remote sensing

MDPI

Article

Background-Aware Cross-Attention Multiscale Fusion for Multispectral Object Detection

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

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China; guorunze14@nudt.edu.cn (R.G.); sunxiaoyong14@nudt.edu.cn (X.S.); zpd05@nudt.edu.cn (P.Z.); sunbei08@nudt.edu.cn (B.S.); ssjing@nudt.edu.cn (S.S.)
* Correspondence: jeanakin@nudt.edu.cn

Abstract: Limited by the imaging capabilities of sensors, research based on single modality is difficult to cope with faults and dynamic perturbations in detection. Effective multispectral object detection, which can achieve better detection accuracy by fusing visual information from different modalities, has attracted widespread attention. However, most of the existing methods adopt simple fusion mechanisms, which fail to utilize the complementary information between modalities while lacking the guidance of a priori knowledge. To address the above issues, we propose a novel background-aware cross-attention multiscale fusion network (BA-CAMF Net) to achieve adaptive fusion in visible and infrared images. First, a background-aware module is designed to calculate the light and contrast to guide the fusion. Then, a cross-attention multiscale fusion module is put forward to enhance inter-modality complement features and intra-modality intrinsic features. Finally, multiscale feature maps from different modalities are fused according to background-aware weights. Experimental results on LLVIP, FLIR, and VEDAI indicate that the proposed BA-CAMF Net achieves higher detection accuracy than the current State-of-the-Art multispectral detectors.

Keywords: multispectral object detection; complementary information; priori knowledge; background aware; cross attention; multiscale fusion

Citation: Guo, R.; Guo, X.; Sun, X.; Zhou, P.; Sun, B.; Su, S. Background-Aware Cross-Attention Multiscale Fusion for Multispectral Object Detection. *Remote Sens.* **2024**, *16*, 4034. https://doi.org/10.3390/rs16214034

Academic Editors: Wei Li, Haiyong Gan, Heng-Chao Li and Wenshuai Hu

Received: 4 September 2024
Revised: 22 October 2024
Accepted: 28 October 2024
Published: 30 October 2024

1. Introduction

Object detection has long been a core technology in computer vision tasks due to its wide range of applications in autonomous driving, patrol, and remote sensing [1–6]. However, limited by the sensing performance and adaptive capability of the sensors, the detection based on single modality often struggles to cope with the sensor's own failures and dynamic perturbations in the environment and lacks robustness. Although visible sensors can provide abundant color and texture information, their images are susceptible to the phenomenon of the same object but different spectra, especially under complex conditions such as inclement weather and changing light conditions, leading to inconsistent performance [7]. In contrast, infrared sensors are more sensitive to temperature and radiation, which perform well under the aforementioned harsh conditions. Yet the low resolution and unclear edges of infrared images make it difficult to capture detailed features [8]. As shown in Figure 1a, the visible spectrum is poorly hierarchical in low-light scenes. The foreground and background cannot be recognized either in the image or in the feature space, while objects in infrared images are clearly distinguished from the background. The features of objects (red) and those of the background (blue) are divided into 2 groups. The opposite case is illustrated for Figure 1b. For this reason, even with the great advances in convolutional neural networks (CNNs), detection techniques using only one single source of data continue to encounter difficulties in increasingly elaborate environments.

To alleviate the limitations of single modality in terms of imaging and realize all-weather monitoring, more researchers are focusing on multispectral object detection, by

which the visual information of different spectra can be fused to achieve higher detection accuracy. The solutions for multispectral object detection are currently divided into manual methods and deep learning methods. On the one hand, manual methods are realized through the traditional approach of feature extraction and classifier. Yet the manually designed multimodal operator has limited capability for feature extraction, making it difficult to obtain reliable detection [9]. On the other hand, due to their strong characterization capabilities, CNN-based feature fusion methods [10–17] have been widely used for multispectral object detection, most of which are based on two-stream CNNs. However, existing fusion strategies based on two-stream CNNs only use element summation, multiplication, and splicing [18]. While they offer higher performance than single-modality detection, the interaction and correlation between modalities are not sufficiently taken into account, which implies less adaptability. Worse still, the lack of long-term dependency and the multiscale of different objects may exacerbate the imbalance of the network, leading to unsatisfactory results. In addition, considering the differences in light sensitivity between visible and infrared images, illumination-aware networks are proposed to learn the weight occupancy of different modalities [15,16]. The main idea of these methods is to predict the illumination-aware weights through a formulated gate function, followed by a proportionally weighted fusion of the two branches to obtain the final detection result. However, it is not sufficient to compute modal weights using only illumination information as a priori knowledge, since this method cannot measure other relevant factors or achieve satisfactory performance in complex weather. We aim to obtain intrinsic and complementary information within and across modalities via designing effective fusion mechanisms and learning reliable priori knowledge.

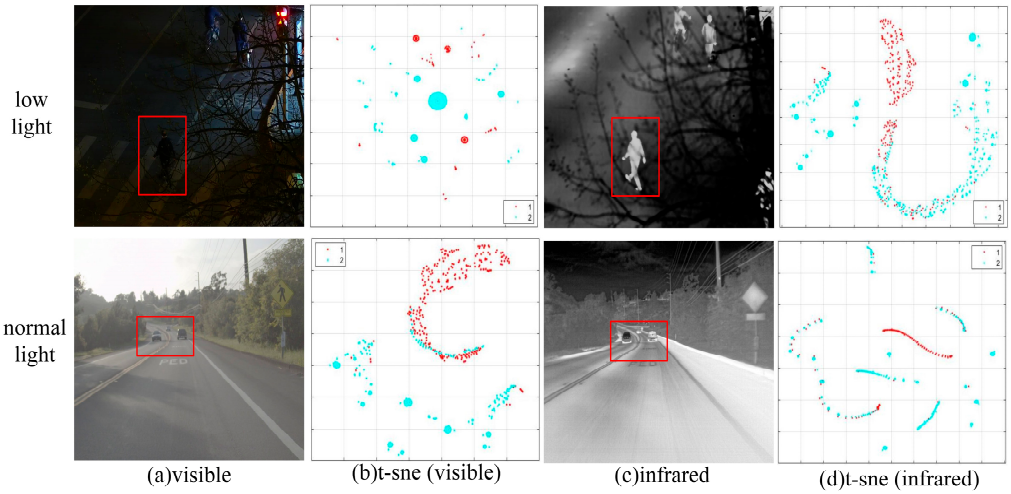


Figure 1. Visualization of foreground and background features in different modes using t-SNE technique. The first and second rows show the original and t-SNE images under low and normal light conditions, respectively. The foreground objects are labeled with red rectangular boxes. In the t-SNE images, the red points 1 indicate the foreground and the blue points 2 identify the background.

Therefore, we propose a novel background-aware guided cross-attention multiscale fusion network (BA-CAMF Net) to address the lack of interaction between modalities and prior knowledge. First, a background-aware module (BA) is designed to measure the light conditions as well as the contrast to obtain adaptive fusion weights. Then, we design a cross-attention multiscale fusion (CAMF) module to focus on inter-modality complement features and intra-modality intrinsic features, which consists of a differential attention (DA) module and a common attention (CA) module. Next, the enhanced features from the two modules are guided to sum according to the background-aware weights to obtain a combined feature map. Finally, the feature maps are fed into the detection head to yield detection results.

The main contributions of this article are as follows:

(1) A two-stream network BA-CAMF Net is proposed for multispectral object detection. The network achieves reliable detection through the guidance of a priori knowledge and the interaction within modalities;
(2) A BA module is designed to guide the fusion of visible and infrared modality, in which we utilize light conditions and contrast to obtain adaptive fusion weights for two branches;
(3) We put forward a CAMF module consisting of DA and CA modules to enhance inter-modality complement features and intra-modality intrinsic features and to achieve adaptive fusion;
(4) Extensive comparative experiments on three typical multispectral detection datasets (LLVIP, FLIR, and VEDAI) have been carried out, and the results show that the proposed BA-CAMF Net achieves higher detection accuracy than the current State-of-the-Art multispectral detectors.

The remainder of this paper is organized as follows: Section 2 briefly describes related work. In Section 3, the proposed BA-CAMF Net is presented. Section 4 illustrates extensive comparison and ablation experiments on three benchmark datasets, followed by some concluding remarks in Section 5.

