*electronics*

MDPI

*Article*

# Multispectral Object Detection Based on Multilevel Feature Fusion and Dual Feature Modulation

**Jin Sun, Mingfeng Yin *, Zhiwei Wang, Tao Xie and Shaoyi Bei**

School of Automobile and Traffic Engineering, Jiangsu University of Technology, Changzhou 213001, China; sun1731997@163.com (J.S.); wang17368155902@163.com (Z.W.); xietao8301997@163.com (T.X.); bsy@jsut.edu.cn (S.B.)

* Correspondence: yinmingfeng@jsut.edu.cn

**Abstract:** Multispectral object detection is a crucial technology in remote sensing image processing, particularly in low-light environments. Most current methods extract features at a single scale, resulting in the fusion of invalid features and the failure to detect small objects. To address these issues, we propose a multispectral object detection network based on multilevel feature fusion and dual feature modulation (GMD-YOLO). Firstly, a novel dual-channel CSPDarknet53 network is used to extract deep features from visible-infrared images. This network incorporates a Ghost module, which generates additional feature maps through a series of linear operations, achieving a balance between accuracy and speed. Secondly, the multilevel feature fusion (MLF) module is designed to utilize cross-modal information through the construction of hierarchical residual connections. This approach strengthens the complementarity between different modalities, allowing the network to improve multiscale representation capabilities at a more refined granularity level. Finally, a dual feature modulation (DFM) decoupling head is introduced to enhance small object detection. This decoupled head effectively meets the distinct requirements of classification and localization tasks. GMD-YOLO is validated on three public visible-infrared datasets: DroneVehicle, KAIST, and LLVIP. DroneVehicle and LLVIP achieved mAP@0.5 of 78.0% and 98.0%, outperforming baseline methods by 3.6% and 4.4%, respectively. KAIST exhibited an MR of 7.73% with an FPS of 61.7. Experimental results demonstrated that our method surpasses existing advanced methods and exhibits strong robustness.

**Keywords:** multispectral object detection; remote sensing; visible-infrared images; multilevel feature fusion; dual feature modulation

check for updates

**Citation:** Sun, J.; Yin, M.; Wang, Z.; Xie, T.; Bei, S. Multispectral Object Detection Based on Multilevel Feature Fusion and Dual Feature Modulation. *Electronics* **2024**, *13*, 443. https://doi.org/10.3390/electronics13020443

Academic Editor: Chiman Kwan

Received: 3 January 2024
Revised: 12 January 2024
Accepted: 19 January 2024
Published: 21 January 2024

**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Object detection [1–5] is critical in computer vision and remote sensing image processing, with applications in autonomous navigation and search and rescue [6]. Prior object detection approaches largely rely on visible light sensors to capture image data, recording characteristics such as color, texture, and edge sharpness. However, these techniques face substantial challenges in low-light nighttime environments [7], often failing to capture detailed object features [8,9]. In contrast, infrared imaging, which relies on thermal radiation, remains largely unaffected by lighting changes or environmental conditions. This makes it particularly useful in challenging low-light scenarios, where infrared thermal emission provides supplementary details to overcome visibility issues [10]. Therefore, combining visible light and infrared modalities for enhanced nighttime object identification becomes an important research area. The successful integration of these distinct data streams to enhance detection performance represents a key focus of contemporary research endeavors.

Image fusion technology uses various approaches to merge visible-infrared images. These are categorized as pixel-level, feature-level, and decision-level [11] fusion based on the level of integration. Traditional methods concentrated on pixel-level fusion, mainly

for high-resolution images, but struggled with low-resolution scenarios and relied heavily on accurate image alignment. Decision-level fusion, while operationally efficient and fast, frequently results in substantial detail loss, reducing system accuracy. Feature-level fusion involves directly combining extracted image features, eliminating registration requirements [12]. This method facilitates the extraction of diverse features based on image characteristics, thereby enhancing the capability to describe image content. Additionally, feature extraction significantly minimizes data input for object recognition networks, thereby boosting efficiency. These methods enable the integration of features from both visual and infrared modalities, effectively reducing redundancy [13]. Due to their substantial impact on real-time performance and accuracy in object recognition networks, feature-level fusion techniques are garnering significant global scholarly interest.

Addressing the challenge of effectively forming and integrating complementary information from visible light and infrared modalities, Hwang et al. [14] introduced a multispectral feature fusion method (ACF + T + THOG). However, their reliance on traditional feature extraction methods did not yield significant enhancements in object recognition. Alternatively, Feichtenhofer et al. [15] presented a method for fusing images at various stages of the deep learning architecture. Similarly, Wagner et al. [16] formulated a dual-branch network structure within the CNN object detection framework, examining accuracy enhancements by comparing early and late-stage network fusion. Their empirical studies indicated a preference for late-stage fusion. Subsequently, research utilizing CNNs to amalgamate visible light and infrared data for improved detection capabilities has progressively gained prominence [17–19].

Fang et al. [20] used an attention-based approach to deduce attention maps from both common and differentiated modalities. These maps are used to intelligently improve certain areas of the input feature maps. Xue et al. [21] employed advanced characteristics to guide lower-level features in acquiring channel attention information, which in turn informs the enhanced features for spatial attention filtering. Zhou et al. [22] used a perception-aware feature alignment module to select complementary features retrieved from a single scale. Liu et al. [23] introduced a multimodal feature learning module to explore the correlation between visible light and infrared images at a single scale. Fang et al. [24] incorporated a transformer to acquire comprehensive contextual information, using self-attention methods to amalgamate information from various modalities.

Despite the effective object detection achieved by contemporary approaches utilizing multispectral feature fusion, they nevertheless encounter significant challenges: (1) The majority of contemporary methods frequently construct complex fusion modules subsequent to a dual-channel feature extraction network, thereby increasing computational demands. (2) Attention-based approaches predominantly extract features at a single scale, employing rudimentary fusion techniques, such as concatenation or addition, to either enhance or reduce these features. This method often overlooks the potential for cross-modal information complementarity across various levels. (3) Although the decoupled head provides separate feature information for classification and localization tasks, it is applied to the same input features. This can lead to incorrect or missed detections of small-sized objects. To address these challenges, a novel dual-channel network architecture is proposed in the present work. Extensive evaluations have been conducted on the DroneVehicle [23], KAIST [24], and LLVIP [25] datasets. The results demonstrate that the method achieves high average precision in model performance.

Contributions of this work can be summarized as:

(1) We develop GMD-YOLO for light-infrared object detection, employing the Ghost module to optimize the dual-channel CSPDarknet53 network. This approach is appropriate for object detection tasks in low-light conditions and requires fewer parameters and less computation.
(2) We propose a multilevel feature fusion module to integrate the different levels of visible-infrared information within the network. This module adopts a top-down,

global-to-local approach, enhancing the representation of multiscale features through the construction of hierarchical residual connections.

(3) We design a novel dual-feature modulation decoupling head, replacing the original coupled head. Generating feature encodings with specific semantic contexts resolves the conflict between classification and localization tasks, thereby increasing the accuracy of small object detection.

(4) Experimental results demonstrated that GMD-YOLO surpasses existing advanced methods and exhibits strong robustness. On the KAIST dataset, the FPS value reached 61.7, surpassing other methods.

## 2. Related Work

### *2.1. Traditional Object Detection Algorithms*

To streamline object detection, traditional approaches primarily analyze visible light images under optimal conditions. Early traditional methodologies in this domain typically comprised three stages: image preprocessing, feature extraction, and classifier recognition. Owing to the reliance on manual feature extraction, these algorithms were limited in terms of recognition efficiency and accuracy, rendering them impractical for real-world applications. This necessitated a shift towards deep learning-based algorithms. The majority of deep learning-based object detection algorithms employ CNN, which is categorized into two types. The first category, exemplified by R-CNN [25] and Fast R-CNN [26], encompasses two-stage object detection algorithms [27]. The second category includes single-stage algorithms, illustrated by the SSD [2] series and the YOLO [3,28,29] series. Single-stage algorithms surpass two-stage algorithms in real-time processing, as they directly extract image features, conduct classification and regression in one step, and eliminate the need for generating candidate region boxes. Among them, the YOLO series of models have the advantage of real-time and accuracy and have long attracted much attention. YOLOv5 demonstrates enhanced accuracy and real-time performance, which is attributed to its improved network architecture and training methods, rendering it more apt for engineering applications.

