Multispectral Object Detection Enhanced by Cross-modal Information Complementary and Cosine Similarity Channel Resampling Modules

Junbo Jang Chanyeong Park Heegwang Kim Jiyoon Lee Joonki Paik

Chung-Ang University, Republic of Korea

Abstract

Images obtained from different modalities can effectively enhance the accuracy and reliability of the detection model by complementing specialized information from visible (RGB) and infrared (IR) images. However, integrating information from multiple modalities faces the following challenges: 1) distinct characteristics of RGB and IR images lead to the problem of modality imbalance, 2) fusing multimodal information can greatly affect the detection accuracy, as some of the unique information provided by each modality is lost during the integration process, and 3) RGB and IR images are fused while preserving the noise of each modality. To address these issues, we propose a novel multispectral object detection network which contains two main components; 1) Cross-modal Information Complementary (CIC) module, and 2) Cosine Similarity Channel Resampling (CSCR) module. The proposed method addresses the modality imbalance problem and efficiently fuses RGB and IR images in the feature level. Extensive experimental results on LLVIP, FLIR, M^3FD, VEDAI and KAIST benchmark datasets, verify the effectiveness and generalization performance of the proposed multispectral object detection network compared with other state-of-the-art methods.

1. Introduction

Multispectral object detection in autonomous driving faces challenges in low-light conditions, where glare affects RGB images. Fusing RGB and infrared (IR) images improves detection by combining their strengths: RGB captures detailed textures in good lighting, while IR excels in low light and occlusion, providing clear outlines despite lacking texture and color.

The challenges associated with modality imbalances in multispectral detection datasets are twofold: illumination modality imbalance and feature modality imbalance [24, 28]. Illumination modality imbalance highlights the variance in lighting conditions between daytime and nighttime images. Feature modality imbalance arises from misalignment and ineffective integration of different modalities, leading to disproportionate feature representation. This imbalance complicates the fusion of RGB and IR feature maps, potentially diminishing essential features or overly highlighting less significant ones.

In this paper, we present a novel multispectral fusion module, the Multispectral Complementary Object-weighted Resampling (MCOR) module, which efficiently combines features through the Cross-modal Information Complementary (CIC) module and the Cosine Similarity Channel Resampling (CSCR) module to address the aforementioned challenges. The CIC module leverages unique information from each modality, distinguishing them effectively to enhance multispectral object detection. The CSCR module emphasizes object-focused feature maps and fuses modalities efficiently by using separate object- and background-weighted maps, adding supplementary information where needed. The proposed multispectral fusion module is demonstrated to be highly efficient, significantly outperforming recently proposed multispectral detection models. Experimental evaluations conducted on FLIR [6], LLVIP [13], M^3FD [25], VEDAI [32] and KAIST [12] datasets demonstrate substantial improvements in detection performance achieved by our proposed module.

Our main contribution can be summarized as follows:

- We introduce the Multispectral Complementary Object-weighted Resampling (MCOR) module, which combines features effectively using the Cross-modal Information Complementary (CIC) and Cosine Similarity Channel Resampling (CSCR) modules to overcome multispectral fusion challenges.

- We present both qualitative and quantitative assessments across multiple benchmark multispectral object detection datasets and models. The proposed technique works better than existing methods.

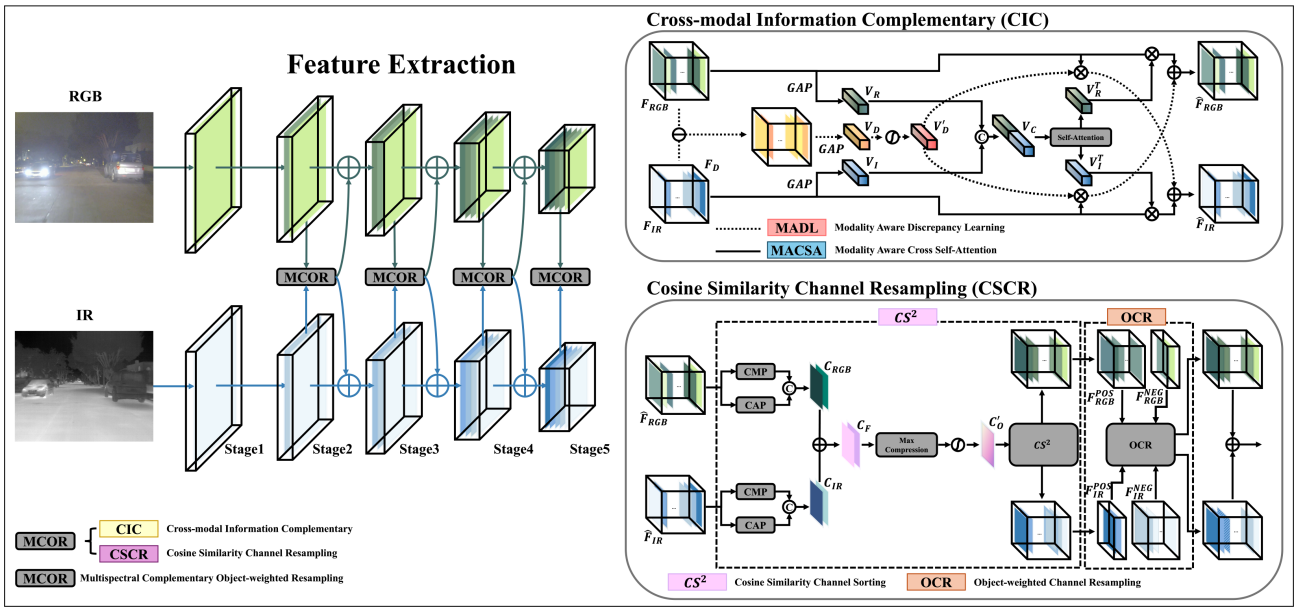

Figure 1. Framework of MCOR Fusion Backbone. This backbone comprises two components: a dual-stream feature extraction network and the Multispectral Complementary Object-weighted Resampling (MCOR) module. The MCOR module incorporates the Cross-modal Information Complementary (CIC) module and the Cosine Similarity Channel Resampling (CSCR) module.