2. Materials and Methods

In this section, we review several algorithms with single-modality object detection, CNN-based multispectral object detection, and illustration-guided multispectral object detection.

2.1. Single-Modality Object Detection

Single-modality object detection is known as generic object detection, which is one of the important branches in the field of remote sensing. It is commonly categorized into two-stage and single-stage methods. The two-stage algorithm is realized through two steps of candidate region selection and identification, the advantage of which is that candidate frames can fully extract the target features with higher accuracy. However, the two steps lead to an increase in model complexity and a decrease in speed, making it even less suitable for the fusion of different modalities. The single-stage algorithm accomplishes classification and positional regression directly by covering the image with a dense set of candidate frames or anchors. Among the single-stage detection algorithms, OverFeat [19] was first proposed to replace the sliding window search with CNNs. The SSD algorithm [20] utilized feature maps from different scales for classification and regression. Subsequently, due to its unrivaled performance, the YOLO algorithm became the mainstream framework for embedded object detection with a series of improvements, which include YOLOv3 [21], YOLOv4 [22], YOLOv5, YOLOX [23], YOLOv6 [24], and YOLOv7 [25]. The innovations include multi-scale training, network structure optimization, loss function modification, anchor adaptation, model reparameterization, and so on, all of which are also widely used in the field of remote sensing. In fact, single-modality object detection is adopted for feature extraction as part of multispectral detection networks.

2.2. CNN-Based Multispectral Object Detection

A single data source is difficult to meet the information needs in industry applications. Multispectral object detection, especially visible-infrared fusion detection, which allows complementary access to richer features, has become one of the future development directions. Currently, the widely used feature fusion architecture is the two-stream CNNs, which aim to solve the problems of weak alignment and large inter-modal variance in the fusion process. Weak alignment problems usually include differences in location, scale, and angle. Considering the factors that exist in sensor locations and internal references, AR-CNN [26] is proposed to align the regional properties of different modalities by means of a regional feature alignment module and to be trained in an end-to-end manner. Zhou et al. [27] propose MBNet, which is designed with a differential modality-aware fusion module to

solve the weakly aligned problem. However, the weak alignment problem can be solved by mapping of matrices, and the above methods do not really mine the differences between modals. In addition, attention structures are used to fuse complementary information across modalities and suppress noise, thus enhancing the feature representation of single-branch networks. Zhang et al. [28] designed a multispectral feature adaptive weighted network using the attention module. Fang et al. [11] achieved lightweight multispectral feature fusion at low computational cost by extracting attention maps from between modalities. However, the above methods do not fully utilize the relevance within modalities, so they perform poorly in situations such as low illumination.

2.3. Illustration-Guided Multispectral Object Detection

Recently, some researchers have taken illumination conditions into account when fusing different modalities. Wang et al. [29] proposed an illumination-aware network for the global enhancement of low-light images. In particular, the illumination-aware network may help to enhance features based on the illumination conditions. Additionally, illumination information is also used to verify the feasibility of multispectral pedestrian detection. The faster RCNN is improved by designing gate functions to measure the illumination conditions [15]. However, the performance of the approaches based on the illumination-aware network is not good in dawn, dusk, and background-mixed scenes, which means that relying only on illumination information for adaptive fusion is not sufficient.

In summary, how to design a two-stream fusion network to learn feature differences across modalities remains a valuable and challenging task. Inspired by attention feature fusion and illumination-aware networks, we propose an effective background-aware guided cross-attention multiscale fusion network that learns intrinsic and complementary intra-modality and inter-modality features by fusing them in a proportionally weighted manner to achieve an adaptive multispectral detection.

3. Methodology

As shown in Figure 2, the proposed BA-CAMF Net consists of a backbone network, a BA, and CAMF.

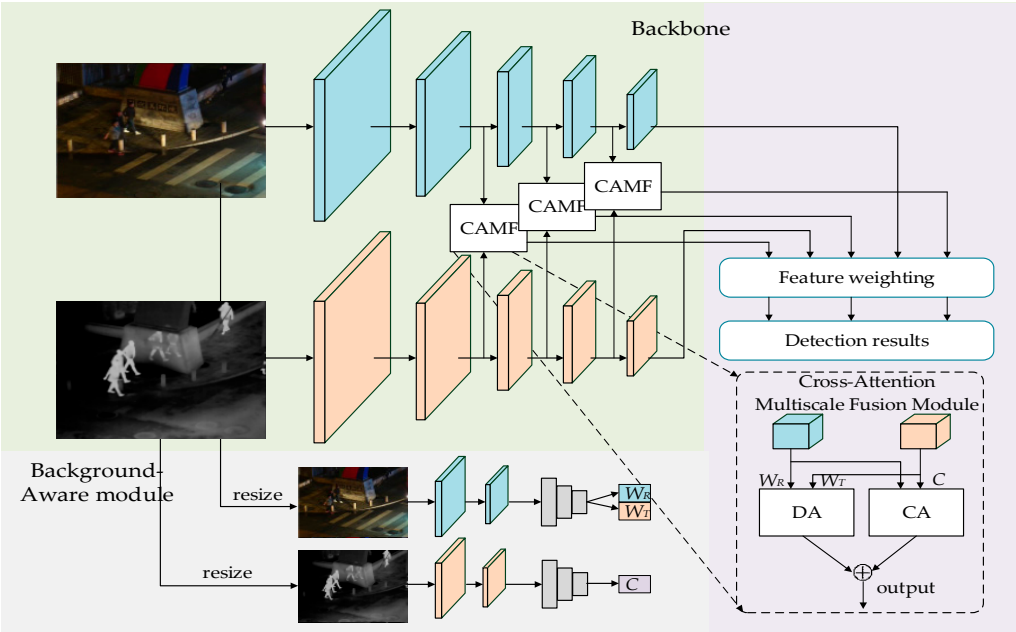


Figure 2. Structure of BA-CAMF Net. The blue and orange cubes represent the visible and infrared branches of the network, respectively. The region connected by the dashed line describes the CAMF structure proposed in this paper.

In order to optimize the feature extraction capability of the network and to improve the robustness of the model in complex scenarios, the backbone network employs a two-branch network consisting of Darknet53 and cross-stage partial structures (CSP). BA utilizes light conditions and the contrast as a priori knowledge to guide the fusion of modalitis. CAMF consists of two parts, the DA module and the CA module. Features from the two parts are fused in a weighted manner to obtain the fused attention map. Finally, the detection results are generated by feature weighting.

3.1. BA Module

The factors affecting the reliability of multispectral detection are complex and diverse. Currently, academics mainly use illumination information as a guide for multimodal fusion. Although illumination information is effective for cross-modality fusion, utilizing only illumination information still has the following limitations: (1) illumination-aware guided methods perform better in stationary scenarios such as daytime and nighttime, but it is difficult to discriminate between objects and the surrounding area in complex backgrounds such as rain, fog, and camouflage; (2) illumination information is usually used to guide the fusion of inter-modality features but is insufficient for intra-modality feature enhancement. However, in foggy or low-visibility scenes, where more factors affect image quality, light conditions play a more limited role. In this case, contrast is crucial for distinguishing objects from the background. Therefore, we propose BA module, in which the light and the contrast are calculated as the prediction weights to guide the adaptive fusion of intra-modality and inter-modality features.

Due to the difference in spectral bands, visible images are more dependent on external light sources than infrared images. Visible images vary greatly from day to night. Infrared images, on the contrary, are passive imaging, reflecting more the difference between the radiation of the target and the background. In general, visible images have higher imaging quality and contain more color and texture information in daytime scenes. Infrared images are sharper and give an outline of objects in dark scenes. Therefore, during the fusion process, when the scene is daytime, visible images contribute more to the detection results. When the scene is dark, the information in visible images fails and infrared images contribute more to the detection results. In order to quantify the effect of light conditions on fusion, we calculate the light conditions. Given a visible image as I_v, the probabilities that the image belongs to day or night are defined as t_d and t_n. Note that, $t_d + t_n = 1$. In natural scenes, daytime light conditions are usually better than those of dark scenes. The better the light conditions, the greater the t_d. We intend to use the probability values to represent the perceptual weights contributed by the different modalities. Due to the binary nature of the light source, t_d and t_n will be close to 0 or 1. If the values are multiplied directly by the results of the branches, the modality with lower probability would be significantly suppressed during the fusion process. In order to optimize the weights used in fusion, a gate function is designed to realign the weights of the two modalities so that their complementary information can be more fully integrated. The function is

$$W_v = \frac{p_d - p_n + 1}{2} \tag{1}$$

$$W_i = \frac{p_n - p_d + 1}{2} \tag{2}$$

where W_v and W_i denote the weights used to guide the fusion of visible and infrared images, and $W_v + W_i = 1$. When the daytime probability p_d is larger, the weight for visible images is greater than $1/2$ and the weight for infrared images is no longer close to 0. And vice versa. Therefore, we design a classification network to predict the light intensity to guide the fusion of inter-modality features. The image pair is resized to 128×128 and fed into BA. The visible image is fed into the light prediction network, which consists of a convolutional layer and a fully connected layer. After the convolutional layer, an activation layer and 2×2 adaptive pooling are added to compress and extract light features. Subsequently,

features are fed into the fully connected and computed to convert the output to the desired weights. In Figure 3, the blue box represents the prediction process of the light conditions. In this case, the blue cubes consist of a convolution module, an activation layer, and a pooling layer. The gray box represents the fully connected layer.

