Cross-Modal Local Calibration and Global Context Modeling Network for RGB–Infrared Remote-Sensing Object Detection

Jin Xie, Jing Nie, Bonan Ding, Mingyang Yu, and Jiale Cao, *Member, IEEE*

***Abstract*—RGB–infrared object detection in remote-sensing images is crucial for achieving around-clock surveillance of unmanned aerial vehicles. RGB–infrared remote-sensing object detection methods based on deep learning usually mine the complementary information from RGB and infrared modalities by utilizing feature aggregation to achieve robust object detection for around-the-clock applications. Most of the existing methods aggregate features from RGB and infrared images by utilizing elementwise operations (e.g., elementwise addition or concatenation). The detection accuracy of these methods is limited. The main reasons can be concluded as follows: local location misalignment across modalities and insufficient nonlocal contextual information extraction. To address the above issues, we propose a cross-modal local calibration and global context modeling network (CLGNet), consisting of two novel modules: a cross-modal local calibration (CLC) module and a cross-modal global context (CGC) modeling module. The CLC module first aligns features from different modalities and then aggregates them selectively. The CGC module is embedded into the backbone network to capture cross-modal nonlocal long-range dependencies. The experimental results on popular RGB–infrared remote-sensing object detection datasets, namely DRoneVehicle and VEDAI, demonstrate the effectiveness and efficiency of our CLGNet.**

***Index Terms*—Multimodal fusion, object detection, remote-sensing object detection.**

I. Introduction

RGB–INFRARED remote-sensing object detection is a crucial remote-sensing task that focuses on classifying and locating arbitrary-oriented objects within a pair of RGB and infrared remote-sensing images, which plays a significant role in remote image understanding. Recently, deep neural networks have pervaded many areas of remote-sensing understanding ranging from remote-sensing image classification [1], [2], [3], [4] and remote-sensing object detection [5], [6], [7], [8], [9], [10], [11] to RGB–infrared remote-sensing object detection [12], [13], [14], [15].

Existing RGB-based remote-sensing object detection methods achieve great success, particularly in well-lit environments. However, the detection performances of these methods are significantly declined in low-light conditions. The main reason is that RGB cameras struggle to capture useful information about objects in low-light scenarios. In contrast, infrared cameras are sensitive to temperature variations and can capture infrared images that provide additional information during nighttime. However, infrared images often suffer from the loss of details and lower resolution, which are the crucial factors for decreasing the detection accuracy. To address the above challenges, remote-sensing object detection methods using RGB–infrared images have emerged as a powerful solution. The RGB–infrared remote-sensing object detection methods aim at combining the strengths of RGB images and infrared images to achieve the accurate object detection across a range of illumination conditions.

RGB–infrared remote-sensing object detection [13], [14], [16] has gained significant popularity in various applications, such as surveillance of unmanned aerial vehicles. These methods typically rely on fusing features extracted from both RGB and infrared input images. UACMDet [13] aggregates multimodal features by utilizing the uncertainty information of two modalities, which mines cross-modal complementary information effectively. The redundant information suppression network (RISNet) [14] is proposed to fuse RGB–infrared features by mitigating cross-modal redundancy and mining the complementary cross-modal information. The aforementioned methods generally consider the effect of modality importance in feature aggregation.

RGB–infrared remote-sensing object detection methods face some key challenges that limit the detection accuracy. The first one is the misalignment in spatial positions between the RGB features and infrared features, caused by various factors, such as time differences in image capturing and distances between different cameras. The second one is lacking of global contextual information, which is proven the clue of reasoning the

Manuscript received 22 May 2023; revised 11 August 2023; accepted 2 September 2023. Date of publication 14 September 2023; date of current version 3 October 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0160404, in part by the National Natural Science Foundation of China under Grant 62206031, Grant 62271346, and Grant 62301092, in part by China Postdoctoral Science Foundation under Grant 2021M700613, Grant 2022M720581, and Grant 2023T160762, in part by Tianjin Natural Science Foundation under Grant 21JCQNJC00420, and in part by the Fundamental Research Funds for the Central Universities under Grant 2023CDJXY-036. *(Corresponding authors: Jin Xie; Jing Nie.)*

Jin Xie, Bonan Ding, and Mingyang Yu are with the School of Big Data and Software Engineering, Chongqing University, Chongqing 400044, China (e-mail: xiejin@cqu.edu.cn; 202224131052@stu.cqu.edu.cn; 20222413101 3t@stu.cqu.edu.cn).

Jing Nie is with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China (e-mail: jingnie@cqu.edu.cn).

Jiale Cao is with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China, and also with Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China (e-mail: connor@tju.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3315544

existence and locations of objects. The elementwise operations for aggregating multimodal features in existing methods [13], [14] make it difficult to effectively address the aforementioned issues.

In this article, we propose a novel feature aggregation manner, terms cross-modal local calibration and global context modeling network (CLGNet), consisting of a cross-modal local calibration (CLC) module and a cross-modal global context (CGC) modeling module, which solves the above two problems to improve the detection accuracy for RGB–infrared remote-sensing object detection. Specifically, the CLC module employs a calibration convolution to align RGB and infrared feature first and utilizes a selective aggregation module (SAM) to filter out irrelevant information and noises. The proposed CLC module can alleviate the misalignment problem in spatial positions and reduce the heterogeneity between the RGB and infrared features. In addition, the CGC module captures global contextual information in RGB and infrared features by mining the interactive information between two modalities, which can improve the detection accuracy.

In summary, the novelty, contribution, and characteristic of the proposed CLGNet are as follows.

1) We propose a novel network for remote-sensing object detection called the CLGNet. It consists of two key modules: the CLC module and the CGC modeling module. Our CLGNet can significantly improve the detection performance by addressing the spatial misalignment issues, reducing abundant information during feature fusion, and enriching cross-modal nonlocal information.
2) The proposed CLC module encompasses a calibration convolution and an SAM. This CLC module is adept at diminishing noise and eliminating irrelevant information during the process of feature aggregation. More specifically, the calibration convolution mitigates noises introduced by spatial offsets between modalities, while the SAM counteracts the presence of irrelevant information caused by the imbalance in modality reliability.
3) The CGC modeling module enhances the enrichment of cross-modal global contextual information, thereby facilitating the inference of object presence and locations under diverse illumination conditions. In addition, we provide experimental and theoretical evidence to demonstrate the superior performance of our CGC compared with single-modal global context modeling operation experimentally and theoretically.
4) Our CLGNet is a generic feature aggregation module that can be flexibly integrated into diverse detectors and consistently improves their detection accuracy (see Fig. 1). Moreover, our CLGNet achieves the state-of-the-art performace on our four widely used RGB–infrared object detection benchmarks (i.e., DroneVehicle, VEDAI, LLVIP, and KAIST).