2. Methodology

2.1. Problem Definition

To reveal relationships between inputs, researchers use sophisticated fusion modules to fully leverage information from both modalities [1, 11, 35, 40, 42]. Multispectral object detection involves three fusion strategies: pixel-level [2, 7, 22, 26, 38], decision-level [3, 9, 20], and feature-level [20, 26, 29, 30, 39, 44, 49].

Combining RGB and IR sensors in multispectral object detection leverages varied spectral information, enhancing detection performance. However, conventional fusion strategies like linear combination or direct concatenation face challenges. These methods often fail to distinguish between valuable information and noise, leading to suboptimal use of RGB and IR data. Traditional approaches treat features from both modalities equally, indiscriminately merging informative and noisy data, causing two primary issues:

1. **Inefficient fusion of complementary information,** limiting the effective integration of distinct modal characteristics to improve detection outcomes.

2. **The exacerbated impact of noise on detection performance,** particularly in environments with high levels of interference.

Moreover, incorporating multiple modalities significantly increases the computational complexity of object detection models. While simpler operations like feature addition are used to reduce computational demands, they introduce challenges such as signal interference, information loss, and difficulty handling complex backgrounds. These challenges compromise the model's ability to accurately identify and classify objects, especially when background features obscure crucial object information or when object features interfere with the background.

To overcome these problems, we propose a novel fusion strategy that promotes the interaction between the two modalities and effectively extracts the inconsistent information, thereby improving the accuracy of multispectral object detection and minimizing the impact of noise. The goal of this work is to reduce the negative effects of signal interference, information loss, and background complexity on object detection performance through a fusion mechanism that goes beyond simple addition operations.

2.2. Framework Overview

In this paper, we utilize YOLOv5 [5], a one-stage detector enabling real-time object detection by identifying multiple objects per frame at high speed. As shown in Figure 1, the model takes IR and RGB images as input. The backbone consists of two CSPDarknets53 with five stages each. We fuse the feature maps generated by each stage using two submodules: the Cross-modal Information Complementary (CIC) module and the Cosine Similarity Channel Resampling (CSCR) module. The CIC module learns the disparity information between modalities while

preserving their independent information, enhancing the model's understanding of their unique importance and interrelation. The CSCR module introduces object-weighted and background-weighted feature maps for each modality, preserving crucial information and providing additional insights into modalities lacking object-weighted maps. The fused feature map is generated by aggregating features from both modalities.

2.3. Cross-modal Information Complementary (CIC) Module

To address the aforementioned challenges in Section 2.1. and effectively integrate the complementary information from both modalities, we propose a Cross-modal Information Complementary (CIC) module. This module exploits the interaction between modalities to capture differences by applying channel attention and self-attention to complement uni-modality features with different information. The CIC module can be divided into two sub-modules: 1) Modality Aware Discrepancy Learning (MADL) module and 2) Modality Aware Cross Self-Attention (MACSA) module.

Modality Aware Discrepancy Learning (MADL) Module. The MADL module extracts distinct information from two modalities by emphasizing the differences in their features. This process involves subtracting modality features to exaggerate differences and selectively incorporating complementary features from the other modality. The explicit modeling of modality differences enhances the capacity of the network to learn and leverage complementary features effectively. The structure of the MADL module is shown in Figure 1. Specifically, the MADL module takes the visible feature F_{RGB} and the IR feature F_{IR} as input. It then computes the difference between F_{RGB} and F_{IR} to obtain the difference feature F_D. F_D is further encoded into a global difference vector V_D by the Global Average Pooling (GAP) [23], which represents the difference between the two modalities across channels. The MADL module utilizes V_D to weight the original features, F_{RGB} and F_{IR}, through channel-wise multiplication and a sigmoid activation function. This results in differentially amplified modality features, F_{RGB}^{MADL} and F_{IR}^{MADL}. Finally, F_{RGB}^{MADL} is combined with F_{IR} to complement F_{RGB}, and F_{IR}^{MADL} is combined with F_{RGB}, to complement F_{IR}. This adaptive learning approach allows the MADL module to effectively capture dependencies between channels across datasets, enhancing the network's generalization performance. By incorporating the weighted features, the MADL module ensures that complementary information is appropriately amplified and integrated from both RGB and IR modalities, contributing to the overall robustness and performance of the network.

Modality Aware Cross Self-Attention (MACSA) Module. The MACSA module integrates information from both modalities by employing self-attention, allowing the network to comprehend the interaction between the two spectra and capture complementary information. Additionally, it facilitates the prioritization of important features while filtering out unnecessary information, and the structure of the MACSA module is shown in Figure 1. Specifically, the MACSA module takes visible features F_{RGB} and IR features F_{IR} as input. Then, F_{RGB} and F_{IR} are encoded by GAP into vector V_R and vector V_I, respectively. Next, V_R and V_I are concatenated, and the resulting vector is passed through self-attention to learn the interaction between them, yielding vector V_C^T. V_C^T is then sliced along the channel to obtain V_R^T and V_I^T, bringing them back to the size of the existing vectors for each modality. The MACSA module can effectively utilize V_R^T and V_I^T to weight the features from F_{RGB}^{MADL} and F_{IR}^{MADL}, respectively, through channel-wise multiplication. This results in differentially amplified modality features F_{RGB}^{MACSA} and F_{IR}^{MACSA}. Finally, F_{RGB}^{MADL} is complemented with F_{RGB}^{MACSA}, and F_{IR}^{MADL} with F_{IR}^{MACSA}. The utilization of self-attention in the MACSA module effectively leverages the distinctive characteristics of multiple spectra across the dataset, leading to a more sophisticated feature representation. By weighting F_{RGB}^{MADL} and F_{IR}^{MADL} with V_R^T and V_I^T, respectively, the module ensures the appropriate amplification and integration of complementary information, contributing to the overall effectiveness of the network in capturing intricate relationships between RGB and IR images.

2.4. Cosine Similarity Channel Resampling (CSCR) Module

The Cosine Similarity Channel Resampling (CSCR) module applies cosine similarity and max compression [16] to align channels based on object-centered feature maps and supplement features for modalities lacking object information. The CSCR module can be divided into two sub-modules: 1) Cosine Similarity based Channel Sorting (CS^2) module and 2) Object-weighted Channel Resampling (OCR) module.