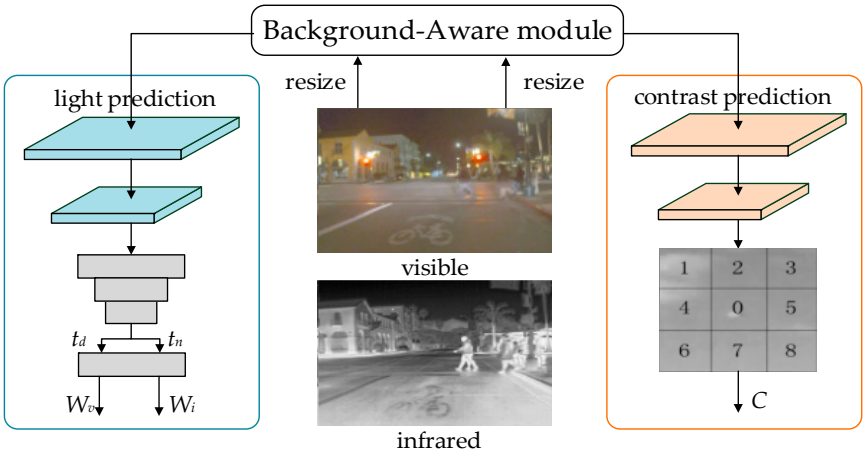


Figure 3. Structure of BA. Visible images are used for light prediction, and infrared images are used for contrast prediction.

As mentioned above, using only light information to guide fusion is not sufficient. For example, the similarity in color in visible images and the similarity in temperature in infrared images between objects and backgrounds can affect the accuracy of detection. The higher the similarity between objects and the background, the more difficult it is to be discriminated. Therefore, this paper introduces the contrast as a priori knowledge to guide the fusion of intra-modality features. The contrast is measured by the gray scale difference between objects and the surrounding background. First, the target region is divided into a 3×3 grid, and a variable m is defined to represent the pixel mean of the remaining grid regions except the object. The calculation process is

$$m_i = \frac{1}{N_i} \sum_{j=1}^{N_i} X_j^i \tag{3}$$

where i represents the ordinal number of the region, and N is the total number of pixel points in each region. j represents the ordinal number of the pixel points and X represents the gray value of the pixel points. Equation (3) calculates the average pixel intensity of the divided region, which is crucial for differentiating them in multispectral images. During rectangular box sliding, the target is usually located in the center region. The variable L is introduced to represent the largest gray value in the object region. The role of L is to quantify the pixel differences between the object and the rest of the region. The contrast difference c can be described by

$$c = \frac{1}{n} \sum_{i=1}^{n} \frac{L}{m_i} \tag{4}$$

The process of calculating the contrast is shown in the right part of Figure 3. Area 0 is the object area. Here, L is the maximum gray value of region 0, i represents the ordinal number of the background region, and n is 8. Therefore, contrast indicates the mean value of the pixel difference between the target region and the remaining 8 background regions. The process is shown on the right side of Figure 3. By calculating the contrast, the network better discriminates objects from the background, which would be used in the subsequent fusion module.

3.2. CAMF Module

Both visible and infrared images have their intrinsic and complementary information, and how to fuse the two modalities is the key to multispectral object detection. However, most of the existing methods based on two-branch networks only use simple fusion schemes, which cannot fully utilize the inter-modality and intra-modality features. In addition, rough combinations and connections also increase the difficulty of network learning, which leads to the degradation of detection performance. Inspired by differential amplification circuits, in which the differential mode and common mode signals are amplified and suppressed respectively, a CAMF module is proposed. CAMF consists of two parts: the inter-modality DA module and intra-modality CA module, as shown in Figure 2. Given visible and infrared convolutional feature maps as M^V and M^I, the differential features M^D and the common features M^C can be represented as

$$M^D = \frac{M^C + M^D}{2} - \frac{M^C - M^D}{2} = M^V - M^I \tag{5}$$

$$M^C = \frac{M^C + M^D}{2} + \frac{M^C - M^D}{2} = M^V + M^I \tag{6}$$

M^D can be viewed as the difference between the two modalities, which is obtained by subtraction to enhance the inter-modality specific feature. On the contrary, M^C can be regarded as the sum of the two modalities, which is obtained by addition to enhance the intra-modality consistency feature. Based on this, our CAMF module defines two new hybrid modals for the final fusion.

3.2.1. Inter-Modality DA Module

Inspired by signals in differential circuits, the inter-modality DA module aims at extracting specific features by computing the difference between visible and infrared modalities. As shown in Figure 4, the differential features are enhanced by the channel attention weighting mechanism.

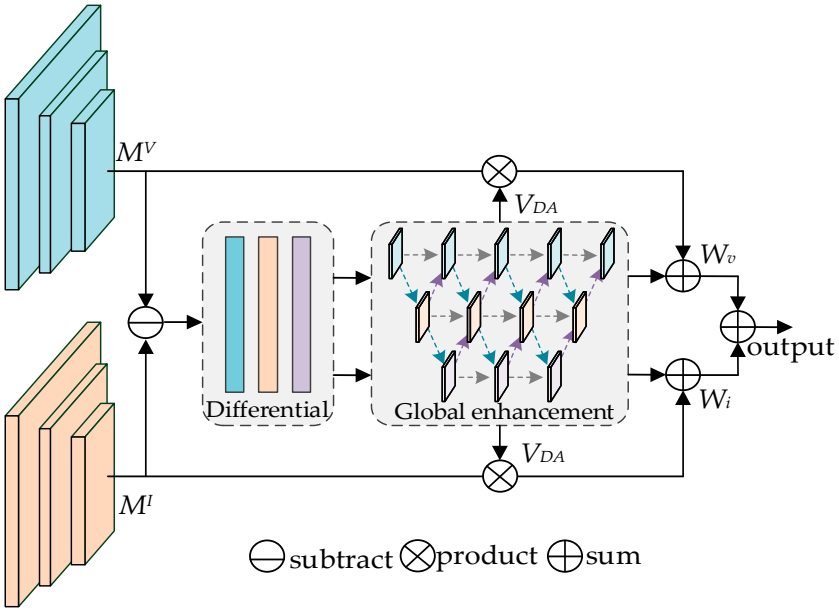


Figure 4. Structure of DA. The differential features are pooled and enhanced to obtain the feature map.

First, visible features M^V and infrared features M^I are input into the module as the initial values and differential features M^D are obtained by direct subtraction. Second, the differential features are encoded into global vectors V_{GAP} and V_{GMP} by global average pooling (GAP) and global maximal pooling to integrate the global spatial information. The

global vectors represent the differences in channel characteristics across modalities. Then, the vectors are sent into a shared convolution, the outputs of which are summed to obtain the channel attention map V_{DA}. Moreover, the attention map is multiplied, respectively, with the visible and infrared feature maps for adaptive aggregating. The results are summed with the input modality to obtain feature maps after differential amplification. Finally, the feature maps are summed according to the weights generated by BA to obtain the output of the DA module. This process can be expressed by

$$M_{DA} = W_v \cdot M^V \otimes (1 + V_{DA}) + W_i \cdot M^I \otimes (1 + V_{DA}) \tag{7}$$

Through differential, compression, excitation, and weighted fusion, the DA module adaptively learns the importance of different channels across modalities. The generalization is also enhanced. Notably, DA draws on residual networks for enhancing the stability of the network. Differential feature maps are added to the input modalities through jump connections, which could avoid the loss of key features.