The YOLOv5 network design consists of four main components. The Input module is mainly responsible for performing data preparation operations. The primary focus of the Backbone is on extracting features. The Neck component of the model utilizes the FPN+PAN design from the PANet. Its purpose is to transmit high-level semantic information, hence improving the ability of the model to handle variations in item size and spatial representation. The Head is devoted to predicting the type and location of the object, leveraging the features extracted in preceding stages.

### *2.2. Multispectral Object Detection Algorithm*

Existing algorithms for object detection primarily depend on unimodal data captured in the visible light spectrum. However, this approach performs inadequately in low-light conditions, prevalent during nighttime. This limitation diminishes the algorithms' effectiveness in diverse weather conditions. In contrast, infrared imaging distinctly identifies objects in low-light conditions, offering additional details absent in visible-light imaging. Consequently, integrating multispectral features into the object recognition framework is essential to address the limited light resistance of conventional object detection algorithms [10]. Leveraging the inherent complementary attributes of high-dimensional visible and infrared light properties can significantly enhance detection accuracy and the models' generalization capabilities.

Object detection requires efficient fusion of information from both visible and infrared senses. Pioneering research by Wagner et al. [16] explored optimal fusion points for multispectral feature integration. Liu et al. [30] investigated four fusion structures in dual-branch CNNs: early, middle, late, and decision-level with score fusion, determining that middle-stage fusion is superior. Expanding on this, Li et al. [31] utilized a two-stream mid-layer fusion architecture in MSDS-RCNN to refine detection accuracy further. Li et al. [32]

extended this approach with input fusion and an alternative confidence fusion strategy, showing via comparative analysis that mid-layer fusion of multispectral features within the network is the most effective in improving detection performance. Subsequent research focuses primarily on multispectral object detection algorithms at the intermediate layers of the network [17]. It additionally investigates the effectiveness of fusing multimodal features post various convolutional layers, building on the concept of mid-level fusion.

In addition to considering fusion locations, determining which fusion methods to adopt for the optimal utilization of the complementary information of features has become a research focus in recent years. Zhang et al. [33] introduced an attention-based cross-modal interactive approach that utilizes global information to direct the network in understanding the relationships between different modalities. This mechanism translates visible and infrared features into more semantically significant fused characteristics. Zheng et al. [34] employed two distinct gated fusion units to capture and assign weight to the contributions of multimodal data. This method yielded superior detection performance relative to cascade fusion and surpassed Faster R-CNN models in terms of speed. Zhang et al. [35] introduced CFR, a method for cyclic integration and enhancement of individual spectral characteristics. An et al. [36] introduced ECISNet, a network aimed at augmenting the feature extraction capabilities of CIS. Sun et al. [37] employed feature-level and decision-level fusion strategies in their research. They introduced UA-CMDet, a method integrating visible, infrared, and bimodal fusion branches for detection. Yuan et al. [38] developed TSFADet, aiming to mitigate the adverse effects of cross-modal misalignment by equalizing the discrepancies between two modal features. Zhang et al. [39] proposed SuperYOLO, incorporating a cross-modal fusion module to extract supplementary information from the data. Wang et al. [40] introduced RISNet, featuring a mutual information minimization module to reduce redundancy in multimodal features. Bao et al. [41] proposed Dual-YOLO, integrating the D-Fusion module to minimize duplicated fused feature information. Fu et al. [42] developed LRAF-Net, a network that combines long-range connections of extended visible and infrared characteristics to improve detection accuracy.

The aforementioned methods enhance multispectral image feature extraction but are limited by large parameter size, increasing computational burden, and insufficient inter-modal information exchange. This study addresses these limitations by rationally designing multispectral feature extraction and fusion. Employing YOLOv5 as a base, we introduce GMD-YOLO, a novel network that utilizes visible and infrared feature fusion to enhance detection accuracy and resilience to illumination variations.

## 3. Methods

In this paper, we introduce GMD-YOLO, which is designed for nighttime object detection, as illustrated in Figure 1. Initially, the dual-channel CSPDarknet53 network efficiently extracts features from both RGB and infrared images. Subsequently, a multilevel feature fusion module produces three integrated features from the interactions between multimodal feature layers. Finally, these fused features are inputted into a feature pyramid to derive multiscale features. These are then processed through an enhanced dual feature modulation decoupling head to predict object probabilities and locations.

### *3.1. Dual-Channel CSPDarknet53 Network*

Currently, common approaches typically employ a dual-channel feature extraction network, followed by a fusion of these features through some means. Building upon the YOLOv5 architecture, this study incorporates a dual-channel network structure consisting of two identical CSPDarknet53 components. However, a drawback of these methods is the added complexity needed to establish interactive relationships, which leads to an increased computational workload.

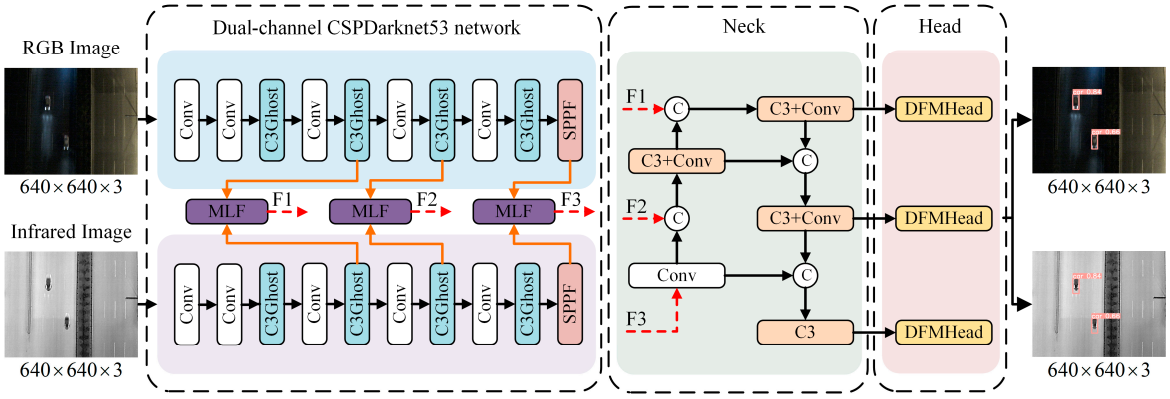


**Figure 1.** Schematic diagram of GMD-YOLO structure. The dual-channel CSPDarknet53 network extracts RGB and IR image features, respectively; the MLF module is the proposed multilevel feature fusion method, the Neck layer fuses different levels of features, and the DFMHead module is an improved detection head for outputting the final detection results.

Inspired by the GhostNet [43] structure, as illustrated in Figure 2, traditional convolution operations generate redundant features in output feature maps, thereby increasing data volume and computational complexity. To tackle this issue and lessen the FLOPs usage in computations, this study introduces the GhostBottleneck module to refine the C3 module in the CSPDarknet53 structure, referred to as C3Ghost. The GhostBottleneck module consists of two layered Ghost modules, with the first acting as an expansion layer to augment channel count. The second Ghost module operates as a contraction layer to align the channel count with the shortcut path, used to link the inputs and outputs of these two Ghost modules, as depicted in Figure 3.