The rest of this article is organized as follows. In Section II, we discuss the relevant literature and prior research in the field. We present our proposed CLGNet in Section III, detailing its key components. In Section IV, we present and analyze the experimental results. Finally, Section V concludes this article.

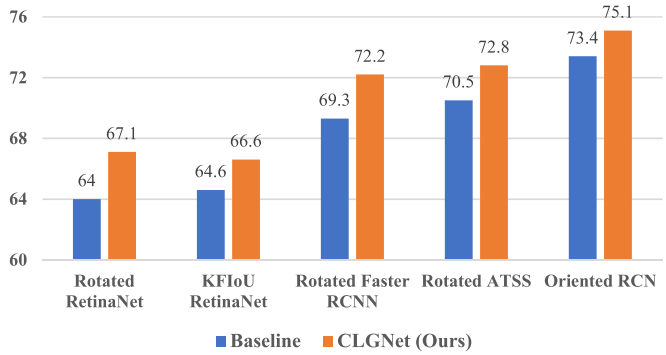


Fig. 1. Detection accuracy comparison on the DroneVechile *test* set. Our CLGNet is generic and obtains consistent improvements while being integrated into diverse detection frameworks.

II. Related Work

In this section, we begin by reviewing the object detection methods, remote-sensing-oriented object detection methods, and RGB–infrared object detection methods that form the basis of RGB–infrared remote-sensing-oriented object detection methods.

A. Object Detection

Object detection aims at classifying and locating objects in images. With the development of deep learning, object detection methods based on deep learning have achieved great success. Existing object detection methods based on deep learning can be roughly divided into two categories: anchor-based object detection methods [17], [18], [19], [20], [21] and anchor-free object detection methods [22], [23], [24], [25]. Anchor-based object detection methods predict the locations and categories of objects by deploying hand-crafted anchor boxes. Anchor-free object detection methods locate and classify objects without deploying default anchor boxes.

B. Remote-Sensing-Oriented Object Detection

Remote-sensing-oriented object detection [9], [26], [27], [28] aims to classify and locate the arbitrary-oriented objects in remote-sensing imagery in which the objects are often arbitrary-oriented, dense-distributed, and of small sized. Most existing oriented object detection methods are modified from general object detection methods [17], [20], [29] by additionally regressing the orientation. Instead of using rotated anchors to generate rotated region proposals, RoI transformer [6] introduces a rotated RoI learner to learn a rotated region of interest from a horizontal region of interest, which reduces the computational complexity. Oriented RCNN [9] directly generates oriented region proposals from the traditional axis-aligned anchors, which further decreases the computational cost. Yang et al. [30] proposed a fully differentiable KFIoU loss for oriented object detection. Huang et al. [31] designed a taskwise sampling convolution to extract specific features of classification and localization tasks for arbitrary-oriented remote-sensing object detection.

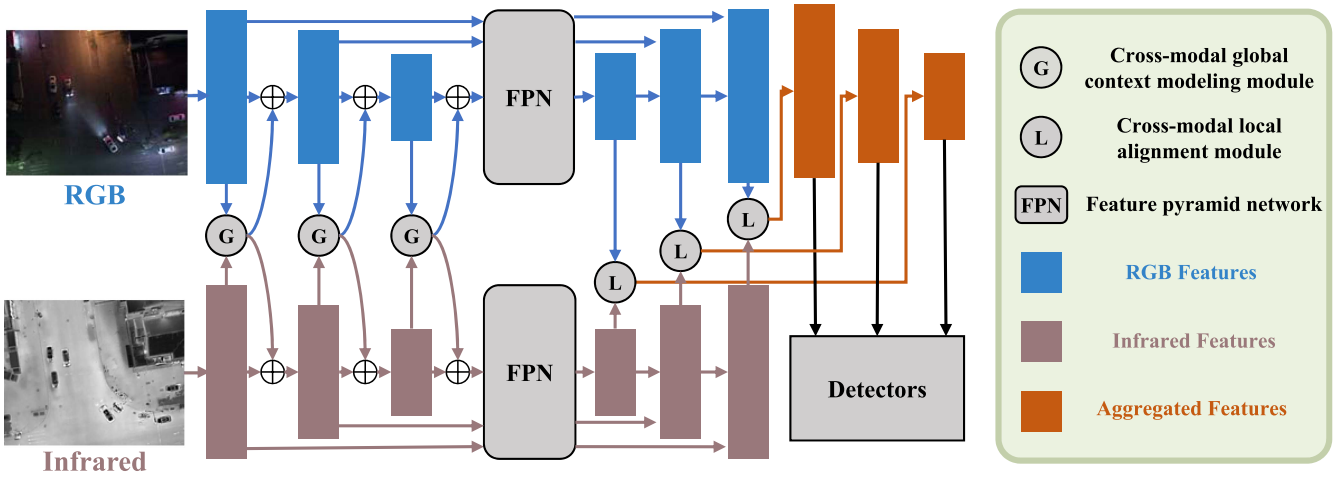


Fig. 2. Overall architecture of the proposed RGB–infrared object detection method. The focus of our proposed method is a novel CGC modeling module and a CLC module. The CGC module is embedded into the two-steam backbone networks to enrich the cross-modal global contextual information. The CLC module takes the RGB and infrared features from the feature pyramid network as inputs and aligns and aggregates RGB and infrared features adaptively.

C. RGB–Infrared Object Detection

RGB–infrared object detection [32], [33], [34], [35], [36], [37], [38] aims at classifying and locating objects around the clock by combining RGB images and infrared images, which serves as a fundamental approach for numerous other tasks, such as RGB–infrared person reidentification [39]. RGB images have rich textures in the daytime, while infrared images provide more effective information at night. RGB–infrared object detection aggregates the complementary information RGB images and infrared images to achieve better object detection accuracy.