3.2.2. Inter-Modality CA Module

The similarity between foreground and background in multispectral object detection affects the detection performance. In addition to complementary features, intrinsic features are also crucial for discriminative feature extraction. Therefore, the CA module is designed to focus on intra-modality shared information guided by contrast weights. As shown in Figure 5, CA sums the features of two branch networks and remixes them into a new feature to achieve an enhanced feature map.

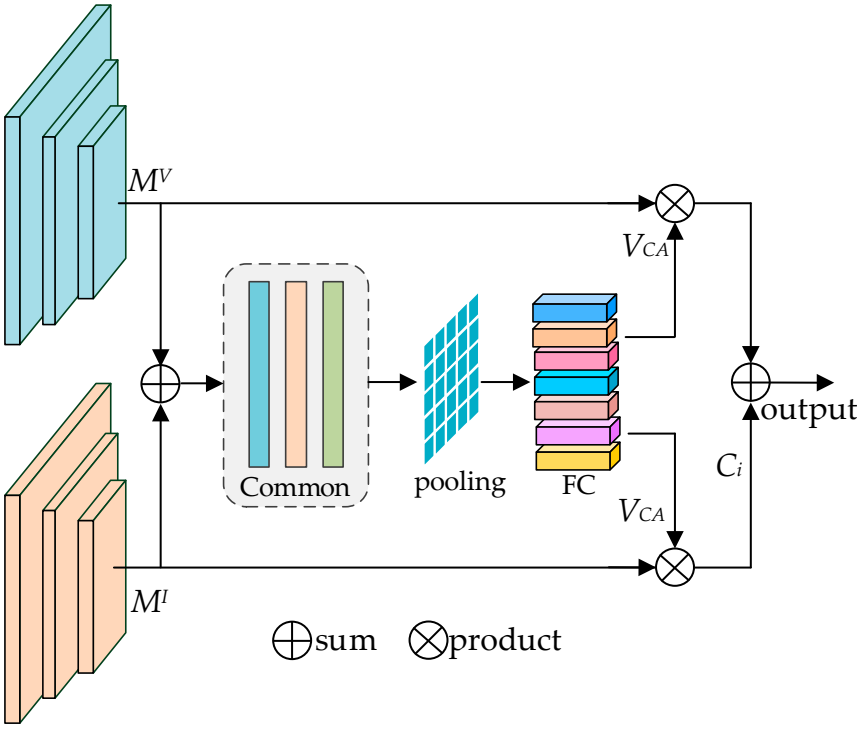


Figure 5. Structure of CA. The common features are pooled and enhanced to add weights to the original features.

First, the visible features M^V and infrared features M^I are used as inputs and directly summed to obtain the common features M^C. Next, the visible attention maps V_{CA}^V and infrared attention maps V_{CA}^I are computed through GAP and a fully connected layer (FC). Subsequently, the attention maps of the two modalities are multiplied, respectively, with

the input features and contrast weights from BA. Finally, the enhanced shared features are summed up to obtain the output of ca within the modality. This process can be expressed by

$$M_{CA} = M^V \otimes V_{CA}^V + C_i \cdot M^I \otimes V_{CA}^I \tag{8}$$

Through summation, weights sharing, compression, and normalization, the model achieves adaptive channel selection. The innovations of this module are as follows: (1) the parameters of FC are shared to reduce the dimension of features and improve the computational efficiency; (2) through channel selecting and the guidance of prior knowledge, the weights of the model are redistributed to the feature channels of the visible and infrared modality to avoid the introduction of redundant features; (3) the use of skip connections enables the network to reuse shallow features and improves the representation of complex features.

3.2.3. Multiscale Cross-Fusion Strategies

Fusing the outputs is the final step in our CAMF module. Generally, two-branch features are fused by adding and subtracting. However, existing studies have demonstrated that the above approaches may exacerbate the imbalance of the network. In addition, variations in scales of objects, especially for some small-sized targets, can also lead to the degradation of the model's performance. Therefore, a multiscale cross-fusion strategy is designed to achieve the fusion of cross-modality images, through which different levels of feature maps are interacted with each other. As shown in Figure 2, we extract feature maps from different modalities and scales (small, medium, and large) out and feed them into our CAMF module, followed by connecting horizontally and feature weighting to achieve the fusion of multiscale features. First, we use a convolutional layer after each result to reduce their dimensions to 1/2 of the original ones. Then, the bilinear interpolation is performed to restore them as the input aspect. Finally, the features at different scales are spliced together as global features for multiscale fusion. The step of fusion is shown as follows:

$$F_{FUSE} = \sum_{i=1}^{3} \left(M_{CA}^i + M_{DA}^i \right) \tag{9}$$

where i takes values in the range of 1, 2, and 3, representing the attention feature maps at three different scales: small, medium, and large, respectively. It is worth noting that the features of both visible and infrared modalities are processed and refined by the attention module to avoid the loss of key information.

3.3. Loss Function

To capture inter-modality and intra-modality information, we propose the multitask perception loss in this paper. It is noted that the contrast needs not to be trained using a separate network, so the loss function does not take contrast into account. Therefore, the multitask loss consists of detection loss, light condition loss and which are used to calibrate the multispectral detection results and the light prediction weights, respectively. The loss is defined as

$$L = L_d + L_l \tag{10}$$

The detection loss consists of the classification loss L_{cls}, the bounding-box regression loss L_{reg} and the confidence loss L_{conf}. The classification loss uses the strong correlation of the states to enable that the labels could better guide the learning of the category. The definition of L_{cls} is shown in Equation (11). The bounding-box regression loss is inspired by GIoU, which is proposed to alleviate the gradient problem of IoU loss. We added a penalty term to the original loss and defined it in Equation (12). The confidence loss is formulated as

Equation (13), which mainly solves the problem of imbalance in the proportion of different kinds of objects. It is suitable for complex scenarios, such as few samples and various scales.

$$L_{cls} = -[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{11}$$

$$L_{reg} = 1 - IoU(B, A) + \frac{|C - (B \cup A)|}{|C|} \tag{12}$$

$$L_{conf} = -\alpha_i (1 - p_i)^{\tau} \log(p_i) \tag{13}$$

where i denotes the ordinal number of the sample, p denotes the probability predicted by the model, and y is a binary variable taking the value of 0 or 1. A denotes the area of the real box, B denotes the area of the predicted box, and C denotes the area of the smallest box that contains both the real and the predicted. α is used to address the imbalance between positive and negative samples, and τ is used to address the imbalance between difficult and easy samples, which have been set to 0.25 and 2, respectively, in this paper.

The fusion of inter-modal complement and intra-modality intrinsic information relies heavily on the guidance of the background-aware module, especially the perception of light conditions. The light condition reflects the strength of the light conditions in the image and can be viewed as a classifier that calculates the probability of belonging to daytime and nighttime. Therefore, we use the cross-entropy loss to constrain its training process with the following equation:

$$L_l = -z \log \sigma(x) - (1 - z) \log(1 - \sigma(x)) \tag{14}$$

where z is the label of the light condition, x denotes the probability that the image belongs to the daytime, and σ is a softmax function that normalizes the light condition probability to [0,1]. In order to fully characterize the strength of the light conditions, the value space of z is set to 0, 0.5, and 1.0 to denote the dark, low light, and daytime scenes, respectively.

4. Experiments

In this section, experiments are conducted on VEDAI [30], FLIR-aligned [31], and LLVIP [32] to vertify our BA-CAMF net. To begin with, the datasets and evaluation metrics are introduced. Second, we describe the experimental setup and deployment. Then, comparative experiments are carried out between the State-of-the-Art methods and our BA-CAMF Net. Subsequently, we conduct ablation studies for the proposed method. Finally, we discuss some limitations.

4.1. Datasets and Evaluation Metrics

We compare the available dual-modality detection datasets, focusing on three datasets with remote sensing, roads, and low illumination as backgrounds: VEDAI, FLIR, and LLVIP. Figure 6 shows the distribution of variables such as width and height of the three datasets and the change of scale, which indicates that the datasets used in this paper cover examples of different scales and have good representation.