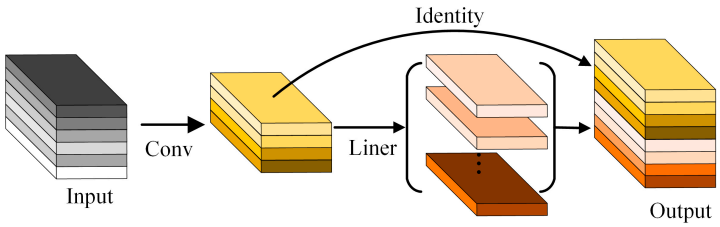


**Figure 2.** Ghost module. Conv with 1 × 1 kernels, Liner is the liner operations.

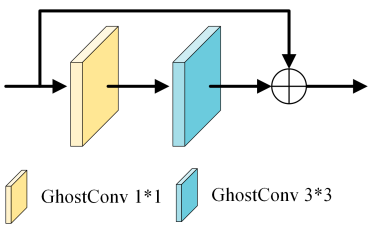


**Figure 3.** GhostBottleneck module.

### *3.2. Multilevel Feature Fusion Module*

In deep learning frameworks, isolated branches often fail to adequately capture the correlation information between different modalities, which to some extent limits the ability of the model to process highly complex tasks [44]. To enhance inter-branch information transmission while curtailing unrestricted noise transfer from specific feature maps to ensuing predictions, a multilevel feature fusion module (MLF) is employed, illustrated in Figure 4.

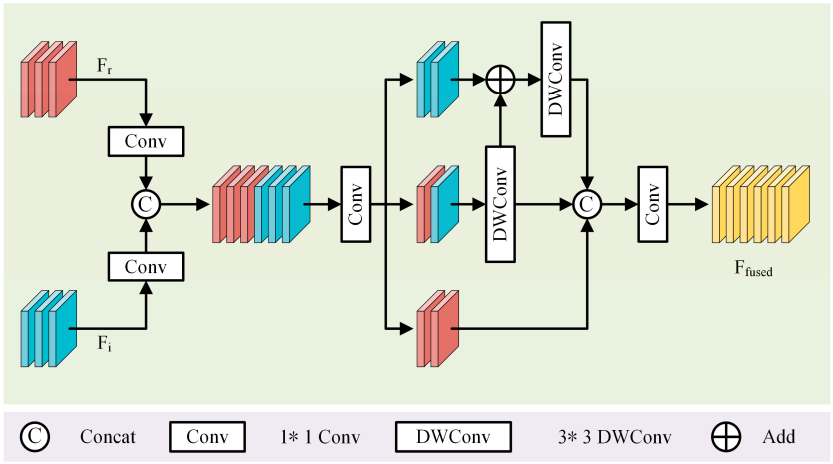


**Figure 4.** Multilevel feature fusion module. This module receives features extracted from both RGB and IR channels as inputs. To capture feature information across various spatial scales, $1 \times 1$ standard convolution is employed, followed by two $3 \times 3$ depth-wise separable convolutions.

The MLF module is utilized to delve into deeper levels of fused feature representation. The MLF module extracts multiscale contextual information with the aim of obtaining a spatial response mapping. This adaptive mapping applies weighted adjustments to the feature mapping at each position. By assigning weights to each pixel, the model increasingly concentrates on the most pertinent sections, rendering it more effective for complex background environments. Within the MLF module, standard convolutions with $1 \times 1$ kernels and depth-wise separable convolutions with $3 \times 3$ kernels are independently used for inter-channel information fusion. This facilitates the network in capturing local or global image features with finer granularity. Furthermore, the smaller parameter size of $1 \times 1$ and $3 \times 3$ convolution kernels ensures that incorporating the MLF module does not markedly escalate the computational cost of the model. The ultimate output is a feature map amalgamated with features from various scales.

Within the backbone network, the multispectral fusion features at down-sampling scales of 8, 16, and 32 layers adeptly merge the high-dimensional properties of both RGB and infrared modalities. This integration, amplified by multiscale convolution processing, yields a superior depiction of prominent object features. Conversely, features acquired at down-sampling scales of 2 and 4 layers predominantly contain low-dimensional information like color and edges, making them less apt for multiscale convolutional operations. This disparity is ascribed to the differing kernel sizes in multiscale convolutions. Larger kernels, equating to a wider receptive field, capture more extensive global information whilst retaining detailed low-level feature representations of the object. As a result, larger kernels may undermine the integrity of these intricate features. Hence, this research refrains from applying the multilevel feature fusion module to the initial two layers, which chiefly encompass low-dimensional semantic information.

### *3.3. Dual Feature Modulation Decoupling Head*

In the domain of object detection, the decoupled head is extensively employed in most one-stage and two-stage object detection algorithms to address classification and localization tasks. However, the original YOLOv5 algorithm employs a coupled head, wherein classification, localization, and confidence predictions are output directly from a convolutional layer following feature fusion. This coupled approach has inherent limitations. Firstly, as the classification and regression tasks share the same feature layer, they may interfere with each other. Secondly, different tasks possess distinct feature requirements, and the coupled approach might not offer optimal feature representations for each, potentially limiting the performance of the model in complex scenarios. The decoupled head is introduced to address these issues. This allows for an independent

feature extraction mechanism for each task, ensuring access to the most appropriate feature representation for each task [45,46]. However, this introduces a potential issue wherein decoupling might restrict the information flow between tasks, potentially diminishing the overall performance of the model. Therefore, this study proposes a dual feature modulation decoupling head, as illustrated in Figure 5.

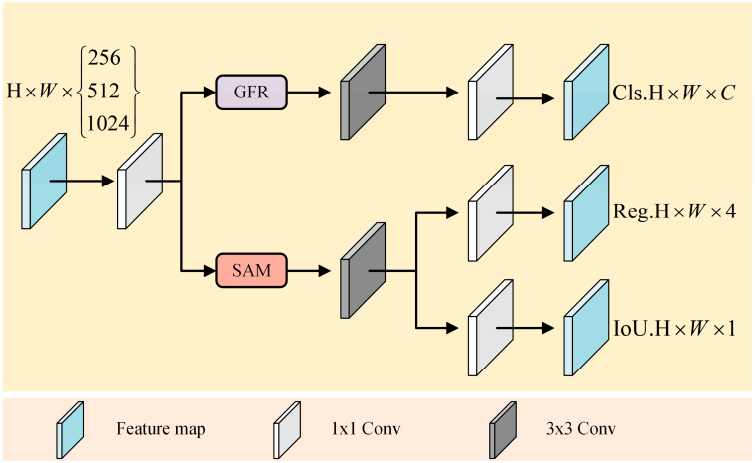


**Figure 5.** Structure of dual feature modulation decoupling head. The GFR module serves as a global feature regulation module tailored for classification challenges, while the SAM module embodies spatial attention specifically introduced for addressing localization issues.

To address the challenges associated with the need for richer contextual semantic information in classification tasks, the global feature regulation (GFR) module is introduced. This module is specifically designed to facilitate the modulation of shallow features using deep ones, enabling the capture of pivotal local image regions; as illustrated in Figure 6 in the preliminary steps leading to the classification task, the GFR harnesses feature maps from two distinct scales, $P_l$ and $P_{l+1}$. The deep feature is subsequently upsampled, followed by a $1 \times 1$ convolution to match the channel count of $P_{l+1}$. These are then concatenated with $P_{l+1}$ to produce the final $G_l^{cls}$:

$$G_l^{cls} = Concat(Conv(Upsample(P_l)), P_{l+1}) \quad (1)$$

where *Concat*, *Conv* and *Upsample* denote concatenation, convolution and upsampling.