D. RGB–Infrared Remote-Sensing-Oriented Object Detection

Sun et al. [13] collected paired drone-based RGB–infrared images from day to night and created the DroneVehicle dataset. Moreover, a UA-CMDet is proposed and fuses complementary information from RGB–infrared images by concatenation to improve the vehicle detection performance in low-light conditions. An uncertainty-aware module is proposed to predict three uncertainty weights for RGB, infrared, and fused modalities. The uncertainty weights are utilized to achieve illumination-aware cross-modal NMS to improve the detection accuracy. RISNet [14] suppresses the redundant information in cross-modal fusion by computing the information entropy and improves the vehicle detection performance in poor illumination conditions. Li et al. [16] proposed a cross-modal knowledge distillation strategy, including selective feature knowledge distillation and adaptive prediction knowledge distillation, to improve the accuracy of RGB–infrared remote-sensing object detection.

Limitation: Many of the existing methods [32], [33] aggregate features from different modalities using simple elementwise operations (e.g., addition or concatenation). These approaches often treat both modalities equally, potentially leading to the incorporation of irrelevant and distracting information from different modalities. Moreover, UA-CMDet [13] relies on RoI-based operations, which restricts its application to two-stage object detectors and limits its versatility. In contrast, our proposed CLGNet serves as a versatile feature aggregation network that can be seamlessly integrated into anchor-free, two-stage, and single-stage detectors. Furthermore, many existing methods [13], [14] overlook the importance of modeling long-range feature dependencies, which play a crucial role in understanding pedestrian presence and locations.

III. Methods

A. Overall Architecture

Fig. 2 shows the overall architecture of the proposed RGB–infrared object detection method. A two-stream backbone network integrated with our proposed CGC modeling module takes RGB and infrared images as input to extract features with enriched global contextual information. And then FPN [40] is employed to construct a feature pyramid to help detect the objects of various sizes. A novel CLC module is utilized to align features with spatial positions and aggregate multimodal features selectively. At last, the aggregated features go through a detector to predict the locations and categories of objects.

B. CGC Modeling Module

Modeling long-range dependencies to capture global contextual information has been demonstrated as an effective way of improving detection accuracy [41], [42]. GCNet [41] is a lightweight nonlocal operation to capture global contextual information, achieving great success in single-modal tasks. Despite its effectiveness in single-modal tasks, it fails to perform well in multimodal tasks. The main reason for the performance gap is the lack of consideration for the reliability of different modalities. The reliability of the modalities may vary with the changes in illumination conditions. Specifically, in the day, the infrared features are unreliable, leading that it is difficult to capture the global contextual information of the infrared features. For the same reason, it is difficult to capture the global contextual

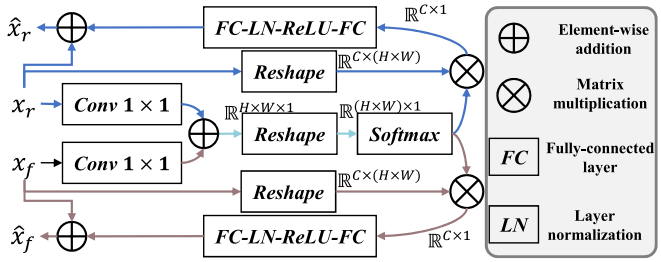


Fig. 3. Architecture of our proposed CGC modeling module.

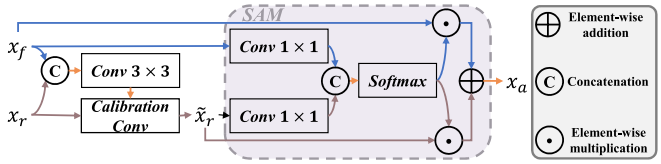


Fig. 4. Architecture of our proposed CLC module. The focus of our CLC module is the calibration convolution and an SAM, which is utilized to align and aggregate features, respectively.

information of RGB features at night because the RGB features are unreliable at night. To address this issue, we propose a novel CGC modeling module. Fig. 3 shows the architecture of the proposed CGC module.

We first pass the RGB features x_r and infrared features x_f through two separate convolutional layers whose kernel size is 1×1. Next, we add the outputs of these two layers together and reshape the resulting features to prepare for matrix multiplication. Finally, we apply a softmax operation to compute the global attention pooling weights w_p. The computational process can be represented by the following equations:

$$w_p = S(R(f^{1\times1}(x_r) + f^{1\times1}(x_f)) \tag{1}$$

where w_p is the global attention pooling weight, which is used to extract global contextual features. $f^{1\times1}$ is the convolutional layer whose kernel size is 1×1. S and R denote the softmax and reshape operations, respectively.

Then, the obtained global context weights are utilized to extract global contextual features. The detailed computation is given as follows:

$$\begin{aligned} g_r &= \mathrm{FC}(\sigma(\mathrm{LN}(\mathrm{FC}(R(x_r)w_p)))) \\ g_f &= \mathrm{FC}(\sigma(\mathrm{LN}(\mathrm{FC}(R(x_f)w_p)))) \end{aligned} \tag{2}$$

where g_r and g_f are the RGB and infrared global contextual features, FC denotes the full-connected layer, R represents the reshape operation, LN is the layer normalization, and σ is the ReLU activation function. It can be noted that the obtained global attention pooling weights w_p depend on both RGB and infrared features. This ensures that the RGB and infrared global contextual features (i.e., g_r and g_f) calculated by the weights w_p are robust and discriminative, regardless of the illumination level being high or low. Finally, the generated RGB and infrared global contextual features are as a residual part to be added into the input RGB and infrared features, respectively. The following equation provides a detailed computation of the process

$$\begin{aligned} \hat{x}_r &= x_r \oplus g_r \\ \hat{x}_f &= x_f \oplus g_f \end{aligned} \tag{3}$$

where $\oplus$ represents the broadcast elementwise addition, and $\hat{x}_r$ and $\hat{x}_f$ denotes the enhanced of the RGB and infrared features, respectively, which contains enriched global contextual information. These two features are used to replace the original features x_r and x_f as input to the subsequent layers of backbone networks.

Our proposed CGC module can be embedded into any stage of the backbone network (e.g., ResNet). In Section IV, we conduct experiments to analyze the effects of the embedded stages. According to the experimental results, we integrate our CGC module into Stage 3, Stage 4, and Stage 5 of ResNet.