Cosine Similarity based Channel Sorting (CS^2) Module. The CS^2 module begins by creating an object-centered feature map, which serves as a reference for subsequent operations. This object-centric feature map is crucial for calculating the cosine similarity per channel for each modality and sorting the channels based on their similarity to the object-centric feature map. The structure of the CS^2 module is shown in Figure 1. Specifically, the CS^2 module takes $\widehat{F}_{RGB}$ and $\widehat{F}_{IR}$ as input, obtained from the CIC module proposed in this paper. Then, Channel Max Pooling (CMP) and Channel Average Pooling (CAP) operations are performed on each modality information to obtain F_{RGB}^{CAP}, F_{RGB}^{CMP}, F_{IR}^{CAP}, and F_{IR}^{CMP}. The information from each

modality is concatenated using C_{RGB} and C_{IR}, obtaining C_F, and max compression is applied to compress the two channels into one channel, resulting in C_O, which represents the object-centered channel and is fed into a sigmoid function to obtain $C_O^{'}$.

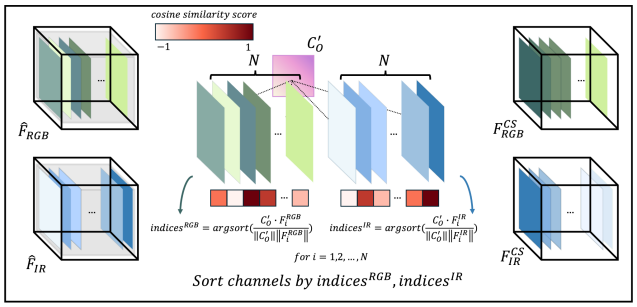

Figure 2. left: The Structure of Cosine Similarity based Channel Sorting (CS^2) module.

Figure 2 shows the cosine similarity-based sorting mechanism, starting with the highest value. After sorting, the feature maps for each modality are organized into object-centric feature maps from the front. This ensures that during the addition operation on two modalities, the object-centric feature maps contribute to each other, emphasizing important information while excluding irrelevant features and noise. Specifically, it computes the cosine similarity between $C_O^{'}$ and $F_i^{\{RGB,IR\}}$, for each $i = 1, 2, ..., N$, obtaining the cosine similarity-based aligned feature maps F_{RGB}^{CS} and F_{IR}^{CS}.

Method	mAP$_{50}$(%)	mAP$_{75}$(%)	mAP(%)
Manhattan	75.9	33.9	39.0
Euclidean	76.3	32.6	38.2
MSE	76.3	33.4	38.7
Pearson	76.4	32.9	38.6
SSIM	77.1	33.2	38.8
Cosine	**78.2 (+1.1)**	**34.4 (+1.2)**	**39.9 (+1.1)**

Table 1. Comparisons of performances using various similarity-based methods with FLIR datasets in terms of mAP$_{50}$, mAP$_{75}$, and mAP.

Why Cosine Similarity? Cosine similarity measures the angle between two non-zero vectors, emphasizing structural and directional features rather than intensity or size. By normalizing vectors, it ensures comparisons are independent of lighting or scale, making it robust to variations in imaging conditions. This focus on vector orientation highlights geometric patterns, making it effective for comparing shapes in images.

The quantitative evaluation demonstrates that cosine similarity outperforms other methods in the proposed model, improving mAP$_{50}$, mAP$_{75}$, and mAP metrics by 1.1%p, 1.2%p, and 1.1%p respectively. Its robustness and accuracy make it the most reliable metric for the CS^2 module.

Object-weighted Channel Resampling (OCR) The OCR module is introduced to address the variability in the number of object-centric and background-centric features in each modality's feature map after passing through the backbone. The aim is to enhance the identification and classification of objects, especially in complex backgrounds or scenarios with multiple overlapping objects. The structure of the OCR module illustrated in Figure 3 is defined as the following equation.

$$\begin{aligned} &F_{RGB}^{pos}, \quad F_{RGB}^{neg} = F_{RGB}^{CS}[:, : N_{RGB}, :, :], \\ &\quad F_{RGB}^{CS}[:, N_{RGB} :, :, :] \\ &F_{IR}^{pos}, \quad F_{IR}^{neg} = F_{IR}^{CS}[:, : N_{IR}, :, :], \\ &\quad F_{IR}^{CS}[:, N_{IR} :, :, :] \\ &N_{RGB}, \quad N_{IR} = \sum \mathbb{1}^{S_{RGB}>0}(F_{RGB}^{CS}), \\ &\quad \sum \mathbb{1}^{S_{IR}>0}(F_{IR}^{CS}). \end{aligned} \tag{1}$$

The OCR module takes two modality feature maps F_{RGB}^{CS} and F_{IR}^{CS}, whose channels are aligned based on cosine similarity. F_{RGB}^{CS} and F_{IR}^{CS} are divided into positive (F_{RGB}^{pos}, F_{IR}^{pos}) and negative samples (F_{RGB}^{neg}, F_{IR}^{neg}) based on the cosine similarity of zero. To determine the number of positive samples for each modality, N_{RGB} and N_{IR}, we utilize the $\mathbb{1}(\cdot)$ function that identifies values greater than zero for each given cosine similarity score, S_{RGB} and S_{IR}, between $C_O^{'}$ and $F_{RGB,IR}^{CS}$.

$$\begin{aligned} &F_{RGB}^{pos}, F_{RGB}^{neg} = concat(F_{RGB}^{pos}, F_F^{extra}), \\ &\quad F_{RGB}^{neg}[:, N_{extra} :, :, :], \quad \text{if } N_{RGB} < N_{IR} \\ &F_{IR}^{pos}, F_{IR}^{neg} = concat(F_{IR}^{pos}, F_F^{extra}), \\ &\quad F_{IR}^{neg}[:, N_{extra} :, :, :], \quad \text{if } N_{RGB} > N_{IR} \\ &F_F^{extra} = max(F_{RGB}^{pos}[:, : N_{extra}, :, :], \\ &\quad F_{IR}^{pos}[:, : N_{extra}, :, :]), \quad \text{if } N_{extra} \neq 0. \end{aligned} \tag{2}$$

To match the number of positive and negative samples for each modality, it takes the positive samples of the two modalities, F_{RGB}^{pos} and F_{IR}^{pos}, as input. N_{RGB} and N_{IR} are used to find the difference, $N_{extra} = |N_{RGB} - N_{IR}|$. If N_{extra} is non-zero, the mismatched object-oriented feature map, F_F^{extra}, is generated using the max compression technique for the two modalities, F_{RGB}^{pos} and F_{IR}^{pos}, by the number of missing channels. The F_F^{extra} is then concatenated to which has insufficient positive feature maps, and channels with zero cosine similarity in the negative sample are dropped for the number of added channels.