The VEDAI dataset serves as a database in the field of optical remote sensing images, which covers nine different types of small and medium-sized transport vehicles under the backgrounds such as cities, roads, fields, forests, etc. The dataset contains 1246 image pairs and 3640 instances. It is characterized by large differences in object sizes, which can validate the algorithm's ability for multi-scale objects.

As a benchmark in the field of autonomous driving, the FLIR-aligned dataset contains a total of 5142 bimodal image pairs for objects as pedestrians, bicycles, and vehicles under the background as streets and highways. Since the weak alignment between modalities is not mentioned in this paper, in order to avoid the network failing to converge, the experiments are performed on the FLIR-aligned dataset, which is made by manually removing unaligned images from the original dataset. For convenience, the FLIR refers to the alignment version.

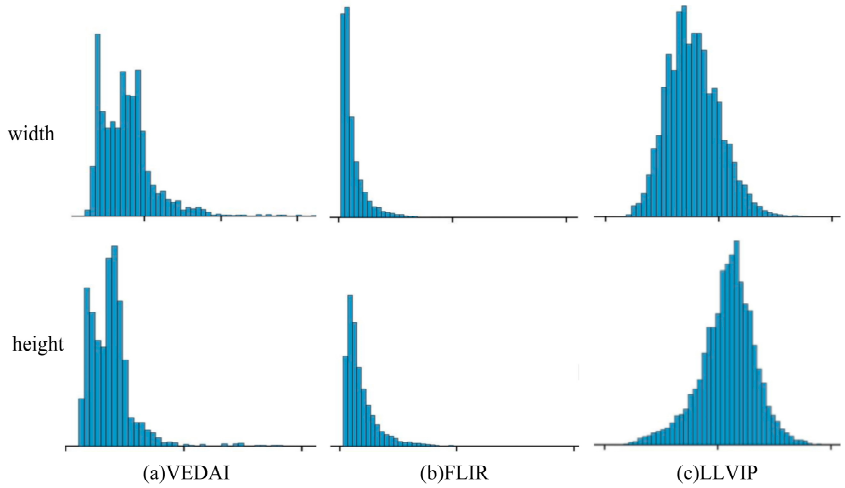


Figure 6. Distribution of datasets and changes in scale. The first column is the width of objects, and the second column is the height of objects.

The LLVIP dataset is designed to solve the problem of multispectral pedestrian detection in low-light scenes, with a ratio of 12:1 between night and day scenes. The dataset contains 15,488 image pairs, of which 60% are used for training and the rest for testing and validation. Compared with other datasets, the image pairs from different modalities in the LLVIP dataset are strictly aligned in time and space.

To evaluate and compare the performance of the network in this paper, P (precision), R (recall), mAP, which are the most recognized metrics are adopted. The metrics are defined as

$$P = \frac{TP}{TP + FP} \tag{15}$$

$$R = \frac{TP}{TP + FN} \tag{16}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i = \frac{1}{n}\sum_{i=1}^{n} P_i(r)dr \tag{17}$$

where TP is true positive and denotes the number of objects that are correctly classified as positive cases. Otherwise, they are considered false-positive (FP) cases. FN is false negative, which means the number of objects incorrectly classified as negative cases. AP is the integral of the precision-recall curve (PRC) for each category. mAP_{50} is the mean of all the AP values for all forms when IoU = 0.5. mAP is the more challenging metric, which is the mean of the AP values when IoU = 0.5:0.95.

4.2. Experimental Setup

The proposed algorithm is implemented in Pytorch 1.9, and all experiments are conducted on a ubuntu 20.04 system with 4 GPUs. To avoid instability, the training warms up with a lower learning rate of 0.01 and the SGD optimizer is used with a momentum of 0.93. In addition, the training parameters are set as follows: the elapsed time is set to 120, the batch size is set to 16, and the numworker is set to 8. For data enhancement, a mosaic enhancement method is also adopted with random flipping and rotation. It should be noted that the settings of the training parameters are consistent across experiments; otherwise, the experimental results may be affected by other factors and the performance of the algorithm cannot be measured.

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

To validate the effectiveness of BA-CAMF Net in this paper, we compare it with baseline detectors and other State-of-the-Art multispectral detection networks, including CFT [11], LRAF-Net [33], YOLO-Fusion [5], ICAFusion [34], GAFF [28], TFDet [35], and

single-modality detection networks Faster R-CNN and YOLOv9. The baseline detector is a two-stream CNN that uses element summation for feature fusion.

4.3.1. On the VEDAI Dataset

In remote sensing scenarios, we compare our BA-CAMF method with other related works. The experimental results are shown in Table 1. It can be observed from the bolded font that on the VEDAI dataset, our method achieves the best detection performance compared to other State-of-the-Art algorithms. For single-modal detection, our method outperforms the one-stage YOLOv9 and Faster R-CNN by 11.3% and 18.2%, respectively, in the mAP, which illustrates the importance of the dual-stream network. For multimodality detection, our method outperforms the best method by 1.3% and 0.7% in the mAP_{50} and the mAP, respectively.

Table 1. Comparison of different methods on the VEDAI dataset.

Model	Modality	Backbone	mAP_{50}	mAP
unimodality networks				
Faster R-CNN	visible	ResNet50	64.5%	38.9%
Faster R-CNN	infrared	ResNet50	71.2%	41.6%
YOLOv9	visible	CSPNet+ELAN	73.4%	42.5%
YOLOv9	infrared	CSPNet+ELAN	75.6%	44.6%
multimodality networks				
CFT	two-stream	CFB	85.3%	56.0%
LRAF-Net	two-stream	Darknet53	85.9%	59.1%
YOLO-Fusion	two-stream	Darknet53	78.6%	49.1%
ICAFusion	two-stream	Darknet53	76.6%	44.9%
Baseline	two-stream	Darknet53+CSP	83.9%	54.2%
Ours	two-stream	Darknet53+CSP	**87.2%**	**59.8%**

In addition to quantitative comparisons, we also perform qualitative analysis on the VEDAI dataset. Figure 7 shows the original input image pairs, the detection results of baseline, and our BA-CAMF. It is noted that the ground truth is labeled in the input IR image. As shown in Figure 7a–c, objects in the remote sensing images have small target sizes, dense arrangements, and weak texture features, which are prone to being missed or detected falsely by using only a simple two-stream network. However, our method fully integrates the complementary and inherent features of the two modalities, which can significantly reduce the occurrence of missed and false detection, as shown in Figure 7d. Therefore, it is proven that our algorithm has excellent detection performance in remote sensing scenarios.

4.3.2. On the FLIR Dataset

In the road scenarios, the proposed method is compared with other detection methods, as shown in Table 2. It can be seen from the bold footer that the performance of our BA-CAMF is optimal on the FLIR dataset. In particular, our algorithm improves the mAP_{50} by 2.2% over the typical two-stream network LRAF-Net. In addition, compared with the State-of-the-Art unimodal detection algorithms YOLOv9, our method improves the mAP_{50} and mAP by 8.6% and 4.0%, respectively. This indicates that the fusion algorithm in this paper is effective and significantly improves the detection performance.

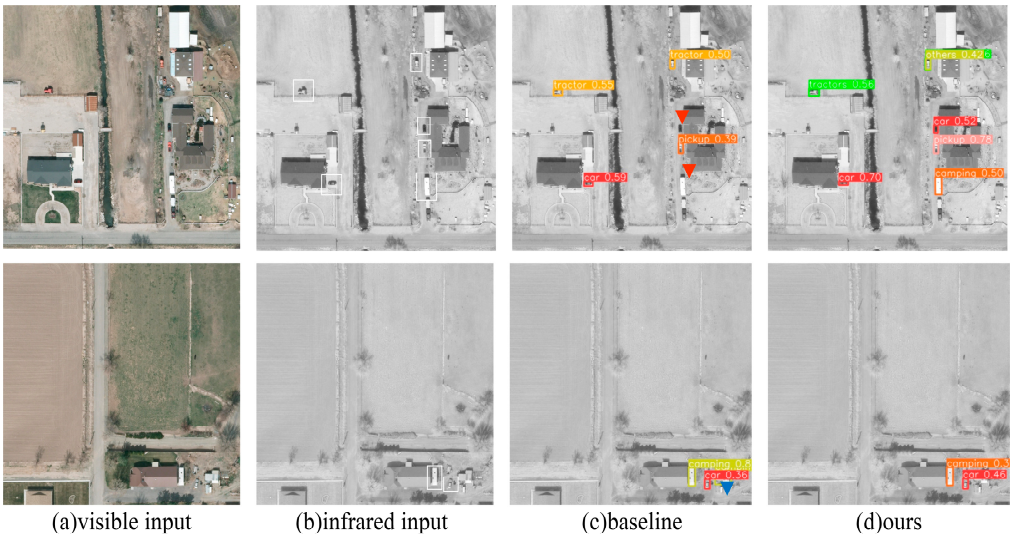


Figure 7. Visualization of detection results on VEDAI. The ground truth is labeled in the infrared input. Note that red triangles mean FNs, and blue triangles imply FPs.