C. CLC Module

The heterogeneity across different modalities would lead to the introduction of noise in the multimodal feature aggregation. And position shifts caused by camera acquisition stage would lead misalignment problems in spatial positions during multimodal feature aggregation. The above two issues make it difficult to obtain discriminative aggregated features. To alleviate the above problems, we propose a novel CLC module, which first employs a calibration convolution to align multimodal features, and then uses an SAM to reduce irrelevant information and noises during feature aggregation. In the next, we describe the details of our proposed CLC module (see Fig. 4).

The input of our CLC module is the features x_{r_i} and x_{f_i} extracted by the feature pyramid network, where i denotes the feature level of the feature pyramids. For each level feature, one CLC module is utilized to aggregate scale-specific multimodal features. For brevity, the feature level i is omitted in the following sections.

Calibration convolution: The calibration convolution is utilized to align multimodal features. The computational process can be denoted as follows:

$$\tilde{x}_r(p) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x_r(p_0 + p_n + \Delta(p_0)) \tag{4}$$

where p_0 denotes the spatial location of the features x_r, $\mathcal{R}$ is the set of sampled positions corresponding to a regular grid (e.g., for a convolution whose kernel size is 3×3, $\mathcal{R} = \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\}$), and w is denoted the weights of the convolution. The input and output features are denoted by x_r and $_r$, respectively. The kernel offsets $\Delta(p_0)$ are used to augment the regular sampling grid, which calibrates the features from the RGB and infrared branches. The computational process for the kernel offsets $\Delta(p_0)$ can be expressed using the following equations:

$$\Delta(p_0) = f^{3\times3}(x_r \| x_f) \tag{5}$$

where $f^{3\times3}$ denotes a convolution layer whose kernel size is 3×3. $\|$ denotes the concatenation operation. x_r and x_f are the input features of RGB branch and infrared branch, respectively.

The aligned RGB features $\tilde{x}_r$ and infrared features x_f are further used to aggregate together by the next SAM.

SAM: The input to the SAM consists of the infrared features x_f and the RGB features $\tilde{x}_r$, which are aligned with the infrared features using the above calibration convolution. Two parallel convolution layers followed by one softmax layer are utilized to predict the selective aggregation weights $a_{r/f}$. The computation process can be denoted by

$$\begin{aligned} z_r &= f^{1\times1}(\hat{x}_f \| x_r) \in \mathbb{R}^{H\times W\times 1} \\ z_f &= f_{1\times1}(\hat{x}_f \| x_r) \in \mathbb{R}^{H\times W\times 1} \\ a_r &= \frac{e^{z_r}}{e^{z_r}+e^{z_f}} \\ a_f &= \frac{e^{z_r}}{e^{z_r}+e^{z_f}} \end{aligned} \tag{6}$$

where $\|$ denotes the concatenation operation, $f^{1\times1}$ denotes the convolution layer whose kennel size is 1×1, and H and W are the height and width of the features. The selective aggregation weights $a_{r/f}$ refer to the spatial importance weights for the multimodal feature maps that being aggregated. These weights are learned through end-to-end training. Once the selective aggregation weights have been learned, they are utilized to aggregate multimodal features. Selective aggregation weights can be used to identify important modalities at specific spatial locations, filtering-out irrelevant information. The aggregation of multimodal features using the selective aggregation weights can be represented mathematically as follows:

$$x_a = x_r \cdot a_r + x_f \cdot a_f \tag{7}$$

where x_a is the aggregated features, and $\cdot$ denotes the elementwise multiplication. Finally, the resulting aggregated features are input into the following detection head network to predict the object categories and locations.

IV. Experiments

A. Datasets

Here, we conduct experiments on two RGB–infrared remote-sensing object detection datasets (DroneVehicle [13] and VEDAI [43]). In addition, to demonstrate the generality and effectiveness of our method, we also conduct experiments on two RGB–infrared pedestrian detection datasets (LLVIP [44] and KAIST [45]). In the next, we would describe the details of these four datasets.

DroneVehicle is a recently proposed RGB–infrared remote-sensing object detection dataset, providing oriented bounding-box annotations with five object categories (i.e., car, freight car, truck, bus, and van). The training, validation, and test sets contain 17 990, 1469, and 8980 RGB–infrared image pairs, respectively.

VEDAI is a popular RGB–infrared vehicle detection dataset in aerial imagery with nine vehicle categories (i.e., car, truck, pickup, tractor, camper, ship, van, plane, and other).

LLVIP is a challenging RGB–infrared pedestrian detection dataset, comprising 12 025 training images and 3463 test images.

KAIST is a popular multispectral pedestrian detection dataset. Following the state-of-the-art methods [35], [37], [46], our method is trained on the annotations provided by [46] and tested on the improved test annotations [32].

B. Evaluation Metrics

For the DroneVehicle dataset, we follow [13] and report results by utilizing the standard mean average precision (mAP) with an intersection over union (IoU) threshold of 0.5.

For the VEDAI dataset, we follow [15], [43] and adopt the tenfold cross-validation protocol. The standard mAP with an IoU threshold of 0.5 is chosen as the evaluation metric.

For the LLVIP dataset, we follow [37] and report results by using the COCO-style average precision AP [47] that is computed with averaged across IoU thresholds from 0.5 to 0.95 and with an interval of 0.05.

For the KAIST dataset, we use the standard log-average miss rates [48] as the performance metric, which has lower values for better detection performance.

C. Implementation Details

We utilize two popular PyTorch [49] based object detection toolboxes MMRotate[1] [50] and MMDetection[2] [51] to implement our method. We train our model on two NVIDIA GeForce RTX 3090 GPUs and set the batch size to 2 per GPU. For all experiments, the backbone networks used in our method are pretrained on the ImageNet dataset [52]. During the training stage, we train the model for a total of 12 epochs. At the 8th and 11th epochs, we decay the learning rate by a factor of 10.

On the DroneVehicle dataset, we follow [13] and remove the white border of the images. In addition, the image size is set to be 640×512 during both the training and test. The oriented bounding-box annotations on the infrared modalities are utilized to train and test our model.

On the VEDAI dataset, we set the image size to 1024×1024 for both training and testing.