The CSCR module emphasizes object-centric feature maps, removes unnecessary feature maps, and effectively enhances the clarity between object and background while reducing noise. By integrating information from both modalities and emphasizing object-centric feature maps, the proposed CSCR module contributes to improving the accuracy of multispectral object detection.

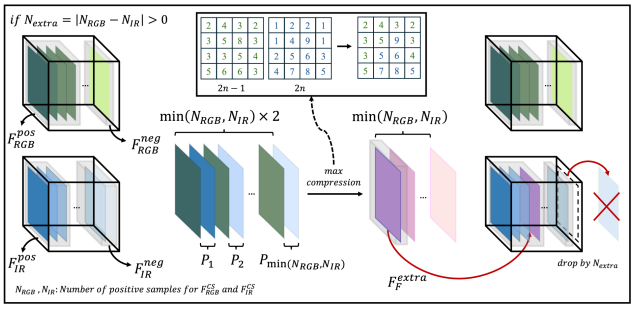

Figure 3. The Structure of Object-weighted Channel Resampling (OCR) module. The diagram depicts a situation where the RGB feature map has more positive samples than the IR feature map.

3. Experiments

3.1. Implementation Details

All our networks are developed using Python, and both the source code and pre-trained models are accessible. At each epoch in the training phase, we reserved 10% of the training images for validation. We used the same hyper-parameter settings from the original YOLOv5. All models were trained for a maximum of 100 epochs, using a batch size of 16 and an input image dimension of 640×640. The performance evaluation was conducted on a single NVIDIA GTX 3090 GPU. This approach guarantees a fair comparison with previous state-of-the-art methods.

3.2. Datasets

LLVIP The LLVIP dataset is important for multispectral object detection, consisting of 15,488 RGB and IR image pairs from surveillance cameras at 26 locations, mostly under low-light conditions. It includes 12,025 pairs for training and 3,463 for testing, with a focus on the "pedestrian" category for urban safety and surveillance.

FLIR The FLIR dataset is widely used in multispectral object detection, featuring 5,142 aligned RGB-IR image pairs from a car driver's perspective, covering day and night scenarios. It includes 4,129 pairs for training and 1,013 for testing, focusing on three object types: people, cars, and bikes, after excluding the dog category.

M^{3}FD The M^{3}FD dataset includes 8,400 images, with 4,200 RGB and IR image pairs for fusion and detection tasks, plus 600 images for fusion only. It has 34,407 labeled instances across six categories: People, Car, Bus, Motorcycle, Lamp, and Truck.

VEDAI The Vehicle Detection in Aerial Imagery (VEDAI) dataset contains over 3,700 annotated targets in 1,268 RGB-infrared image pairs, categorizing vehicles like cars, trucks, planes, and boats. Images are standardized at 1024×1024 resolution, with annotations in horizontal boxes.

KAIST The KAIST dataset includes 95,328 visible and infrared image pairs, with 103,128 bounding boxes identifying 1,182 pedestrians. Due to issues with initial annotations, improved annotations were used for training [26, 46], enhancing data reliability and model accuracy. The test set has 2,252 frames, selected from every 20th video frame, with 1,455 daytime and 797 nighttime images to evaluate algorithm performance based on the time of day.

3.3. Metrics

The evaluation of object detection models is measured by mean Average Precision (mAP), which assesses detection accuracy. A true positive is a correctly identified object with an Intersection over Union (IoU) above a threshold, while a false positive is an incorrect detection, and a false negative is a missed detection. mAP involves precision (ratio of correctly predicted objects to all predictions) and recall (fraction of actual objects detected). Specific IoU thresholds like mAP$_{.5}$, mAP$_{.75}$, and mAP (averaging precision across IoU thresholds from 0.5 to 0.95) are used across datasets like LLVIP, FLIR, M^{3}FD, and VEDAI.

For the KAIST dataset, we use Log-Average Miss Rate (MR^{-2}) to measure pedestrian detection performance [12]. MR^{-2} averages the Miss Rate over nine False Positives Per Image (FPPI) in the range of $[10^{-2}, 10^{0}]$ uniformly distributed in log space.

3.4. Quantitative Results

Table 2 presents a comparative analysis of state-of-the-art (SOTA) models using FLIR, LLVIP, and M^{3}FD datasets. The proposed MCOR method outperforms YOLOv5 (RGB), the baseline network, by 9.5%p on FLIR, 6.8%p on LLVIP, and 3.7%p on M^{3}FD in terms of AP50. It also surpasses YOLOv5 (IR) by 3.4%p on FLIR and 3.0%p on LLVIP and M^{3}FD datasets. These results demonstrate MCOR's effectiveness in integrating information from two modalities, enhancing detection accuracy.

Compared to multi-modal models, MCOR exceeds the YOLOv5-based simple fusion model by 3.1%p on FLIR, 1.8%p on LLVIP, and 3.4%p on M^{3}FD at AP50. This improvement is due to the efficient interaction between modalities, effective extraction of inconsistent information, and noise reduction.