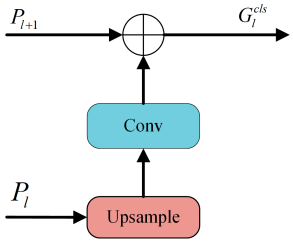


**Figure 6.** Structure of global feature regulation module.

For tasks requiring precise spatial details, it is essential to incorporate a spatial attention mechanism, as illustrated in Figure 7. This approach filters out distractions or background noise in the image and assigns varying degrees of importance to different areas of the feature map, focusing primarily on regions most relevant to the task. In the model discussed, the multiscale feature maps fed into the detection head are used as input feature maps for the spatial attention mechanism module. The spatial attention mechanism (SAM) [47] compresses the channel domain features of the input through global max pooling and global average pooling. It then reduces the multi-channel features to a single

channel via convolution, mitigating the impact of channel-wise information distribution on the spatial attention mechanism. Subsequently, the spatial weight information is normalized through an activation function. Finally, the spatial weight information is element-wise multiplied by the input feature map, resulting in feature maps with varied weights. The comprehensive operation of the spatial attention module can be observed in Equation (2).

$$M_s(F) = \sigma(f^{3\times3}[AvgPool(F); MaxPool(F)]) \quad (2)$$

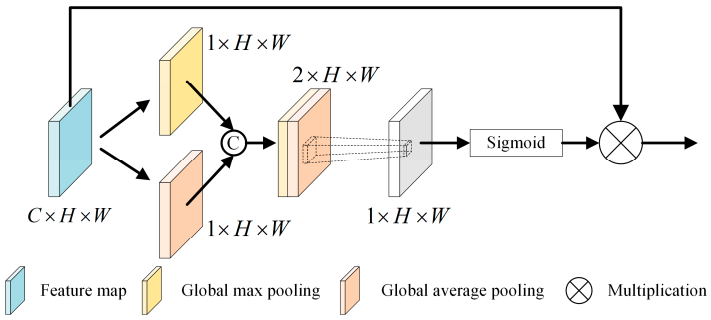


**Figure 7.** SAM structure diagram.

Let $F$ represent the input feature map. The sigmoid activation function is represented by $\sigma$. The convolutional layer employing a kernel of size $3 \times 3$ is denoted by $f$. The pooled feature map is signified by the notation $[AvgPool(F); MaxPool(F)]$, while $M_s(F)$ stands for the spatial attention parameter matrix.

## 4. Experiments

### *4.1. Experimental Environment*

In the experimental setup, we employ a Windows 11 operating system. The model described herein is implemented using the PyTorch framework, specifically tailored for the YOLOv5 multispectral night detection algorithm. We carry out the computations on a system equipped with an Intel i7-12700K CPU and an NVIDIA RTX 3070 GPU.

### *4.2. Datasets and Evaluation Metrics*

The DroneVehicle Dataset [37] is meticulously compiled from images synchronously captured by drones equipped with both visible light and infrared cameras, encompassing a total of 28,439 pairs of visible/infrared images. This dataset is distinguished by its temporal diversity, featuring images taken during various times of the day, including daytime (14,478 pairs), evening (8493 pairs), and nighttime (5468 pairs). Notably, images captured under low-light conditions at night constitute approximately half of the entire dataset. The dataset categorizes objects into several vehicle types: cars, freight cars, trucks, buses, and vans. To address the discrepancy in resolution between the original visible light images ($1024 \times 768$ pixels) and infrared images ($640 \times 512$ pixels), Sun and colleagues have undertaken a series of image processing steps, including scaling, cropping, padding, and alignment, standardizing the image dimensions to $840 \times 712$ pixels.

The KAIST Dataset [48] is a widely recognized multispectral dataset-tailored dataset. It encompasses 95,324 paired visible-infrared images; these images, of dimensions $640 \times 512$, span both day and night scenarios.

The LLVIP Dataset [49] offers high-resolution visible-temperature infrared-paired pedestrian detection. It comprises 16,836 precisely aligned infrared and visible image pairs. For training, it utilizes 12,025 pairs, whereas 3463 pairs are designated for testing. A notable characteristic is the temporal and spatial alignment of these images, ensuring their coherence.

In this study, we employ a suite of performance metrics for object detection assessment, GFLOPS, Precision, *mAP*, and *MR*. Precision quantifies the percentage of correctly predicted

positive instances among all positive predictions. The *mAP* provides an average of precision values across various recall levels. *MR*, particularly pivotal in pedestrian detection tasks, inversely correlates with performance, with lower scores indicating superior performance, as shown in Equations (3)–(6).

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$\begin{cases} AP = \int_0^1 \mathrm{Precision}(\mathrm{Recall}) d\mathrm{Recall} \\ mAP = \frac{\sum_{m=1}^{M} AP(n)}{M} \end{cases} \tag{5}$$

$$FPS = \frac{FigureNumber}{TotalTime} \tag{6}$$

In Equations (3) and (4) *TP* is the positive sample predicted by the model to be in the positive category; FP is the negative sample predicted by the model to be in the positive category; and *FN* is the positive sample predicted by the model to be in the negative category. Samples with the IOU greater than 0.5 are considered positive.

### *4.3. Ablation Experiments*

#### 4.3.1. Impact of the Multilevel Feature Fusion Module

The experimental results of the MLF module on the LLVIP dataset are presented in Table 1. The baseline utilizes the original backbone network of YOLOv5 to form a dual-branch network structure. It employs an element-wise addition approach to merge visible light and infrared features. The first two rows of this table clarify that in the LLVIP dataset, the detection efficacy of infrared features surpasses that of RGB features. This superiority arises primarily from the high-definition quality of the dataset. The dataset contains fewer small-sized pedestrian objects and provides more distinct visibility of pedestrians in thermal infrared imagery compared to visible light images.

**Table 1.** Effect of the MLF module. (The best indicator is bold).

| Method | Modal | Precision (%) | mAP@0.5 (%) |
|---|---|---|---|
| Baseline-RGB | RGB | 90.2 | 90.0 |
| Baseline-Thermal | IR | 92.3 | 94.9 |
| Baseline | RGB + IR | 92.9 | 93.6 |
| Baseline + MLF | RGB + IR | **93.2** | **94.0** |

Consequently, implementing the MLF module on both RGB and infrared branches aids in gathering complementary information from each. The thus enhanced fusion features are subsequently employed for further detection tasks. In the context of the LLVIP dataset, this method has shown an improvement of 0.3% and 0.4% in performance over the baseline methods, as shown in the final row of the table.

The intuitive attention maps presented in Figure 8 qualitatively validate the efficacy of the MLF module. Figure 8b illustrates the feature maps extracted by the Backbone, which serve as the input for the feature maps in Figure 8c,d. Upon observing Figure 8c, it becomes apparent that the features of the detected objects are markedly enhanced following single-scale fusion. This enhancement intensifies the features across both modalities, directing the focus of the network to the foreground in the modal feature maps. Comparing Figure 8d to Figure 8c, it is revealed that the use of the MLF module leads to the diminution of several ineffective background features. The scenarios depicted in the first two rows of Figure 8 are especially significant, where nearby vehicles could potentially confuse the detector if the model overemphasizes these features. Figure 8d shows that the MLF module aids in

reducing the interference of redundant features, thereby enabling the model to focus more sharply on pedestrian features.

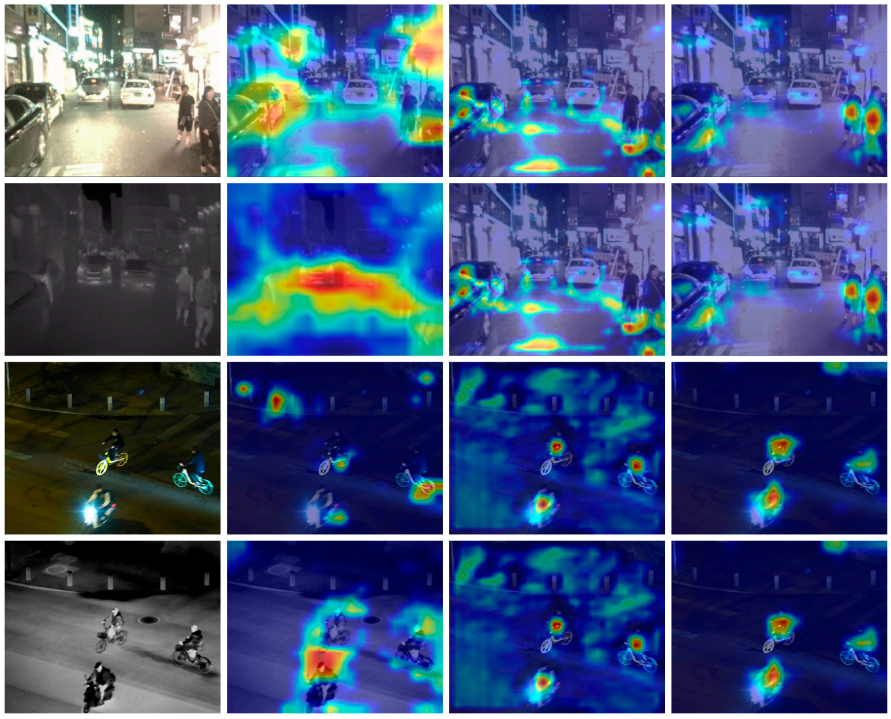


**Figure 8.** Feature visualization results. From (**a**–**d**) columns: original RGB and IR images, feature maps extracted from the backbone, heat maps from the single-scale fusion method, and the MLF fusion method.