On the LLVIP dataset, the image size is set to be 1280×1024 for both the training and test stages.

On the KAIST dataset, we use an image size of 640×512 for both training and testing stages.

We set the other experimental settings (such as learning rates and optimizer) by following the default settings in MMRotate and MMDetection.

D. Results on DroneVehicle

Baseline comparison: First, we demonstrate the effects of our proposed modules: CLC module and CGC modeling module on the DroneVehicle *val* set in Table I. For a fair comparison, the detector used for all methods in Table I is RotatedRetinaNet, and the backbone network is ResNet50. We utilize the elementwise addition as the baseline method for feature aggregation. The baseline achieves an mAP score of 65.0%. By integrating the CLC module, which aligns and aggregates the RGB and infrared features adaptively, we observe an improvement of 2.1%, resulting in an mAP score of 67.1%. In addition, the introduction of the CGC module, which enriches the cross-modal global

[1][Online]. Available: https://github.com/open-mmlab/mmrotate
[2][Online]. Available: https://github.com/open-mmlab/mmdetection

TABLE I
ANALYZING THE IMPACTS OF OUR PROPOSED CLC AND CGC MODULES ON THE DRONEVEHICLE *VAL* SET

Methods	mAP
Baseline	65.0
Baseline + CLC	67.1
Baseline + CGC	66.9
Our CLGNet: Baseline+CLC+CGC	**67.6**

All the methods use the same training data, input scale, and experimental settings.
Best results are boldfaced.

TABLE II
ANALYZING THE IMPACTS OF STAGES IN CGC ON THE DRONEVEHICLE *VAL* SET

Stage3	Stage4	Stage4	mAP
-	-	-	65.0
✓	-	-	65.7
-	✓	-	66.6
-	-	✓	66.4
✓	✓	✓	66.9

TABLE III
COMPARISON OF THE PROPOSED METHOD WITH THE BASELINE IN TERMS OF INFERENCE SPEED

Fusion methods	Inference speed(FPS)	mAP	
		val	*test*
Baseline	16.7	65.0	64.0
CLGNet	15.5	67.6	67.1

TABLE IV
COMPARISON [IN AVERAGE PRECISION (%)] WITH DIFFERENT FEATURE AGGREGATION METHODS ON THE DRONEVEHICLE *VAL* SET

Methods	mAP
GCNet [41]	65.7
CGC (ours)	66.9 (↑ 1.2)

contextual information, improves the detection accuracy from 65.0% to 66.9% mAP. Our novel CLGNet combining the CLC module and the CGC module provides a significant gain of 2.6% in mAP over the baseline. These experimental results serve as evidence to demonstrate the effectiveness of our proposed CLC and CGC modules.

Impacts of the integrating CGC module at different stages: Here, we conduct experiments to analyze the impacts of integrating the CGC module at different stages of backbone network ResNet on the DroneVehicle *val* set, and the results are given in Table II. We consider four situations: Stage 3 (i.e., conv3_x in ResNet), Stage 4 (i.e., conv4_x in ResNet), Stage 5 (i.e., conv5_x in ResNet), and the combination of the above stages. It can be observed that integrating our CGC module at any single stage can improve the detection accuracy. Additionally, integrating our CGC at all of Stage 3, Stage 4, and Stage 5 simultaneously can achieve higher detection accuracy than integrating our CGC module at a single stage. Therefore, we choose to integrate our CGC module at all three stages due to the better detection performance.

Computational complexity analysis: Here, we examine the computational complexity of our CLGNet. Table III presents the inference speed comparison between our CLGNet and the baseline feature fusion method (elementwise addition). The measurements in Table III are tested on a single NVIDIA RTX3090 GPU. To ensure a fair comparison, both our CLGNet and the baseline are integrated into the RotatedRetinaNet with ResNet50. When the input resolution is set to 640×512, compared with the baseline, our CLGNet can obtain absolute gains of 2.6% and 3.1% on the DroneVehicle *val* and *test* sets, respectively, while only reducing the speed by 1.2 FPS. It can be concluded that the proposed CLGNet obtains improved detection accuracy albeit with a slight increase in computational burden.

Comparison with the single-modal global context method: We compare our CLGNet with the single-modal global context method GCNet [41]. GCNet is widely recognized as one of the leading approaches to capture global context by modeling long-range dependencies. Instead of utilizing our proposed CGC modeling module, we employ the GCNet on the RGB and infrared backbone networks, individually. The experimental results are given in Table IV. It can be observed that compared with GCNet, our CGC improves the detection accuracy from 65.7% to 66.9% in terms of detection accuracy on the DroneVehicle *val* set. The superior detection performance dues to the following main reason: The GCNet only captures global contextual information within each modality separately, leading that the extracted global contextual information is susceptible to being influenced by variations in illumination. On the contrary, our CGC effectively models the global context by exploring the complementary information between RGB and infrared modalities. This enables us to capture discriminative global contextual information that remains robust regardless of illumination variations. The experimental results demonstrate the importance of exploring the complementary information between RGB and infrared modalities when extracting global contextual information.

Robustness on different detectors: Here, we integrate our CLGNet into different popular detectors: Rotated RetinaNet [20], oriented RCNN [27], rotated faster RCNN [17], RoI transformer [6], KFIoU RetinaNet [30], and rotated ATSS [53]. To ensure a fair comparison, all methods except for the detectors utilize the same backbone (ResNet50) and experimental settings. The results are presented in Table V. Compared with the baseline feature fusion method (elementwise addition), integrating our CLGNet into oriented RCNN, rotated faster RCNN, rotated RetinaNet, KFIoU RetinaNet, and rotated ATSS achieves the absolute gains of 1.7%, 2.9%, 3.1%, 2.0%, and 2.3% in terms of detection accuracy, on the DroneVehicle *test* set. Additionally, integrating our CLGNet into the above detectors outperforms integrating the baseline feature fusion method into these detectors with a significant margin on the DroneVehicle *val* set. The experimental results provide evidence of the effectiveness and versatility of our CLGNet.