MCOR also outperforms other SOTA multi-modal models, achieving superior results in AP50, AP75, and mAP

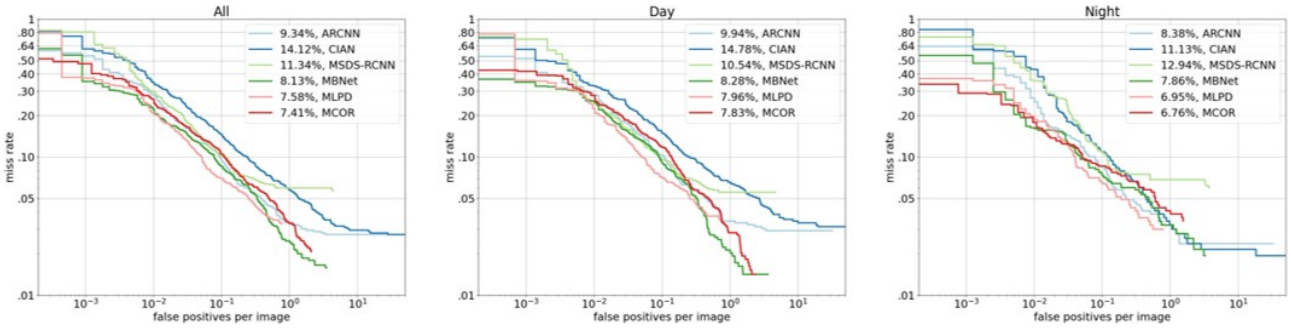

Figure 4. Comparisons of detection results reported on the test set of KAIST dataset, in terms of ALL (left), Day (middle), and Night (right).

across all datasets. Specifically, on the FLIR dataset, MCOR surpasses GAFF by 4.6%p and CFT by 3.2%p. On LLVIP, it exceeds UA-CMDet by 1.3%p and CFT by 0.8%p. On M^{3}FD, it outperforms CFT by 4%p and ADCNet by 2.8%p. These outcomes underscore MCOR's robust generalizability and SOTA performance across various datasets.

As shown in Table 3, the MCOR achieves the highest mAP$_{50}$ of 76.2% on the VEDAI dataset, outperforming models like Input Fusion, Mid Fusion, SuperYOLO, and CFT. This superior detection precision at an IoU threshold of 0.5 highlights MCOR's leading role in this metric. Its advantage over models such as Mid Fusion demonstrates MCOR's robust capability in high-precision scenarios, affirming its status as a top solution in object detection requiring rigorous accuracy.

We evaluate the proposed MCOR by comparing it with CIAN [45], MSDS-RCNN [18], ARCNN [46], MBNet [48], and MLPD [15] on the KAIST test set. Figure 4 shows the comparison of results in terms of MR under reasonable settings. The evaluation illustrates the performance in terms of miss rate against false positives per image, segmented into all conditions, day, and night. The MCOR model exhibits significant reductions in miss rates across all conditions: 7.41%, 7.83%, and 6.76%, respectively. This consistent outperformance compared to methods like ARCNN and CIAN, which show higher miss rates, underscores MCOR's effectiveness. During the day, while CIAN has a higher miss rate at 14.78%, MCOR maintains a lower rate of 7.83%. At night, MCOR achieves the lowest miss rate, highlighting its robustness in low-light environments, which are challenging for pedestrian detection.

Table 4 presents a comparative analysis of the performance between uni- and multi-modal state-of-the-art (SOTA) models using the KAIST dataset. MCOR outperforms most SOTA models, achieving a reduction in miss rate (MR) by 7.41%, 7.83%, and 6.76% for the day, night, and all-day subsets,respectively, under an IoU threshold of 0.5. These results demonstrate the effectiveness of MCOR in seamlessly integrating information from both modalities and emphasizing objects to improve detection accuracy.

The MCOR module, as shown in Table 5, significantly improves pedestrian detection across various distances and occlusion levels using the KAIST datasets. It excels in detecting distant pedestrians, achieving the lowest miss rate of 34.18%, highlighting the model's capability to accurately identify pedestrians far from the sensor, crucial for safety in urban environments.

Additionally, the proposed method performs well under varying occlusion levels, achieving the lowest miss rates in 'None' and 'Heavy' occlusion scenarios with 19.81% and 53.58%, respectively. This reliability in both clear and occluded conditions is essential for real-world applications, such as crowded urban scenes.

3.5. Qualitative Results

A qualitative assessment using the FLIR dataset is shown in Figure 5. These results highlight the complementary nature of infrared and RGB images. The infrared detector struggles with capturing a bicycle due to subtle temperature differences, while the RGB image captures texture and color details, enabling precise detection by YOLOv5 (RGB). Conversely, the RGB detector has limitations in detecting pedestrians compared to the infrared detector.

The multi-modal detector showed an increased false positive rate. However, our method achieves a lower false positive rate than CFT. The MCOR retains relevant details from each modality, enhancing the understanding of each. Meanwhile, the CSCR module retains both object and background information and emphasizes the object-weighted feature map, resulting in the lowest false positive rate among the three models.

3.6. Ablation Study

Table 6 and Table 7 present a comparative analysis of the individual contributions of CIC and CSCR modules within the proposed method, as tested on FLIR, LLVIP, and M^{3}FD datasets. The results show that both sub-blocks are effec-