Table 2. Comparison of different methods on the FLIR dataset.

Model	Modality	Backbone	mAP_{50}	mAP
unimodality networks				
Faster R-CNN	visible	ResNet50	64.9%	28.9%
Faster R-CNN	infrared	ResNet50	74.4%	37.6%
YOLOv9	visible	CSPNet+ELAN	69.5%	33.4%
YOLOv9	infrared	CSPNet+ELAN	73.9%	38.9%
multimodality networks				
CFT	two-stream	CFB	78.7%	40.2%
LRAF-Net	two-stream	Darknet53	80.5%	42.8%
GAFF	two-stream	ResNet18	72.9%	37.5%
ICAFusion	two-stream	Darknet53	79.2%	41.4%
Baseline	two-stream	Darknet53+CSP	80.3%	42.2%
Ours	two-stream	Darknet53+CSP	**82.5%**	**43.9%**

Similarly, we qualitatively analyze the FLIR dataset and select two scenarios, daytime and nighttime. The detection results are shown in Figure 8. The ground truth is labeled by white rectangular boxes in infrared images. Objects in the road scene may occlude each other, which leads to the degradation of the algorithm's performance. As shown in Figure 8c, the pedestrian overlapping with utility poles is not detected in the daytime scene, and the vehicle occluded by pedestrians is detected twice in the nighttime scene. In contrast, in Figure 8d, the above objects are detected. The reason is that our method takes into account both the global and local information, which effectively enhances the features in the case of occlusion.

4.3.3. On the LLVIP Dataset

The comparison results of the proposed algorithm with other algorithms in low-light scenarios are shown in Table 3. It can be seen from the bold that our BA-CAMF has State-of-the-Art performance on the LLVIP dataset compared to other algorithms, with the mAP_{50} reaching 97.9% and the mAP reaching 69.2%. Especially, the mAP is also 7.3% and 2.9% higher compared to YOLOv9 and LRAF-Net, respectively. In addition, from

the unimodal detection results, the mAP of infrared images is significantly higher than that of visible images, which indicates that infrared images have a better ability of feature representation in low-light scenarios.

Figure 8. Visualization of detection results on FLIR. The ground truth is labeled in the infrared input. Note that red triangles mean FNs, and blue triangles imply FPs.

Table 3. Comparison of different methods on the LLVIP dataset.

Model	Modality	Backbone	mAP_{50}	mAP
unimodality networks				
Faster R-CNN	visible	ResNet50	91.4%	49.2%
Faster R-CNN	infrared	ResNet50	96.1%	61.1%
YOLOv9	visible	CSPNet+ELAN	90.8%	50.0%
YOLOv9	infrared	CSPNet+ELAN	94.6%	61.9%
multimodality networks				
CFT	two-stream	CFB	97.5%	63.6%
LRAF-Net	two-stream	Darknet53	97.9%	66.3%
TFDet	two-stream	ResNet18	96.0%	59.4%
ICAFusion	two-stream	Darknet53	97.8%	64.1%
Baseline	two-stream	Darknet53+CSP	95.9%	63.5%
Ours	two-stream	Darknet53+CSP	**97.9%**	**69.2%**

Figure 9 shows some experimental results of the baseline and the proposed algorithm. It can be seen that infrared images in low illumination scenes have better detection performance and can visualize the difference between foreground and background. As shown in Figure 9b,c, when objects are overlapped or occluded, the baseline is prone to miss detection. Especially in Figure 9c, there is light source interference beside the pedestrian, which also exacerbates the missed detection. Our method effectively solves the above problem by deeply fusing the information of the two modalities without simply adding the two modalities as baseline does.

(a)visible input (b)infrared input (c)baseline (d)ours

Figure 9. Visualization of detection results on LLVIP. The ground truth is labeled in the infrared input. Note that red triangles mean FNs, and blue triangles imply FPs.

4.3.4. Inference Time

To validate the real-time performance of the algorithm, we conduct experiments on the FLIR dataset with a resolution of 512 × 512 multispectral image pairs. The inference time of our BA-CAMF Net and other dual-stream detection algorithms is shown in Table 4. All experiments in this section are performed with the same setting on the NVDIA RTX 3090 GPU from the USA. As can be seen from the table, our BA-CAMF Net runs slower than the unimodal networks, like Faster R-CNN, YOLOv5, and YOLOv9. The reason is that the model parameters and dimensions of a two-stream network are usually larger than those of a single-stream network. In Table 4, we also compare the algorithm with other multispectral networks. Although the speed of our BA-CAMF Net is slower than that of ICAFusion (38.5 FPS), it still outperforms CFT and TFDet and is greater than 20 FPS. Overall, we have designed a lightweight fusion module that can satisfy the real-time requirements in embedded applications.

Table 4. Comparison of inference time for different methods.

Model	Modality	Backbone	FPS
unimodality networks			
Faster R-CNN	visible	ResNet50	21.7
YOLOv5	visible	Darknet53	42.8
YOLOv9	visible	CSPNet+ELAN	25.6
multimodality networks			
TFDet	two-stream	ResNet18	7.7
CFT	two-stream	CFB	21.0
ICAFusion	two-stream	Darknet53	**38.5**
Ours	two-stream	Darknet53+CSP	21.2

4.4. Ablation Studies

To validate the effectiveness of the algorithm, we evaluate the performance of the proposed method and measure the impact of the proposed module in this paper. Due to its data covering different weathers, experiments are done on the FLIR dataset. First, we explore the impact of different fusion strategies on the detection performance and verify the effectiveness of the multiscale fusion strategy. Subsequently, we test the gain of BA and CAMF modules.

4.4.1. Necessity of Multiscale Cross-Fusion

This section focuses on the effectiveness of different stages and levels of fusion on detection performance. A diagram of different fusion methods is shown in Figure 10. Early fusion, also called pixel-level fusion, can be achieved by superimposing pixel values. Simple early fusion methods may disturb the original features, resulting in poor performance that is inferior to the performance of unimodal detection. Unlike early fusion, middle fusion can deeply interact features of different modalities, which is advantageous in capturing the correlation between different modalities and helps to improve the overall performance. Late fusion, also called decision-level fusion, focuses on the fusion of results and may ignore the possible inter-modality correlations. In addition, the arrangement of different attention modules also affects the performance. In this paper, a multiscale cross-fusion strategy is adopted. Moreover, the serial connection is also worth discussing.

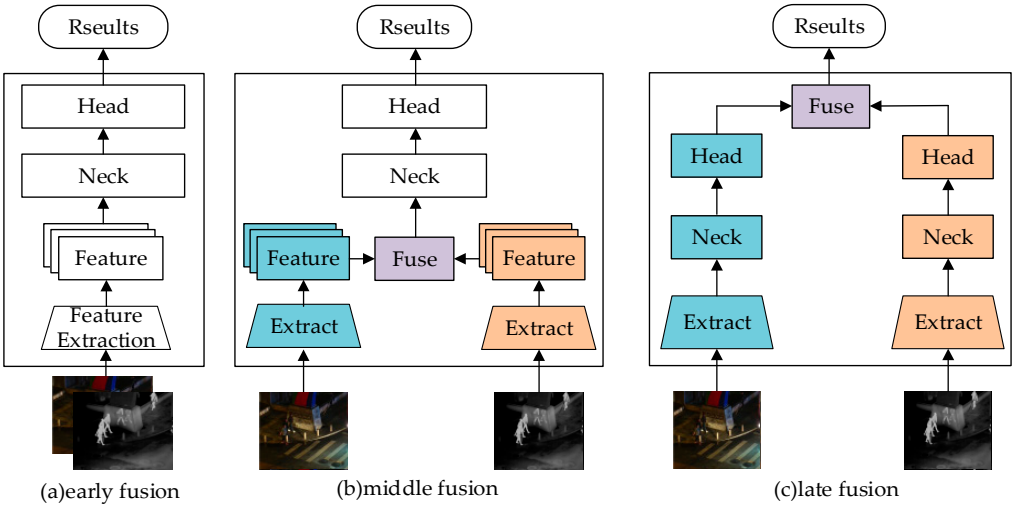


Figure 10. Different fusion strategies. (**a**) early fusion; (**b**) middle fusion; (**c**) late fusion.