Comparison with other feature fusion methods: We compare with our CLGNet with two recent attention-based feature aggregation methods (i.e., CMAFF [15] and CSSA [54]). The results are given in Table V. It can be observed that compared with other feature fusion methods, our CLGNet obtains consistent

TABLE V
COMPARISON [IN AVERAGE PRECISION (%)] OF OUR METHOD WITH DIFFERENT DETECTORS AND DIFFERENT FEATURE FUSION MANNERS ON THE DRONEVEHICLE *VAL* AND *TEST* SETS

Detectors	Fusion methods	mAP	
		val	*test*
Oriented RCNN [27]	Baseline	74.0	73.4
	CMAFF [15]	73.9	73.2
	CSSA [54]	73.9	73.2
	CLGNet	**75.5**	**75.1**
Rotated Faster RCNN [17]	Baseline	71.3	69.3
	CMAFF [15]	72.2	70.4
	CSSA [54]	72.2	70.5
	CLGNet	**73.6**	**72.2**
Rotated RetinaNet [20]	Baseline	65.0	64.0
	CMAFF [15]	66.1	64.8
	CSSA [54]	65.9	64.8
	CLGNet	**67.6**	**67.1**
KFIoU RetinaNet [30]	Baseline	65.1	64.6
	CMAFF [15]	65.8	65.1
	CSSA [54]	65.2	64.8
	CLGNet	**66.9**	**66.6**
Rotated ATSS [53]	Baseline	71.7	70.5
	CMAFF [15]	72.2	70.6
	CSSA [54]	71.7	70.1
	CLGNet	**74.1**	**72.8**

For a fair comparison, all the methods use the same backbone (ResNet50), training data, input scale, and experimental settings. Best results are boldfaced.

TABLE VI
STATE-OF-THE-ART COMPARISONS [IN TERMS OF DETECTION ACCURACY (%)] ON THE DRONEVEHICLE *VAL* AND *TEST* SETS

Methods	mAP	
	val	*test*
UA-CMDet [13]	70.8	70.3
CLGNet	73.1	71.5

improvements on all detectors, demonstrating the superiority of our CLGNet.

Comparison with the state-of-the-art methods: Here, we conduct experiments to compare our proposed with the state-of-the-art methods UA-CMDet [13] on the DroneVehicle *val* and *test* sets. The results are given in Table VI. The result of UA-CMDet is obtained by using the official code provided by the authors, and the detection accuracy is higher than the result reported in the original article. This improvement is from the updating made by the authors in the processes of data annotation and data preprocessing in the realized code.[3] For a fair comparison, we implement our CLGNet based on the code provided by UA-CMDet and utilize the same experimental settings (data processing process, learning schedule, optimizer, etc.) as the UA-CMDet. In addition, the detector of UA-CMDet is based on RoI-transformer [6] to guarantee the fairness, we also chose the RoI-transformer as the detector for our CLGNet. It can be observed that our CLGNet outperforms UA-CMDet [13] with consistent improvements of 2.3% and 1.2% on the *val* and *test* sets, respectively, demonstrating the superiority of our proposed CLGNet.

[3][Online]. Available: https://github.com/SunYM2020/UA-CMDet

TABLE VII
COMPARISON [IN AVERAGE PRECISION (%)] OF OUR METHOD WITH OTHER FEATURE FUSION MANNERS BY INTEGRATING THEM INTO FASTER RCNN ON THE VEDAI DATASET

Methods	Baseline	CMAFF [15]	CSSA [54]	**CLGNet**
mAP	77.9	79.0	78.2	**80.2**

Best results are boldfaced.

TABLE VIII
COMPARISON [IN AVERAGE PRECISION (%)] OF OUR METHOD WITH THE STATE-OF-THE-ART FEATURE FUSION MANNERS BY INTEGRATING THEM INTO VARIOUS DETECTORS ON THE LLVIP *TEST* SET

Detectors		Fusion methods	AP
Single-stage	RetinaNet	Concatenation	56.9
		CMAFF [15]	57.8
		CSSA [54]	57.1
		DCMNet [37]	58.9
		CLGNet	**60.4**
Two-stage	Cascade RCNN	Concatenation	59.6
		CMAFF [15]	61.1
		CSSA [54]	60.7
		DCMNet [37]	61.5
		CLGNet	**62.5**
Anchor-free	Reppoints	Concatenation	57.6
		CMAFF [15]	58.2
		CSSA [54]	59.0
		DCMNet [37]	58.7
		CLGNet	**59.9**

Best results are boldfaced.

Qualitative comparison: Fig. 5 illustrates a qualitative comparison of the input and output feature maps of our proposed CLC module. It can be observed that the input feature maps contain irrelevant information. After the feature aggregation, the aggregated features (the output of our CLC module) significantly enhance the discrimination of object regions from the background, thereby resulting in improved detection accuracy. The qualitative comparison demonstrates that our CLC can reduce the irrelevant information by aligning and aggregating the RGB and infrared features adaptively.

E. Results on VEDAI

Here, our CLGNet is compared with the following recent feature aggregation methods: elementwise addition (baseline) CMAFF [15] and CSSA [54]. To ensure fair comparisons, we use faster RCNN as the detector for all feature aggregation methods. The results are reported in Table VII. Our CLGNet achieves an average precision of 80.2%, which is higher than other feature aggregation methods.

F. Results on LLVIP

Here, we conduct experiments on the challenge RGB–infrared pedestrian detection dataset LLVIP [44]. The results are given in Table VIII. The detection performance of concatenation and DCMNet are reported in [37]. Compared with the elementwise operations (i.e., concatenation), integrating our CLGNet into