Dataset	Model	Modality	mAP$_{50}$(%)	mAP$_{75}$(%)	mAP(%)
uni-modality networks					
FLIR	Faster R-CNN [34]	RGB	64.9	21.1	28.9
	Faster R-CNN [34]	IR	74.4	32.5	37.6
	SDD [27]	RGB	52.2	15.8	21.8
	SDD [27]	IR	65.5	22.4	29.6
	YOLOv3 [33]	RGB	58.3	19.8	25.7
	YOLOv3 [33]	IR	73.6	31.3	36.8
	YOLOv5 [5]	RGB	67.8	25.9	31.8
	YOLOv5 [5]	IR	73.9	35.7	39.5
LLVIP	Faster R-CNN [34]	RGB	91.4	48.0	49.2
	Faster R-CNN [34]	IR	96.1	68.5	61.1
	SDD [27]	RGB	82.6	31.8	39.8
	SDD [27]	IR	90.2	57.9	53.5
	YOLOv3 [33]	RGB	85.9	37.9	43.3
	YOLOv3 [33]	IR	89.7	53.4	52.8
	YOLOv5 [5]	RGB	90.8	51.9	50.0
	YOLOv5 [5]	IR	94.6	72.2	61.9
M^{3}FD	YOLOv5 [5]	RGB	83.5	-	55.2
	YOLOv5 [5]	IR	84.2	-	56.1
multi-modality networks					
FLIR	MMTOD-CG [4]	RGB+IR	61.4	-	-
	MMTOD-UNIT [4]	RGB+IR	61.5	-	-
	Halfway [26]	RGB+IR	71.2	-	-
	CFR [42]	RGB+IR	72.4	-	-
	GAFF [41]	RGB+IR	72.7	30.9	37.3
	SMPD [21]	RGB+IR	73.6	-	-
	BU-ATT [14]	RGB+IR	73.1	-	-
	BU-LTI [14]	RGB+IR	73.2	-	-
	YOLOv5 [30]	RGB+IR	74.2	29.1	35.6
	CFT [30]	RGB+IR	74.5	30.1	36.5
	ICAFusion [35]	RGB+IR	77.3	32.5	38.1
	MCOR(Ours)	RGB+IR	**78.2**	**34.4**	**39.9**
LLVIP	YOLOv5 [30]	RGB+IR	95.8	71.4	62.3
	BU-ATT [14]	RGB+IR	92.6	-	-
	BU-LTI [14]	RGB+IR	92.9	-	-
	UA-CMDet [36]	RGB+IR	96.3	-	-
	YOLOv7 [37]	RGB+IR	96.9	73.1	64.4
	CFT [30]	RGB+IR	96.8	70.7	62.9
	ICAFusion [35]	RGB+IR	97	71.3	62.4
	MCOR(Ours)	RGB+IR	**97.6**	**73.7**	**64.9**
M^{3}FD	DDFusion [8]	RGB+IR	65.7	38.9	38.4
	DIDFuse [47]	RGB+IR	63.5	36	36.6
	TarDAL [25]	RGB+IR	78	48	46.9
	CFR [42]	RGB+IR	82.7	50.8	50.3
	CFT [30]	RGB+IR	83.2	55.1	52.2
	ADCNet [10]	RGB+IR	84.4	55.7	53.3
	ICAFusion [35]	RGB+IR	86.6	58.4	56.9
	MCOR(Ours)	RGB+IR	**87.2**	**62.5**	**57.3**

Table 2. Comparisons of performances with different datasets in terms of mAP50, mAP75, and mAP. The best scores highlighted in **bold** and the second best scores highlited in underline.

tive in improving the prediction accuracy compared to the uni-modal object detector mentioned in Table 2. In Table 7, the CSCR module also proven effective in improving the prediction accuracy when experimented with in addition to other multi-modal object detection models. We conclude that the CIC module preserves valuable information from both modalities, enhancing the comprehension of each modality. Simultaneously, the CSCR module effectively preserves both object and background information while emphasizing the object-weighted feature map. Therefore, the two modules complement each other, further improving the overall performance.

Dataset	Model	mAP$_{50}$(%)	mAP(%)
VEDAI	Input Fusion [31]	74.4	45.7
	Mid Fusion [31]	74.8	**46.3**
	SuperYOLO [43]	73.6	-
	CFT [30]	74.6	44.1
	ICAFusion [35]	75.7	46.2
	MCOR (Ours)	**76.2**	**46.3**

Table 3. Comparisons of performances with VEDAI datasets in terms of mAP$_{50}$, and mAP. The best scores highlighted in **bold** and the second best scores highlited in underline.

Model	Miss Rate (%) ↓		
	Day	Night	All
ACF+T+THOG [12]	47.32	42.65	56.18
Halfway Fusion [26]	25.77	24.91	26.67
Fusion RPN [17]	19.88	22.12	20.77
Fusion RPN+BF [17]	18.29	19.57	16.27
IAF-RCNN [19]	14.55	18.26	15.73
IATDNN+IASS [9]	14.67	14.18	15.28
CIAN [45]	14.13	14.78	11.14
MSDS-RCNN [18]	10.60	13.73	11.63
AR-CNN [46]	10.22	10.80	9.02
MBNet [48]	8.40	8.62	8.27
ICAFusion [35]	7.97	8.14	7.05
MLPD [15]	7.58	7.96	6.95
MCOR (Ours)	**7.41**	**7.83**	**6.76**

Table 4. Comparisons of performances with KAIST datasets in terms of MR^{-2} under IoU threshold of 0.5. The lowest MR highlighted in **bold** and the second best scores highlited in underline.

Model	Miss Rate (%) ↓					
	Near	Medium	Far	None	Partial	Heavy
CIAN [45]	3.71	19.11	69.01	31.40	38.63	55.73
MSDS-RCNN [18]	1.29	16.28	63.73	30	38.71	63.37
AR-CNN [46]	**0**	16.08	69.01	31.40	38.63	55.73
MBNet [48]	0.03	16.09	55.99	27.75	35.43	59.14
ICAFusion [35]	0.04	15.05	42.89	23.03	36.73	60.77
MLPD [15]	0.04	**11.93**	50.86	24.15	**28.75**	53.97
MCOR (Ours)	0.03	14.05	**34.18**	**19.81**	28.81	**53.58**

Table 5. Comparisons of performances with KAIST datasets in terms of MR^{-2} under IoU threshold of 0.5. The pedestrian distances consist of "Near" (115 $\leq$ height), "Medium" (45 $\leq$ height ¡ 115) and "Far" (1 $\leq$ height ¡ 45), while occlusion levels consist of "None" (never occluded), "Partial" (occluded to some extent up to one half) and "Heavy" (mostly occluded). The lowest MR highlighted in **bold** and the second best scores highlited in underline.

Table 8 compares the contributions of CIC and CSCR modules in our method on the KAIST dataset. Adding the CIC module to YOLOv5 reduces the Miss Rate (MR) from 10.04% to 7.97%, a 2.07%p improvement. The CSCR

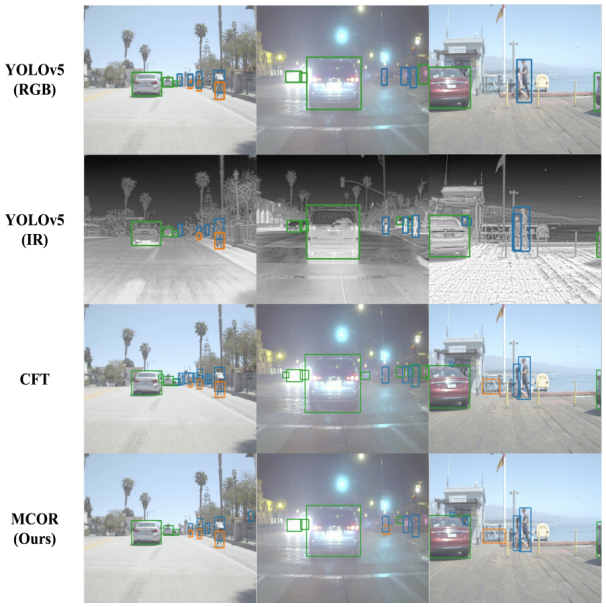

Figure 5. The detection results of the four models mentioned in the experiment. The cars, pedestrians, and bicycles detected by the model are represented by green, blue, and orange bounding boxes, respectively.