Therefore, we carry out ablation studies in four ways: early fusion, cross-fusion, serial fusion, and late fusion. The results are shown in Table 5. It should be noted that Darknet53 is used for the backbone, and all models are trained for 100 epochs. It can be seen from the bold footer that the detection performance of middle fusion is optimal (including both cross-fusion and serial fusion), followed by late fusion. In addition, the performance of the cross-fusion method is better than that of the serial fusion method. The reason is that as the network passes forward, the cross-connection can retain different levels of features, while the serial connection can only retain deep semantic information and cannot learn shallow details. Therefore, in this paper, we use cross-connectivity to fuse information from different modalities.

Table 5. Comparison of different fusion levels on the FLIR dataset.

Method	Backbone	mAP_{50}	mAP
Visible	Darknet53+CSP	68.9%	33.8%
Infrared	Darknet53+CSP	74.2%	38.5%
Early Fusion	Darknet53+CSP	75.2%	38.3%
Cross Fusion (ours)	Darknet53+CSP	**82.5%**	**43.9%**
Series Fusion	Darknet53+CSP	79.5%	41.6%
Late Fusion	Darknet53+CSP	78.3%	39.3%

4.4.2. Necessity of the Proposed Module

The BA module adaptively guides the visible and infrared modalities to be fused based on the light information and contrast weights of the input images. Table 6 shows the performance of the networks before and after using the background-aware module, before and after using the CAMF module. √ indicates that the method is used in the model. It can be seen from the bold footer that after using the background-aware module, mAP_{50} and mAP are improved by 3.3% and 1.4%, respectively, which outperforms the performance of the network when only light information or only contrast is used. This suggests that the BA module can use the prior knowledge to better guide the network for fusion, which is adaptive to more complex backgrounds.

Table 6. Comparison of different modules on the FLIR dataset.

Method				Backbone	mAP_{50}	mAP
BA		CAMF				
Light	Contrast	DA	CA			
				Darknet53+CSP	76.5%	38.8%
√				Darknet53+CSP	78.6%	39.6%
	√			Darknet53+CSP	77.3%	39.1%
√	√			Darknet53+CSP	79.8%	40.2%
		√		Darknet53+CSP	80.1%	39.8%
			√	Darknet53+CSP	79.5%	39.4%
		√	√	Darknet53+CSP	81.2%	41.4%
√	√	√	√	Darknet53+CSP	**82.5%**	**43.9%**

Figure 11 lists the weights of the lighting conditions predicted by the background-aware module in the upper right corner of the image, with higher weights for visible modalities in daytime scenarios and for infrared modalities in nighttime scenarios. In the first graph visible weights can reach 0.557.

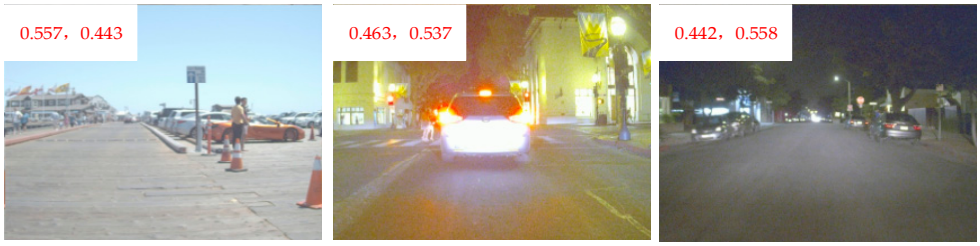
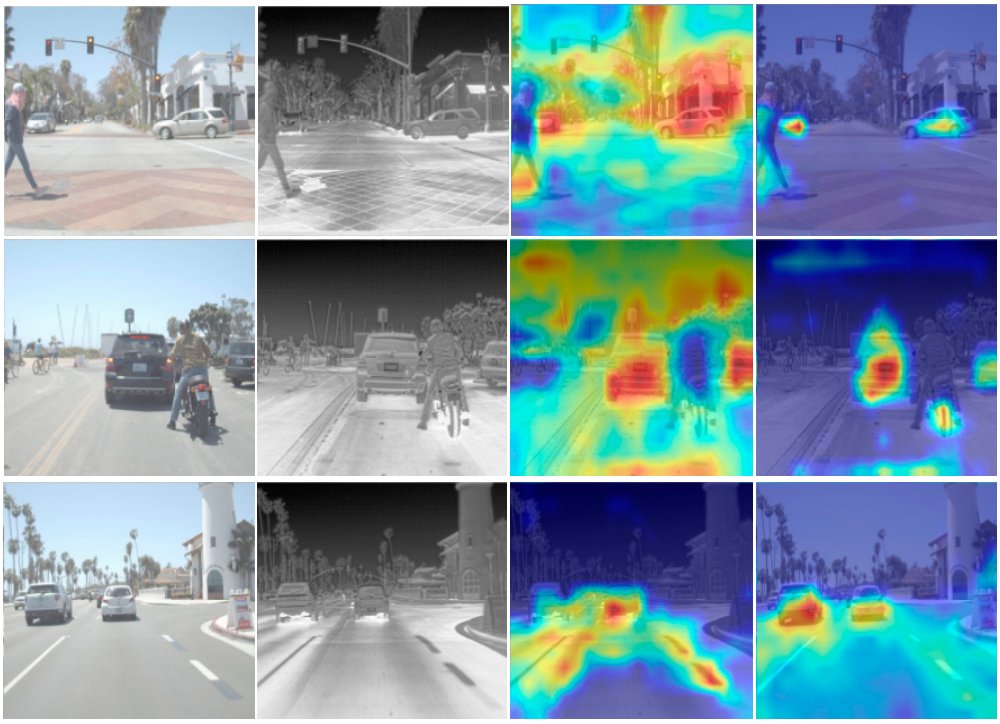


Figure 11. Demonstration of light conditions.

Our CAMF consists of a DA module, which focuses on the complementary features between modalities, and a CA module, which focuses on the inherent features of the modalities themselves. Table 6 shows that the DA module alone improves the mAP by 1.0% while the whole CAMF module improves the model's mAP by 2.6%. Figure 12 shows the visualization results of the feature map. The heatmap of single-branch features and fused features is shown on Figure 12c,d. It can be seen that the unimodal features are more dispersed while the fused features are more concentrated, in which objects are significantly enhanced.

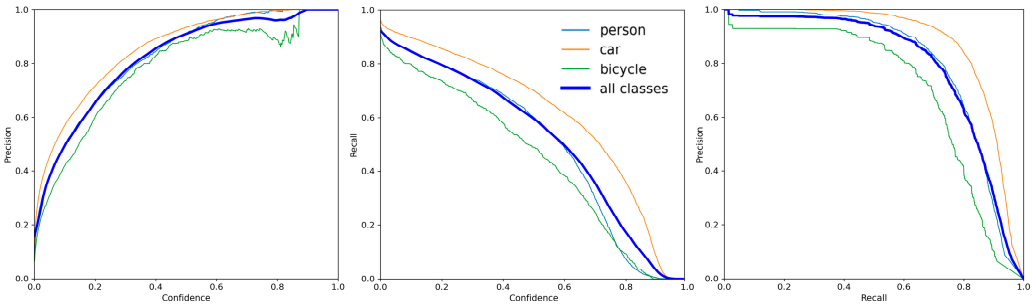
In addition, we explain the intrinsic relationship between the P, R, and PR metrics on the FLIR dataset. Figure 13a–c show the P-curve, R-curve, and PR-curve of this paper's algorithm, respectively. It can be seen that as the confidence increases, the precision of the model also increases, while the recall of the model shows a decreasing trend. When the

confidence level is greater than 0.3, the precision rate of the model is more stable. From Figure 13c, it can be seen that the proposed network maintains a high accuracy under different recall levels, which indicates that it has a good performance in dealing with different-sized objects. Therefore, the BA-CAMF Net designed in this paper is conducive to the interaction and fusion of multimodality information.