#### 4.3.2. Impact of Improved Detection Head

To meet the preferences of different tasks for coarse and fine-grained features, a dual feature modulation decoupling head is proposed. In nighttime detection tasks, the absence of ambient light renders object features less distinct than in daylight, necessitating reliance on contextual clues from the surroundings for recognition. Therefore, spatial information becomes crucial for differentiating objects from the background and comprehending inter-object relationships. An experimental evaluation is conducted to assess the performance of the decoupled head using different attention mechanisms and their impact on model performance. In these experiments, the model parameters remain constant, and only the type of attention mechanism varies. This allows for an analysis of how different scenarios affect the detection performance of the model.

Table 2 shows that the SAM (Spatial Attention Mechanism) demonstrates superior performance in the decoupled head localization task compared to other attention mechanisms. Although the mAP@0.5 for SAM (Spatial Attention Mechanism) decreased by 0.7% compared to GAM (Global Attention Mechanism), SAM has the advantage of introducing fewer parameters and achieving a higher Precision of 93.5%. Despite both CBAM (Convolutional Block Attention Module) and GAM addressing channel and spatial attention, their efficacy in spatial attention is not as pronounced as SAM. Conversely, ECA (Efficient Channel Attention) and CA (Coordinate Attention) mechanisms focus mainly on channel information, leading to shortcomings in capturing spatial details. Ultimately, the approach that incorporates SAM achieves optimal performance, yielding a precision of 94% and mAP@0.5 of 95.5%.

**Table 2.** Effect of decoupled head with different attention mechanisms. (Lower GFLOPs are better, higher Precision is better, higher mAP@0.5 is better. The best indicator is bold).

| Method | GFLOPs | Precision (%) | mAP@0.5 (%) |
|---|---|---|---|
| SAM | 46.0 | **94.0** | 95.5 |
| CBAM [47] | 46.2 | 93.1 | 95.3 |
| GAM [50] | 50.2 | 93.5 | **96.2** |
| CA [51] | 46.1 | 92.8 | 95.1 |
| ECA [52] | 46.0 | 93.2 | 95.0 |

#### 4.3.3. Impact of Different Modules

To validate that the inclusion of each module in the proposed method enhances model performance, ablation experiments on the LLVIP dataset were performed. In addition, identical parameters are assigned to each variable to ensure the fairness of the assessment. The experimental results are displayed in Table 3, where a '√' denotes the inclusion of modules for improvement.

**Table 3.** The ablation experiments for GMD-YOLO. (The best indicator is bold).

| Method | Ghost | MLF | DFMHead | GFLOPs | Precision (%) | mAP@0.5 (%) |
|---|---|---|---|---|---|---|
| Baseline | | | | 32.3 | 92.9 | 93.6 |
| 1 | √ | | | 26.4 | 92.4 | 92.9 |
| 2 | | √ | | 29.7 | 93.2 | 94.0 |
| 3 | | | √ | 46.0 | 94.0 | 95.5 |
| 4 | √ | √ | | **22.9** | 93.4 | 95.9 |
| 5 | √ | | √ | 39.2 | 94.2 | 96.8 |
| 6 | | √ | √ | 42.5 | 95.2 | 96.7 |
| 7 | √ | √ | √ | 35.7 | **96.0** | **98.0** |

In this study, the baseline model utilizes the original YOLOv5 network, with CSPDarknet53 as the primary feature extraction backbone, attaining a mAP of 93.6%. A significant reduction in network parameters was noted when the Ghost module partially supplanted the C3 modules. This alteration, known as Method 1, reduced the GFLOPs from 32.3 to 26.4. Method 2 incorporates a novel multilevel feature fusion (MLF) module, ensuring the preservation of essential feature information while enabling the interaction of intrinsic modal information, resulting in an improved mAP of 94.0%. Method 3 substitutes the coupled head with an enhanced dual feature modulation decoupling head (DFMHead), resulting in a 1.9% increase in the mAP of the network compared to the original configuration. Methods 3, 4, and 5 illustrate that a hybrid approach, integrating two types of enhancements, results in superior mAP improvements compared to employing a single strategy. The simultaneous application of all three enhancements—Ghost, MLF, and DFMHead—led to the creation of the optimal model. This model demonstrated an exceptional precision rate of 96.0%, a 3.1% increment over the baseline, and a 4.4% rise in mAP. As a result, this study adopts the integration of Ghost, MLF, and DFMHead into the model to attain the most efficient performance outcomes.

### *4.4. Comparison Experiment*

#### 4.4.1. Experiments on the DroneVehicle Dataset

In this study, the efficacy of the proposed GMD-YOLO model is evaluated on the DroneVehicle dataset, benchmarking it against representative unimodal and multispectral object detection algorithms from recent years. The evaluation results, as presented in Table 4, demonstrate significant improvements achieved by GMD-YOLO.

**Table 4.** Comparison of experimental results for the DroneVehicle. (The best indicator is bold).

| Method | Modal | Backbone | Precision (%) | mAP@0.5 (%) |
|---|---|---|---|---|
| YOLOv5 | RGB | CSPDarknet53 | 59.5 | 64.8 |
| YOLOv5 | IR | CSPDarknet53 | 70.1 | 75.9 |
| YOLOv5 | RGB + IR | CSPDarknet53 | 74.1 | 74.4 |
| CFR [35] | RGB + IR | ResNet | - | 73.9 |
| UA-CMDet [37] | RGB + IR | YOLO | - | 64.0 |
| ECISNet [36] | RGB + IR | ResNet | - | 76.0 |
| TSFADet [38] | RGB + IR | ResNet | - | 73.0 |
| GMD-YOLO | RGB + IR | YOLO | **80.3** | **78.0** |

Specifically, with the use of either visible light or infrared unimodal imagery, GMD-YOLO demonstrates a 13.2% and 2.1% enhancement in the mAP@0.5 metric, respectively, compared to YOLOv5. In scenarios utilizing fused visible and infrared imagery, GMD-YOLO surpasses YOLOv5, CFR, UA-CMDet, ECISNet, and TSFADet in the mAP@0.5 metric by margins of 3.6%, 4.1%, 14%, 2%, and 5%, respectively. This improvement is attributed to GMD-YOLO's innovative approach to capturing rich contextual information within images. Contrary to existing bimodal object detection algorithms that mainly concentrate on enhancing or suppressing input image features through weighted mechanisms, often overlooking the exploration of dependencies across various levels of cross-modal information, GMD-YOLO effectively amalgamates features of objects at multiple scales. This integration, utilizing the synergistic relationship between modalities, considerably improves the detection of smaller-scale objects.

#### 4.4.2. Experiments on the KAIST Dataset

To rigorously evaluate the performance of GMD-YOLO, this section juxtaposes the model against a cohort of contemporary multispectral object detection methodologies that have emerged in recent years. These comparisons are conducted using the KAIST dataset.