Dataset	CIC	CSCR	mAP$_{50}$(%)	mAP$_{75}$(%)	mAP(%)
FLIR	×	×	74.2	29.1	35.6
	✓	×	75.7	31.1	37.3
	×	✓	75.0	31.0	37.3
	✓	✓	**78.2 (+4.0)**	**34.4 (+5.3)**	**39.9 (+4.3)**
LLVIP	×	×	96.7	71.3	62.3
	✓	×	97.2	73.5	64.2
	×	✓	96.7	72.4	63.9
	✓	✓	**97.6 (+0.9)**	**73.7 (+2.4)**	**64.9 (+2.6)**
M^{3}FD	×	×	82.6	55.8	52.3
	✓	×	86.1	62.4	57.8
	×	✓	85.2	60.3	56.6
	✓	✓	**87.2 (+4.6)**	**60.9 (+5.1)**	**57.2 (+4.9)**

Table 6. Comparisons of performances with different datasets in terms of mAP$_{50}$, mAP$_{75}$, and mAP.

module lowers the MR from 10.04% to 8.76%, a 1.28%p gain. These improvements are consistent across different times of day; with the CIC module, the daytime MR drops from 10.01% to 8.32% and nighttime MR from 8.75% to 7.17%. The CSCR module reduces daytime MR to 8.41% and nighttime MR to 7.87%.

The CIC module enhances understanding by preserving critical information across modalities, while the CSCR module maintains object and background features, emphasizing context in object detection. As shown in Table 8, our method achieves 35.4 FPS on the RTX 3090, highlighting its utility in scenarios demanding high detection speed and accuracy.

Dataset	Method	CSCR	mAP$_{50}$(%)	mAP$_{75}$(%)	mAP(%)
FLIR	Two Stream	×	74.2	29.1	35.6
	Two Stream	✓	**75.0 (+0.8)**	**31.0 (+1.9)**	**37.3 (+1.7)**
	CFT [30]	×	74.5	30.1	36.5
	CFT [30]	✓	**75.6 (+1.1)**	**30.9 (+0.8)**	**37.2 (+0.7)**
LLVIP	Two Stream	×	96.7	71.3	62.3
	Two Stream	✓	**96.8 (+0.1)**	**71.4 (+0.1)**	**62.8 (+0.2)**
	CFT [30]	×	96.8	70.7	62.9
	CFT [30]	✓	**97.3 (+0.5)**	**71.3 (+0.6)**	**62.3 (-0.6)**
M^{3}FD	Two Stream	×	82.6	55.8	52.3
	Two Stream	✓	**85.2 (+2.6)**	**60.3 (+5.5)**	**56.6 (+4.3)**
	CFT [30]	×	83.2	55.1	52.2
	CFT [30]	✓	**86 (+2.8)**	**59.3 (+4.2)**	**57.2 (+5.0)**

Table 7. Ablation study resultf for CSCR on the different datasets and different multi-modality networks in term of mAP$_{50}$, mAP$_{75}$, and mAP.

CIC	CSCR	Miss Rate (%) ↓ All	Day	Night	FPS (Hz)
×	×	10.04	10.01	8.75	45.7
✓	×	7.97 (-2.07)	8.32 (-1.69)	7.17 (-1.58)	39.2
×	✓	8.76 (-1.28)	8.41 (-1.6)	7.87 (-0.88)	41.9
✓	✓	**7.41 (-2.18)**	**7.83 (-2.21)**	**6.76 (-1.99)**	35.4

Table 8. Ablation study results for CIC and CSCR on the KAIST dataset. (Bold represents the best result in each column.)

4. Conclusions

In this paper, we propose the MCOR framework to address modality imbalance and integrate information from both RGB and IR modalities for multi-modal object detection. The MCOR framework demonstrates superior information capture compared to other approaches. The CIC module enhances uni-modal features with complementary information, promoting balanced contributions from RGB and IR modalities. The CSCR module uses an object-centric feature map to weigh objects and compensate for modality misalignments. Our results show improved prediction accuracy when both modules are used. Additionally, the framework adapts well across different models, showcasing strong generalization capabilities.

Acknowledgment This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2021-0-01341, Artificial Intelligence Graduate School Program(Chung-Ang University)), MSIT(Ministry of Science and ICT), Korea, under the Graduate School of Metaverse Convergence support program(IITP-RS2024-00418847) and Korea Research Institute for defense Technology planning and advancement through Defense Innovation Vanguard Enterprise Project, funded by Defense Acquisition Program Administration(R230106).

