. 3098–3102.
43. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal loss for dense object detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
44. Dhanaraj, M.; Sharma, M.; Sarkar, T.; Karnam, S.; Chachlakis, D.G.; Ptucha, R.; Markopoulos, P.P.; Saber, E. Vehicle detection from multi-modal aerial imagery using YOLOv3 with mid-level fusion. *Proc. SPIE* **2020**, *11395*, 1139506.
45. Wang, C.Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.

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.