The evaluative metrics, presented in Table 5, reveal that GMD-YOLO attains a logarithmic average miss rate (MR) of 7.73%, a figure that notably exceeds that of the nearest competitor, MBNet, by a margin of 0.13%. This statistical superiority, especially in nighttime environments, signifies the robust performance offered by the GMD-YOLO framework. The improvement in detection accuracy can be attributed to the integration of a dual feature modulation decoupling head and an MLF module within GMD-YOLO. The integrated components significantly improve the capability of the model to extract features from target objects while reducing the incidence of false positives.

**Table 5.** Comparison of experimental results for the KAIST. (The best indicator is bold).

| Method | MR (%) | FPS (Hz) | Platform |
|---|---|---|---|
| ACF + T + THOG [14] | 56.17 | - | - |
| Halfway Fusion [30] | 26.67 | 2.33 | TITAN X |
| Fusion RPN + BN [17] | 16.27 | - | - |
| MSDS-RCNN [31] | 13.73 | 4.55 | GTX 1080Ti |
| IAF-RCNN [32] | 16.70 | 4.76 | TITAN X |
| CIAN [33] | 11.13 | 16.67 | GTX 1080Ti |
| AR-CNN [53] | 9.02 | 8.33 | TITAN X |
| MBNet [22] | 7.86 | 14.29 | GTX 1080Ti |
| GMD-YOLO | **7.73** | **61.7** | GTX 3070 |

#### 4.4.3. Experiments on the LLVIP Dataset

In practical applications, pedestrian data is characterized by its richness and diversity. To ascertain the robustness of the methodology proposed in this article across different pedestrian datasets, identical experimental validations are performed on the LLVIP multimodal dataset. The detection outcomes and comparisons are presented in Table 6.

**Table 6.** Comparison of experimental results for the LLVIP. (The best indicator is bold).

| Method | Modal | Backbone | Precision (%) | mAP@0.5 (%) |
|---|---|---|---|---|
| YOLOv3 | RGB | Darknet53 | 90.7 | 81.2 |
| YOLOv3 | IR | Darknet53 | 91.4 | 88.8 |
| YOLOv5 | RGB | CSPDarknet53 | 90.2 | 90.0 |
| YOLOv5 | IR | CSPDarknet53 | 92.3 | 94.9 |
| YOLOv5 | RGB + IR | CSPDarknet53 | 92.9 | 93.6 |
| CFT [24] | RGB + IR | CFB | - | 97.5 |
| SuperYOLO [39] | RGB + IR | YOLO | - | 96.7 |
| GMD-YOLO | RGB + IR | YOLO | **96.0** | **98.0** |

As evident from Table 6, within the LLVIP dataset, the proposed method achieves the mAP@0.5 of 98.0%. In the unimodal modality, the mAP@0.5 obtained with YOLOv3 and YOLOv5 on the visible light subset of the LLVIP dataset is markedly lower than that of the method. This observation intuitively suggests that the approach surpasses unimodal visible light-based detection methods in terms of efficacy and performance. In the infrared subset, the mAP@0.5 of the method significantly exceeds that achieved with YOLOv3 and YOLOv5 methods, and it also surpasses the mAP@0.5 in the visible light subset. This improvement is attributed to the high-definition quality of the dataset, characterized by a reduced proportion of small-sized pedestrian objects and enhanced visibility of pedestrian objects in thermal infrared imagery as opposed to visible light imagery.

In terms of multispectral detection, the single-scale feature-level fusion in YOLOv5, while extracting features from multispectral images, overlooks the retention of detailed information, leading to an mAP@0.5 of only 93.6%, lower than the algorithm. Compared to advanced methods like SuperYOLO and CFT, the method still exhibits an increase of 1.3% and 0.5% in mAP@0.5, respectively. This result substantiates the considerable enhancement in accuracy and robust generalizability of the algorithm on this dataset.

### *4.5. Qualitative Analysis*

In order to facilitate a more intuitive comparison, Figure 9 illustrates the visualized results of the Baseline and GMD-YOLO on the DroneVehicle dataset. The Baseline model exhibits instances of false detections, evidenced by the misidentification of freight cars as trucks in the first and second images. Furthermore, the Baseline model demonstrates partial omissions in detection, as seen in the third and fourth images where cars were not detected. Significantly, GMD-YOLO shows exceptional proficiency in detecting smaller-sized objects in complex scenes. This superior detection capability is due to the ability of GMD-YOLO to extract deeper, integrated feature representations. The enhanced decoupled head in GMD-YOLO affords task-specific feature preferences, effectively mitigating the effects of substantial variations in object sizes.

In the experiments, the GMD-YOLO framework also demonstrates high potential in detecting obscured objects and in scenarios with blurred object representations. To demonstrate the performance of the method under conditions of obstruction and object blurriness, researchers carefully select and visualize results from the KAIST and LLVIP datasets. These results are subsequently compared with those obtained using the Baseline method.

Figure 10 showcases nighttime scenes where the baseline method, while capable of identifying pedestrian objects in columns a and b, misses detection in cases of obstructed pedestrians. Conversely, the algorithm presented in this study successfully detects and recognizes the objects. In columns c and d, featuring visible light and infrared images, the Baseline method fails to detect distant, blurry pedestrians, likely due to the complexity of the scene and the dense presence of pedestrians. However, the algorithm in this study effectively utilizes visible light and infrared feature information, accurately detecting and recognizing pedestrian objects.

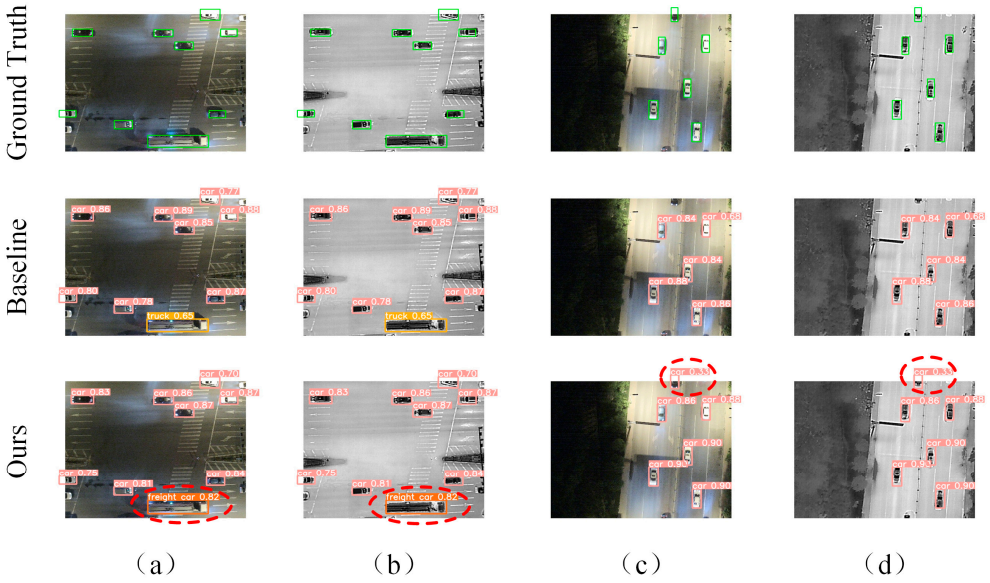


**Figure 9.** Detection results on the DroneVehicle. (**a**,**b**) are RGB and IR images of a false detection scenario. (**c**,**d**) are RGB and IR images of a missed detection scenario. Objects encircled in red denote instances where the baseline algorithm either misses or falsely detects objects, whereas the GMD-YOLO algorithm accurately identifies them.