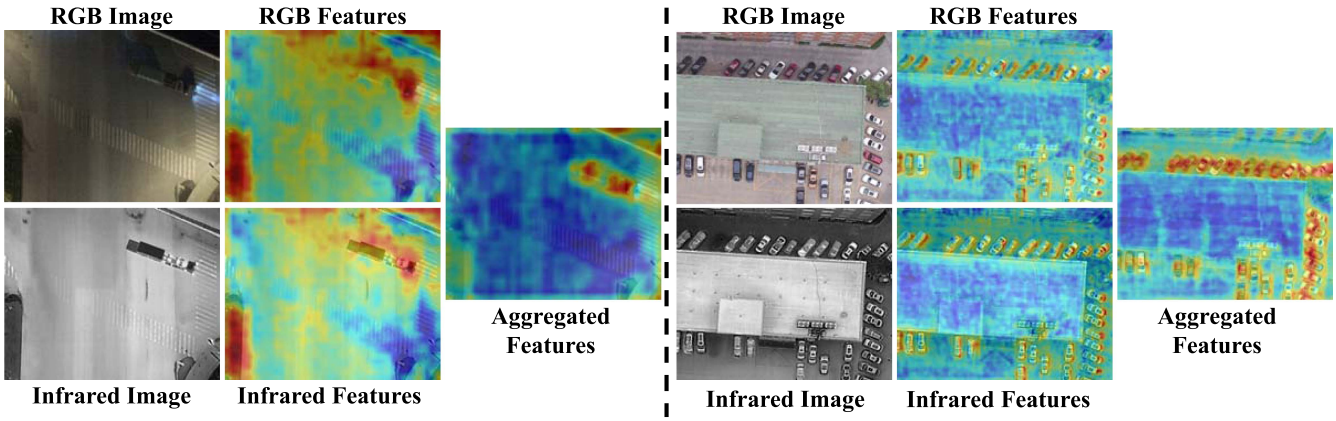


Fig. 5. Visualizations of the input and output feature maps of our CLC module.

RetinaNet [20], Cascade RCNN [55], and Reppoints [29] detectors yields improvements of 3.5%, 2.9%, and 2.3% in terms of average precision, respectively. The main reason for the improved performance of our CLGNet is attributed to its ability of capturing enriched global contextual information, aligning the spatial position between modalities, and balancing the modality importance.

In addition, integrating the state-of-the-art method DCMNet [37], which performs multimodal feature aggregation utilizing dynamic cross-modal modules that effectively mine local and nonlocal complementary information between modalities, into RetinaNet, Cascade RCNN, and Reppoints achieves AP scores of 58.9%, 61.5%, and 58.7%, respectively, as shown in Table VIII. Integrating out CLGNet into these detectors provides a superior detection accuracy of AP scores of 60.4%, 62.5%, and 59.9%, compared with DCMNet, respectively. It demonstrates that, compared with DCMNet, our CLGNet mines local and nonlocal complementary information more effectively, leading to superior detection accuracy.

To sum up, the experimental results on the LLVIP dataset show that integrating our CLGNet into different detectors yields improvements in the detection accuracy across various types of detectors, including single-stage, two-stage, and anchor-free detectors, demonstrating the generality and superiority of our method.

G. Results on KAIST

We compare our method with the following recent state-of-the-art methods: ACF [45], halfway fusion [32], IAF-RCNN [56], IATDNN+IAMSS [57], CIAN [34], MSDS-RCNN [33], AR-CNN [46], MBNet [35], BAANet [58], and UFF+UCG [59]. The results are given in Table IX. We choose the RetinaNet as the detector. It can be observed that our method outperforms the state-of-the-art methods on the *All*, *Day*, and *Night* sets. Among the existing methods, the UFF-UCG [59] reports log-average miss rates of 7.89%, 8.18%, and 6.96% on the *All*, *Day*, and *Night* sets, respectively. Our CLGNet achieves superior results with log-average miss rates of 6.67%, 7.48%, 4.80% on the *All*, *Day*, and ıNight sets, respectively. The results on the KAIST dataset show the superior and general performance of our CLGNet.

TABLE IX
COMPARISON OF THE PROPOSED AANET WITH THE STATE-OF-THE-ART METHODS IN TERMS OF THE LOG-AVERAGE MISS-RATE ON THE KAIST *TEST* SET

Methods	Backbone	All	Day	Night
ACF [45]	-	47.32	42.57	56.17
Halfway Fusion [32]	VGG	25.75	24.88	26.59
IAF-RCNN [56]	VGG	15.73	14.55	18.26
IATDNN+IAMSS [57]	VGG	14.95	14.67	15.72
CIAN [34]	VGG	14.12	14.77	11.13
MSDS-RCNN [33]	VGG	11.34	10.53	12.94
AR-CNN [46]	VGG	9.34	9.94	8.38
MBNet [35]	ResNet50	8.13	8.28	7.86
BAANet [58]	ResNet50	7.92	8.37	6.98
UFF+UCG [59]	ResNet50	7.89	8.18	6.96
CLGNet (ours)	VGG	**6.67**	**7.48**	**4.80**

Best results are boldfaced.

V. CONCLUSION

We have proposed the CLGNet as an effective feature aggregation network for RGB–infrared remote-sensing object detection. The CLGNet comprises a CLC module and a CGC modeling module. The CLC module employs a calibration convolution and an SAM to adaptively align and aggregate multimodal features by reducing spatial misalignment and effectively handling irrelevant information. The CGC module effectively captures global contextual information by exploring complementary information between RGB and infrared modalities, regardless of the illumination level being high or low. We conduct extensive experiments on one RGB–infrared remote-sensing object detection benchmark and one RGB–infrared pedestrian detection benchmark. The experimental results demonstrate that integrating CLGNet into various detectors can consistently improve detection accuracy, highlighting the effectiveness and superiority of our CLGNet.

REFERENCES

[1] H. Sun, X. Zheng, and X. Lu, "A supervised segmentation network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 30, pp. 2810–2825, Feb. 2021.

[2] X. Zheng, H. Sun, X. Lu, and W. Xie, "Rotation-invariant attention network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 4251–4265, May 2022.

[3] Y. Zhang, X. Zheng, Y. Yuan, and X. Lu, "Attribute-cooperated convolutional neural network for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8358–8371, Dec. 2020.

[4] X. Chen, X. Zheng, Y. Zhang, and X. Lu, "Remote sensing scene classification by local–global mutual learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Feb. 2022, Art. no. 6506405.

[5] W. Liu, K. Quijano, and M. M. Crawford, "YOLOv5-tassel: Detecting tassels in RGB UAV imagery with improved YOLOv5 based on transfer learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8085–8094, Sep. 2022.

[6] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2849–2858.

[7] J. Ding et al., "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7778–7796, Nov. 2022.

[8] C. Xu, X. Zheng, and X. Lu, "Multi-level alignment network for cross-domain ship detection," *Remote Sens.*, vol. 14, no. 10, 2022, Art. no. 2389.

[9] W. Li, Y. Chen, K. Hu, and J. Zhu, "Oriented reppoints for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1829–1838.

[10] W. Lu et al., "A CNN-transformer hybrid model based on CSWin transformer for UAV image object detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1211–1231, Jan. 2023.