References

[1] Yue Cao, Junchi Bin, Jozsef Hamari, Erik Blasch, and Zheng Liu. Multimodal object detection by channel switching and spatial attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 403–411, 2023. 2

[2] Yanpeng Cao, Xing Luo, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Locality guided cross-modal feature aggregation and pixel-level fusion for multispectral pedestrian detection. *Information Fusion*, 88:1–11, 2022. 2

[3] Zhiwei Cao, Huihua Yang, Juan Zhao, Shuhong Guo, and Lingqiao Li. Attention fusion for one-stage multispectral pedestrian detection. *Sensors*, 21(12):4184, 2021. 2

[4] Chaitanya Devaguptapu, Ninad Akolekar, Manuj M Sharma, and Vineeth N Balasubramanian. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 7

[5] Glenn Jocher et. al. ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support, Oct. 2021. 2, 7

[6] TELEDYNE FLIR. Free flir thermal dataset for algorithm training. 1

[7] Lei Fu, Wen-bin Gu, Yong-bao Ai, Wei Li, and Dong Wang. Adaptive spatial pixel-level feature fusion network for multispectral pedestrian detection. *Infrared Physics & Technology*, 116:103770, 2021. 2

[8] Yu Fu, Xiao-Jun Wu, and Josef Kittler. Deep decomposition network for image processing: A case study for visible and infrared image fusion. *arXiv preprint arXiv:2102.10526*, 2021. 7

[9] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019. 2, 7

[10] Mingzhou He, Qingbo Wu, King Ngi Ngan, Feng Jiang, Fanman Meng, and Linfeng Xu. Misaligned rgb-infrared object detection via adaptive dual-discrepancy calibration. *Remote Sensing*, 15(19):4887, 2023. 7

[11] Ziyue Huang, Qingjie Liu, Huanyu Zhou, Guangshuai Gao, Tao Xu, Qi Wen, and Yunhong Wang. Building detection from panchromatic and multispectral images with dual-stream asymmetric fusion networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:3364–3377, 2023. 2

[12] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 1, 5, 7

[13] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3496–3504, 2021. 1

[14] My Kieu, Andrew D Bagdanov, and Marco Bertini. Bottom-up and layerwise domain adaptation for pedestrian detection in thermal images. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1):1–19, 2021. 7

[15] Jiwon Kim, Hyeongjun Kim, Taejoo Kim, Namil Kim, and Yukyung Choi. Mlpd: Multi-label pedestrian detector in multispectral domain. *IEEE Robotics and Automation Letters*, 6(4):7846–7853, 2021. 6, 7

[16] Mingi Kim, Heegwang Kim, Junghoon Sung, Chanyeong Park, and Joonki Paik. High-resolution processing and sigmoid fusion modules for efficient detection of small objects in an embedded system. *Scientific Reports*, 13(1):244, 2023. 3

[17] Daniel Konig, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch. Fully convolutional region proposal networks for multispectral person detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 49–56, 2017. 7

[18] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Multispectral pedestrian detection via simultaneous detection and segmentation. *arXiv preprint arXiv:1808.04818*, 2018. 6, 7

[19] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019. 7

[20] Qing Li, Changqing Zhang, Qinghua Hu, Huazhu Fu, and Pengfei Zhu. Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection. *IEEE Transactions on Multimedia*, 2022. 2

[21] Qing Li, Changqing Zhang, Qinghua Hu, Pengfei Zhu, Huazhu Fu, and Lei Chen. Stabilizing multispectral pedestrian detection with evidential hybrid fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 7

[22] Shutao Li, Xudong Kang, Leyuan Fang, Jianwen Hu, and Haitao Yin. Pixel-level image fusion: A survey of the state of the art. *information Fusion*, 33:100–112, 2017. 2

[23] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 3

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1

[25] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022. 1, 7

[26] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N Metaxas. Multispectral deep neural networks for pedestrian detection. *arXiv preprint arXiv:1611.02644*, 2016. 2, 5, 7

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer*

Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pages 21–37. Springer, 2016. 7

[28] K Oksuz, BC Cam, S Kalkan, and E Akbas. Imbalance problems in object detection: A review. arxiv e-prints p. *arXiv preprint arXiv:1909.00169*, 2019. 1

[29] Peiran Peng, Tingfa Xu, Bo Huang, and Jianan Li. Hafnet: Hierarchical attentive fusion network for multispectral pedestrian detection. *Remote Sensing*, 15(8):2041, 2023. 2

[30] Fang Qingyun, Han Dapeng, and Wang Zhaokui. *arXiv preprint arXiv:2111.00273*, 2021. 2, 7, 8

[31] Fang Qingyun and Wang Zhaokui. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognition*, 130:108786, 2022. 7

[32] Sebastien Razakarivony and Frederic Jurie. Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34:187–203, 2016. 1

[33] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 7

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 7

[35] Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145:109913, 2024. 2, 7

[36] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022. 7

[37] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 7

[38] Mengliang Xing, Gang Liu, Haojie Tang, Yao Qian, and Jun Zhang. Multi-level adaptive perception guidance based infrared and visible image fusion. *Optics and Lasers in Engineering*, 171:107804, 2023. 2

[39] Lijuan Xu and Xuemiao Xu. Rgb-d visual saliency detection algorithm based on information guided and multimodal feature fusion. *IEEE Access*, 2023. 2

[40] Maoxun Yuan, Xiaorong Shi, Nan Wang, Yinyan Wang, and Xingxing Wei. Improving rgb-infrared object detection with cascade alignment-guided transformer. *Information Fusion*, 105:102246, 2024. 2

[41] Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International conference on image processing (ICIP)*, pages 276–280. IEEE, 2020. 7

[42] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Guided attentive feature fusion for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 72–80, 2021. 2, 7

[43] Jiaqing Zhang, Jie Lei, Weiying Xie, Zhenman Fang, Yunsong Li, and Qian Du. Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 7

[44] Jiaqi Zhang, Dan Zhang, Wenping Ma, and Licheng Jiao. Deep self-paced residual network for multispectral images classification based on feature-level fusion. *IEEE Geoscience and Remote Sensing Letters*, 15(11):1740–1744, 2018. 2

[45] Lu Zhang, Zhiyong Liu, Shifeng Zhang, Xu Yang, Hong Qiao, Kaizhu Huang, and Amir Hussain. Cross-modality interactive attention network for multispectral pedestrian detection. *Information Fusion*, 50:20–29, 2019. 6, 7

[46] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5127–5137, 2019. 5, 6, 7

[47] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Pengfei Li, and Jiangshe Zhang. Didfuse: Deep image decomposition for infrared and visible image fusion. *arXiv preprint arXiv:2003.09210*, 2020. 7

[48] Kailai Zhou, Linsen Chen, and Xun Cao. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 787–803. Springer, 2020. 6, 7

[49] Xin Zuo, Zhi Wang, Yue Liu, Jifeng Shen, and Haoran Wang. Lgadet: Light-weight anchor-free multispectral pedestrian detection with mixed local and global attention. *Neural Processing Letters*, 55(3):2935–2952, 2023. 2