(a)visible input (b)infrared input (c)single branch (d)fusion

Figure 12. Visualization results of the feature map on the FLIR dataset.



(a)P curve (b)R curve (c)PR curve

Figure 13. Visualization results of the train process on the FLIR dataset.

5. Discussion

5.1. Limitations

As seen in Figures 7–9, the BA-CAMF Net in this paper has better performance and robustness under different light conditions or environments. However, there are still duplicate detection results in the case of densely arranged. When the arrangement is dense, the network tends to recognize the one overlapping object as more than one, which leads to false detection. In addition, the running time of the algorithm in this paper, although better than most of the dual-stream networks, is still lower than that of ICAFusion. Its real-time performance needs to be further optimized.

5.2. Ideas for Future Researches

For those problems, data augmentation and the improvement of loss function are better approaches. On public datasets, it is not easy to add new occluded samples, and mosaic enhancement can be used. Specialized loss functions can be designed to focus on the overlapping situation, such as repulsion loss, which can make the detection frame as far away as possible from the non-object region, thus reducing the false detections in the case of occlusion. Additionally, pruning, quantization, and embedded deployment are the next steps of the algorithm in this paper.

6. Conclusions

In this paper, we propose a visible-infrared fusion detection network, BA-CAMF Net, to solve the problems of poor correlation within modalities and lack of a priori knowledge in multispectral detection. BA-CAMF Net consists of the backone, BA, and CAMF. The backbone network employs a two-branch network consisting of Darknet53 and CSP. The BA module calculates the adaptive weights based on lighting conditions and contrast. The CAMF enhances module inter-modality complement features and intra-modality intrinsic features by multiscale cross-fusion of DA and CA guided by adaptive weights. In addition, we design a multitask function to balance the detection loss and background perception loss for fusion detection. Extensive comparative experiments on LLVIP, FLIR, and VEDAI have been carried out, and the results show that the proposed BA-CAMF Net achieves higher detection accuracy than the current State-of-the-Art multispectral detectors.

Author Contributions: Conceptualization, R.G. and X.G.; methodology, R.G. and B.S.; validation, X.G., X.S. and S.S.; formal analysis, X.G.; investigation, S.S.; resources, P.Z.; data curation, B.S.; writing—original draft preparation, X.S.; writing—review and editing, R.G.; visualization, P.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Hunan Provincial Postgraduate Research Innovation Programme, grant number XJCX2023041.

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

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

References

1. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
2. Zhang, C.; Chen, B.Y.; Lam, W.H.; Ho, H.W.; Shi, X.; Yang, X.; Ma, W.; Wong, S.C.; Chow, A.H. Vehicle re-identification for lane-level travel time estimations on congested urban road networks using video images. *IEEE Trans. Intell. Transp. Syst.* **2022**, *8*, 12877–12893. [CrossRef]
3. Feng, D.; Haase-Schütz, C.; Rosenbaum, L.; Hertlein, H.; Glaeser, C.; Timm, F.; Wiesbeck, W.; Dietmayer, K. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Trans. Intell. Transp. Syst.* **2021**, *3*, 1341–1360. [CrossRef]
4. Liu, Y.; Zhang, X.Y.; Bian, J.W.; Zhang, L.; Cheng, M.M. SAMNet: Stereoscopically attentive multi-scale network for lightweight salient object detection. *IEEE Trans. Image Process.* **2021**, *30*, 3804–3814. [CrossRef]
5. Li, C.; Cong, R.; Hou, J.; Zhang, S.; Qian, Y.; Kwong, S. Nested network with two-stream pyramid for salient object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9156–9166. [CrossRef]
6. Ren, X.; Bai, Y.; Liu, G.; Zhang, P. YOLO-Lite: An Efficient Lightweight Network for SAR Ship Detection. *Remote Sens.* **2023**, *15*, 3771. [CrossRef]
7. Fang, Q.; Wang, Z. Cross-Modality Attentive Feature Fusion for Object Detection in Multispectral Remote Sensing Imagery. *arXiv* **2021**. [CrossRef]
8. Zhang, T.; Wu, H.; Liu, Y.; Peng, L.; Yang, C.; Peng, Z. Infrared small target detection based on non-convex optimization with Lp-norm constraint. *Remote Sens.* **2019**, *11*, 559. [CrossRef]
9. Pang, S.; Ge, J.; Hu, L.; Guo, K.; Zheng, Y.; Zheng, C.; Zhang, W.; Liang, J. RTV-SIFT: Harnessing Structure Information for Robust Optical and SAR Image Registration. *Remote Sens.* **2023**, *15*, 4476. [CrossRef]
10. Song, K.; Bao, Y.; Wang, H.; Huang, L.; Yan, Y. A potential visionbased measurements technology: Information flow fusion detection method using RGB-thermal infrared images. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 5004813. [CrossRef]

11. Qing, F.; Da, H.; Zhao, W. Cross-modality fusion transformer for multispectral object detection. *arXiv* **2021**, arXiv:2111.00273.
12. Liu, J.; Zhang, S.; Wang, S. Multispectral deep neural networks for pedestrian detection. In Proceedings of the British Machine Vision Conference (BMVC), York, UK, 19–22 September 2016; pp. 1–13.
13. Wagner, J.; Fischer, V.; Herman, M.; Behnke, S. Multispectral pedestrian detection using deep fusion convolutional neural networks. In Proceedings of the 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2016), Bruges, Belgium, 27–29 April 2016; Volume 587, pp. 509–514.
14. 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.
15. 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]
16. Guan, D.; Cao, Y.; Yang, J.; Cao, Y. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Inf. Fusion* **2019**, *50*, 148–157. [CrossRef]
17. 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. [CrossRef]
18. Zhang, Y.; Yu, H.; He, Y.; Wang, X.; Yang, W. Illumination-guided RGBT object detection with inter- and intra-modality fusion. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 2508013. [CrossRef]
19. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *arXiv* **2013**, arXiv:1312.6229.
20. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016.
21. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. In *Computer Vision and Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2018; Volume 1804, pp. 1–6.
22. Bochkovskiy, A.; Wang, C.; Liao, H. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
23. Ge, Z.; Liu, S.; Wang, F. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
24. Li, C.; Li, L.; Jiang, H. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
25. Wang, C.; Bochkovskiy, A.; Liao, H. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
26. Zhang, L.; Zhu, X.; Chen, X.; Yang, X.; Lei, Z. Weakly Aligned Cross-Modal Learning for Multispectral Pedestrian Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
27. Zhou, K.; Chen, L.; Cao, X. Improving Multispectral PedestrianDetection by Addressing Modality Imbalance Problems. In Proceedings of the European Conference on Computer Vision (ECCV), Online, 23–28 August 2020.
28. Zhang, H.; Fromont, E.; Lefevre, S.; Avignon, B. Guided attentive feature fusion for multispectral pedestrian detection. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 72–80.
29. Wang, W.; Wei, C.; Yang, W.; Liu, J. GLADNet: Low-light enhancement network with global awareness. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 751–755.
30. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery: A small target detection benchmark. *J. Vis. Commun. Image Represent.* **2016**, *34*, 187–203. [CrossRef]
31. FREE Teledyne FLIR Thermal Dataset for Algorithm Training. Available online: https://www.flir.com/oem/adas/adas-dataset-form/ (accessed on 1 January 2020).
32. Jia, X.; Zhu, C.; Li, M.; Tang, W. Llvip: A visible-infrared paired dataset for low-light vision. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW 2021), Montreal, BC, Canada, 11–17 October 2021; pp. 3489–3497.
33. Fu, H.; Wang, S.; Duan, P.; Xiao, C.; Dian, R.; Li, S. LRAF-Net: Long-Range Attention Fusion Network for Visible–Infrared Object Detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 13232–13245. [CrossRef] [PubMed]
34. Ji, S.; Yi, C.; Yue, L.; Xin, Z. ICAFusion: Iterative Cross-Attention Guided Feature Fusion for Multispectral Object Detection. *arXiv* **2023**, arXiv:2308.07504v1.
35. Zhang, X.; Zhang, X.; Sheng, Z. TFDet: Target-aware Fusion for RGB-T Pedestrian Detection. *arXiv* **2023**, arXiv:2308.06361. [CrossRef]