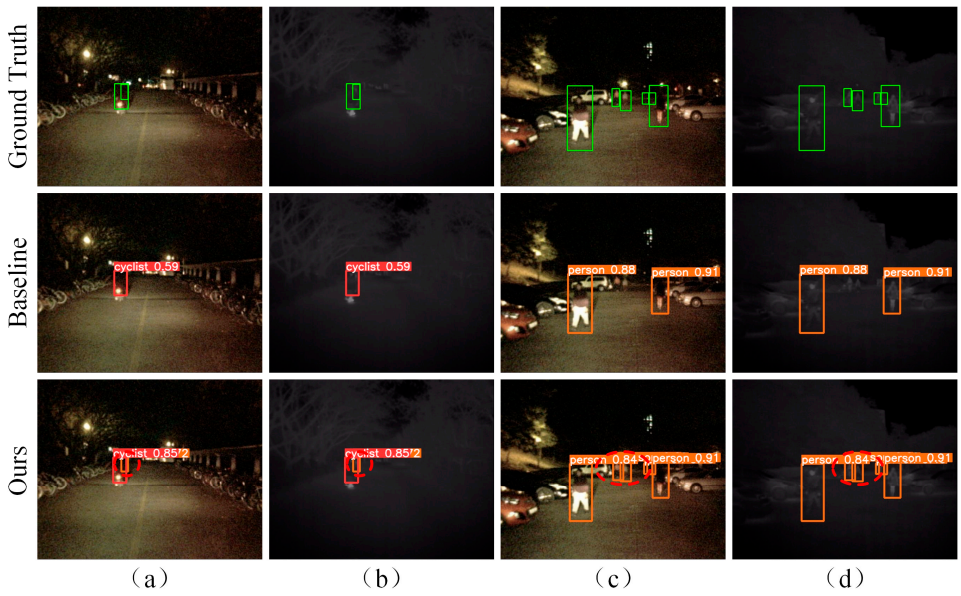


**Figure 10.** Detection results on the KAIST. (**a**,**b**) are RGB and IR images of an occluded detection scenario. (**c**,**d**) are RGB and IR images of a distant detection scenario. Objects encircled in red denote instances where the baseline algorithm misses, whereas the GMD-YOLO algorithm accurately identifies them.

In Figures 11 and 12, set in nocturnal scenes, the Baseline algorithm consistently fails to detect pedestrians obscured by obstacles, whereas GMD-YOLO effectively concentrates on object features, accurately identifying pedestrian objects in the images. In Figure 11a,b, due to the Baseline lack of multiscale information, false detections occur. In Figure 11c,d, the pedestrians are situated in low-light environments with blurred objects. Due to the use of single-scale convolutions in the Baseline, noise from some feature maps is transmitted to subsequent predictions in the network, resulting in missed detections.

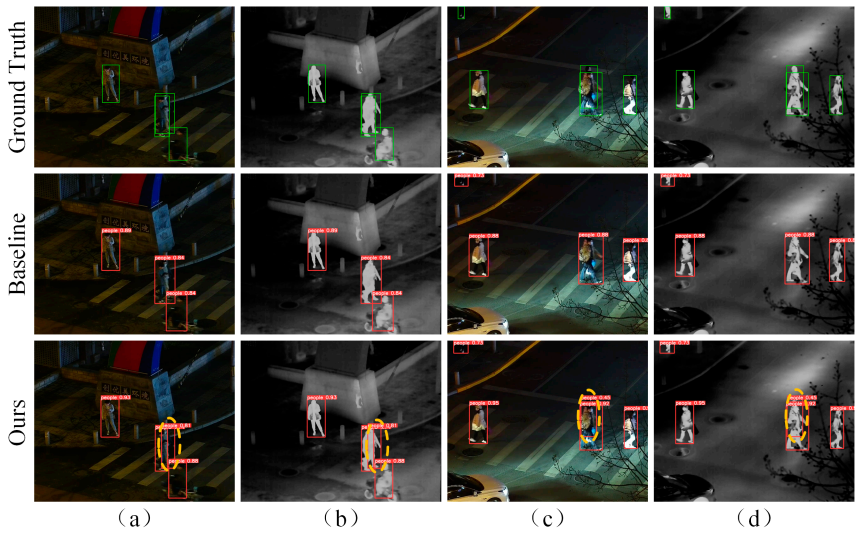


**Figure 11.** Detection results on the LLVIP. (**a**,**b**) are RGB and IR images of an occluded detection scenario. (**c**,**d**) are RGB and IR images of a missed detection scenario. Objects encircled in yellow denote instances where the baseline algorithm missed, whereas the GMD-YOLO algorithm accurately identified them.

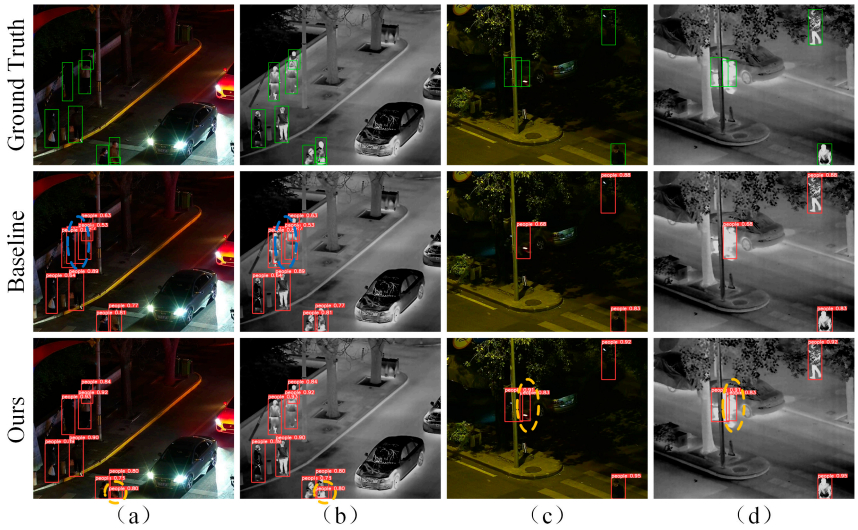


**Figure 12.** Detection results on the LLVIP. (**a**,**b**) are RGB and IR images of a false detection scenario. (**c**,**d**) are RGB and IR images of a darker detection scenario. Objects encircled in yellow denote instances where the baseline algorithm missed, whereas the GMD-YOLO algorithm accurately identified them. Objects encircled in blue denote instances where the baseline algorithm has misdetections.

Results from Figures 9–12 suggest that the algorithm proposed in this study outperforms the baseline algorithms, accurately detecting objects under various conditions. Addressing the issues of false positives and missed detections common in baseline algorithms, the method demonstrates significant improvement. It effectively locates object positions even in adverse conditions such as poor lighting, severe obstruction, or object blurring. Moreover, the size of the object box is appropriate and accurately corresponds to the actual dimensions of the object.

## 5. Conclusions

In this study, a multispectral object detection network based on multilevel feature fusion and dual feature modulation (GMD-YOLO) is proposed. This algorithm utilizes

a novel yet structurally simple backbone network to extract and integrate multispectral features. It employs a specially designed MLF module to enhance the interaction of information across diverse modalities. The algorithm is tailored with a dual feature modulation decoupling head to address classification and regression problems, providing optimal feature representation for each task. Experimental results show that GMD-YOLO surpasses current advanced methods in detection performance on the DroneVehicle dataset, achieving an mAP@0.5 of 78.0%. It also displays enhanced detection performance on the KAIST dataset, with an MR improvement to 7.73%. On the LLVIP dataset, GMD-YOLO achieves notable accuracy enhancements compared to the most advanced multispectral models, indicating its exceptional generalizability. Furthermore, owing to its efficient use of visible light and infrared information, GMD-YOLO shows robustness against sudden changes in light intensity, maintaining superior detection performance in scenarios challenging for human vision.

In the future, our goal is to develop more universal and lightweight models. While leveraging the benefits of both, our intention is to minimize the consumption of excessive resources. Moreover, we plan to implement these streamlined models on edge computing platforms for real-time multispectral object detection.