[11] J. Xue, D. He, M. Liu, and Q. Shi, "Dual network structure with interweaved global-local feature hierarchy for transformer-based object detection in remote sensing image," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6856–6866, Aug. 2022.

[12] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Mar. 2023, Art. no. 5605415.

[13] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6700–6713, Oct. 2022.

[14] Q. Wang, Y. Chi, T. Shen, J. Song, Z. Zhang, and Y. Zhu, "Improving RGB-infrared object detection by reducing cross-modality redundancy," *Remote Sens.*, vol. 14, no. 9, 2022, Art. no. 2020.

[15] F. Qingyun and W. Zhaokui, "Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery," *Pattern Recognit.*, vol. 130, 2022, Art. no. 108786.

[16] A. Li et al., "Cross-modal object detection via UAV," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 10894–10905, Aug. 2023.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[18] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 43, no. 5, pp. 1483–1498, May 2021.

[19] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.

[21] J. Nie, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao, "Enriched feature guided refinement network for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9536–9545.

[22] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.

[23] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[24] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 765–781.

[25] J. Wang, L. Song, Z. Li, H. Sun, J. Sun, and N. Zheng, "End-to-end object detection with fully convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15849–15858.

[26] J. Han, J. Ding, N. Xue, and G.-S. Xia, "ReDet: A rotation-equivariant detector for aerial object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2786–2795.

[27] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3520–3529.

[28] Q. Ming, Z. Zhou, L. Miao, H. Zhang, and L. Li, "Dynamic anchor learning for arbitrary-oriented object detection," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2021, pp. 2355–2363.

[29] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 9657–9666.

[30] X. Yang et al., "The KFIoU loss for rotated object detection," 2022, *arXiv:2201.12558*.

[31] Z. Huang, W. Li, X.-G. Xia, H. Wang, and R. Tao, "Task-wise sampling convolutions for arbitrary-oriented object detection in aerial images," 2022, *arXiv:2209.02200*.

[32] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *Proc. Brit. Mach. Vis. Conf.*, 2016.

[33] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2018.

[34] L. Zhang et al., "Cross-modality interactive attention network for multispectral pedestrian detection," *Inf. Fusion*, vol. 50, pp. 20–29, 2019.

[35] K. Zhou, L. Chen, and X. Cao, "Improving multispectral pedestrian detection by addressing modality imbalance problems," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 787–803.

[36] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Guided attentive feature fusion for multispectral pedestrian detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 72–80.

[37] J. Xie et al., "Learning a dynamic cross-modal network for multispectral pedestrian detection," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 4043–4052.

[38] N. Chen, J. Xie, N. Jing, J. Cao, Z. Shao, and Y. Pang, "Attentive alignment network for multispectral pedestrian detection," in *Proc. 31th ACM Int. Conf. Multimedia*, 2023.

[39] X. Zheng, X. Chen, and X. Lu, "Visible-infrared person re-identification via partially interactive collaboration," *IEEE Trans. Image Process.*, vol. 31, pp. 6951–6963, Nov. 2022.

[40] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[41] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, 2019, pp. 1971–1980.

[42] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[43] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Vis. Commun. Image Representation*, vol. 34, pp. 187–203, 2016.

[44] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "LLVIP: A visible-infrared paired dataset for low-light vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2021, pp. 3496–3504.

[45] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1037–1045.

[46] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 5127–5137.

[47] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[48] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[49] A. Paszke et al., "Automatic differentiation in PyTorch," in *Proc. Int. Conf. Neural Inf. Process. Syst. Workshop*, 2017.

[50] Y. Zhou et al., "MMRotate: A rotated object detection benchmark using PyTorch," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 7331–7334.

[51] K. Chen et al., "MMDetection: Open mmlab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.

[52] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[53] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9759–9768.
[54] Y. Cao, J. Bin, J. Hamari, E. Blasch, and Z. Liu, "Multimodal object detection by channel switching and spatial attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2023, pp. 403–411.
[55] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.
[56] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *Pattern Recognit.*, vol. 85, pp. 161–171, 2019.
[57] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Inf. Fusion*, vol. 50, pp. 148–157, 2019.
[58] X. Yang, Y. Qian, H. Zhu, C. Wang, and M. Yang, " BAANet: Learning bidirectional adaptive attention gates for multispectral pedestrian detection," in *Proc. Int. Conf. Robot. Autom.*, 2022, pp. 2920–2926.
[59] J. U. Kim, S. Park, and Y. M. Ro, "Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1510–1523, Mar. 2022.

Jin Xie received the Ph.D. degree in information and communication engineering from Tianjin University, Tianjin, China, in 2021.

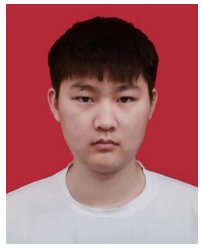
He is currently an Associate Professor with Chongqing University, Chongqing, China. His research interests include machine learning and computer vision, in which he has authored or coauthored 10+ papers in CVPR, ICCV, ECCV, IEEE TPAMI, IEEE TIP, and IEEE TMM.

Jing Nie received the Ph.D. degree in information and communication engineering from Tianjin University, Tianjin, China, in 2022.

She is currently an Assistant Researcher with Chongqing University, Chongqing, China. Her research interests include visual perception and image restoration.

Bonan Ding received the B.S. degree in computer science and technology from Northeast Petroleum University, Heilongjiang, China, in 2022. He is currently working toward the M.S. degree in electronic information with Chongqing University, Chongqing, China.

His research interests include computer vision and deep learning.



Mingyang Yu received the B.S. degree in network engineering from Henan University, Kaifeng, China, in 2022. He is currently working toward the M.S. degree in electronic information with Chongqing University, Chongqing, China.

His research interests include machine learning and computer vision.

Jiale Cao (Member, IEEE) received the Ph.D. degree in information and communication engineering from Tianjin University, Tianjin, China, in 2018.

He is currently an Associate Professor with Tianjin University. He has authored or coauthored more than 20+ IEEE Transactions and CVPR/ICCV/ECCV articles in his research areas. His research interests include object detection and image analysis. He serves as a Regular Program Committee Member for leading computer vision and artificial intelligence conferences, such as CVPR, ICCV, and ECCV.