**Author Contributions:** Conceptualization, J.S. and M.Y.; methodology, J.S.; software, J.S.; validation, J.S., T.X. and Z.W.; formal analysis, M.Y.; investigation, J.S.; resources, S.B.; data curation, J.S.; writing—original draft preparation, J.S.; writing—review and editing, J.S.; visualization, J.S.; supervision, M.Y.; project administration, M.Y.; funding acquisition, S.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by the National Natural Science Foundation of China under Grant 62103192 and Grant 52172367, Natural Science Research Project of Colleges and Universities of Jiangsu Province (20KJB520015), Changzhou Applied Basic Research Project (Medium subsidy) (CJ20200039).

**Data Availability Statement:** All data underlying the results are available as part of the article and no additional source data are required.

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

## References

1. Zhao, Z.; Zheng, P.; Xu, S.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef] [PubMed]
2. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
3. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
4. Singh, A.; Bhambhu, Y.; Buckchash, H.; Gupta, D.K.; Prasad, D.K. Latent Graph Attention for Enhanced Spatial Context. *arXiv* **2023**, arXiv:2307.04149.
5. Biswas, M.; Buckchash, H.; Prasad, D.K. pNNCLR: Stochastic Pseudo Neighborhoods for Contrastive Learning based Unsupervised Representation Learning Problems. *arXiv* **2023**, arXiv:2308.06983.
6. Gu, J.; Su, T.; Wang, Q.; Du, X.; Guizani, M. Multiple Moving Targets Surveillance Based on a Cooperative Network for Multi-UAV. *IEEE Commun. Mag.* **2018**, *56*, 82–89. [CrossRef]
7. Kim, J.H.; Batchuluun, G.; Park, K.R. Pedestrian detection based on faster R-CNN in nighttime by fusing deep convolutional features of successive images. *Expert Syst. Appl.* **2018**, *114*, 15–33. [CrossRef]
8. Zou, T.; Yang, S.; Zhang, Y.; Ye, M. Attention guided neural network models for occluded pedestrian detection. *Pattern Recognit. Lett.* **2020**, *131*, 91–97. [CrossRef]
9. He, X.; Chen, Z.; Dai, L.; Liang, L.; Wu, J.; Sheng, B. Global-and-local aware network for low-light image enhancement. *Eng. Appl. Artif. Intell.* **2023**, *126*, 106969. [CrossRef]
10. Zheng, A.; Ye, N.; Li, C.; Wang, X.; Tang, J. Multi-modal foreground detection via inter-and intra-modality-consistent low-rank separation. *Neurocomputing* **2020**, *371*, 27–38. [CrossRef]
11. Zhang, X.; Zhang, Y.; Guo, Z.; Zhao, J.; Tong, X. Advances and perspective on motion detection fusion in visual and thermal framework. *J. Infrared Millim. Waves* **2011**, *30*, 354–359. [CrossRef]
12. Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; Tian, Q. Deep modular co-attention networks for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6281–6290.

13. Wanchaitanawong, N.; Tanaka, M.; Shibata, T.; Okutomi, M. Multi-Modal Pedestrian Detection with Large Misalignment Based on Modal-Wise Regression and Multi-Modal IoU. In Proceedings of the 2021 17th International Conference on Machine Vision and Applications (MVA), Virtual, 25–27 July 2021; pp. 1–6.
14. 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.
15. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.
16. 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, Bruges, Belgium, 5–7 October 2016; pp. 509–514.
17. Konig, D.; Adam, M.; Jarvers, C.; Layher, G.; Neumann, H.; Teutsch, M. Fully convolutional region proposal networks for multispectral person detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 49–56.
18. Park, K.; Kim, S.; Sohn, K. Unified multi-spectral pedestrian detection based on probabilistic fusion networks. *Pattern Recognit.* **2018**, *80*, 143–155. [CrossRef]
19. Sharma, M.; Dhanaraj, M.; Karnam, S.; Chachlakis, D.G.; Ptucha, R.; Markopoulos, P.P.; Saber, E. YOLOrs: Object detection in multimodal remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2020**, *14*, 1497–1508. [CrossRef]
20. Fang, Q.; Wang, Z. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognit.* **2022**, *130*, 108786. [CrossRef]
21. Xue, Y.; Ju, Z.; Li, Y.; Zhang, W. MAF-YOLO: Multi-modal attention fusion based YOLO for pedestrian detection. *Infrared Phys. Technol.* **2021**, *118*, 103906. [CrossRef]
22. Zhou, K.; Chen, L.; Cao, X. Improving multispectral pedestrian detection by addressing modality imbalance problems. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 787–803.
23. Liu, T.; Lam, K.; Zhao, R.; Qiu, G. Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 315–329. [CrossRef]
24. Fang, Q.; Han, D.; Wang, Z. Cross-modality fusion transformer for multispectral object detection. *arXiv* **2021**, arXiv:2111.00273. [CrossRef]
25. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
26. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
27. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [CrossRef]
28. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
29. Bochkovskiy, A.; Wang, C.; Liao, H.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
30. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral deep neural networks for pedestrian detection. *arXiv* **2016**, arXiv:1611.02644.
31. Li, C.; Song, D.; Tong, R.; Tang, M. Multispectral pedestrian detection via simultaneous detection and segmentation. *arXiv* **2018**, arXiv:1808.04818.
32. 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]
33. 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]
34. Zheng, Y.; Izzat, I.H.; Ziaee, S. GFD-SSD: Gated fusion double SSD for multispectral pedestrian detection. *arXiv* **2019**, arXiv:1903.06999.
35. Zhang, H.; Fromont, E.; Lefevre, S.; Avignon, B. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 276–280.
36. An, Z.; Liu, C.; Han, Y. Effectiveness Guided Cross-Modal Information Sharing for Aligned RGB-T Object Detection. *IEEE Signal Process. Lett.* **2022**, *29*, 2562–2566. [CrossRef]
37. Sun, Y.; Cao, B.; Zhu, P.; Hu, Q. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6700–6713. [CrossRef]
38. Yuan, M.; Wang, Y.; Wei, X. Translation, Scale and Rotation: Cross-Modal Alignment Meets RGB-Infrared Vehicle Detection. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 509–525.
39. Zhang, J.; Lei, J.; Xie, W.; Fang, Z.; Li, Y.; Du, Q. SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5605415. [CrossRef]

40. Wang, Q.; Chi, Y.; Shen, T.; Song, J.; Zhang, Z.; Zhu, Y. Improving RGB-infrared object detection by reducing cross-modality redundancy. *Remote Sens.* **2022**, *14*, 2020. [CrossRef]
41. Bao, C.; Cao, J.; Hao, Q.; Cheng, Y.; Ning, Y.; Zhao, T. Dual-YOLO Architecture from Infrared and Visible Images for Object Detection. *Sensors* **2023**, *23*, 2934. [CrossRef]
42. 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**, 1–14. [CrossRef]
43. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.
44. You, S.; Xie, X.; Feng, Y.; Mei, C.; Ji, Y. Multi-Scale Aggregation Transformers for Multispectral Object Detection. *IEEE Signal Process. Lett.* **2023**, *30*, 1172–1176. [CrossRef]
45. Chen, Z.; Yang, C.; Li, Q.; Zhao, F.; Zha, Z.; Wu, F. Disentangle your dense object detector. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 4939–4948.
46. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*.
47. Woo, S.; Park, J.; Lee, J.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
48. Choi, Y.; Kim, N.; Hwang, S.; Park, K.; Yoon, J.S.; An, K.; Kweon, I.S. KAIST multi-spectral day/night data set for autonomous and assisted driving. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 934–948. [CrossRef]
49. Jia, X.; Zhu, C.; Li, M.; Tang, W.; Zhou, W. LLVIP: A visible-infrared paired dataset for low-light vision. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3496–3504.
50. Liu, Y.; Shao, Z.; Hoffmann, N. Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv* **2021**, arXiv:2112.05561.
51. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 13713–13722.
52. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
53. Zhang, L.; Zhu, X.; Chen, X.; Yang, X.; Lei, Z.; Liu, Z. Weakly aligned cross-modal learning for multispectral pedestrian detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 16–19 June 2019; pp. 5127–5137.


**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.