IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 2, 2022, accepted May 11, 2022, date of publication May 16, 2022, date of current version May 24, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3175303

Attention-Based Cross-Modality Feature Complementation for Multispectral Pedestrian Detection

QUNYAN JIANG, JUYING DAI, TING RUI, FAMING SHAO,
JINKANG WANG, AND GUANLIN LU
Department of Mechanical Engineering, College of Field Engineering, Army Engineering University of PLA, Nanjing 210007, China
Corresponding author: Juying Dai (dinajy2001@126.com)

This work was supported by the China National Key Research and Development Program under Grant 2016YFC0802904.

ABSTRACT Multispectral pedestrian detection based on deep learning can provide a robust and accurate detection under different illumination conditions, which has important research significance in safety. In order to reduce the log-average miss rate of the object under different illumination conditions, a new one-stage detector suitable for multispectral pedestrian detection is proposed. First, in order to realize the complementarity between the information flows of the two modalities in the feature extraction stage to reduce the object loss, a low-cost cross-modality feature complementary module (CFCM) is proposed. Second, in order to suppress the background noise in different environments and enhance the semantic information and location information of the object, so as to reduce the error detection of the object, an attention-based feature enhancement fusion module (AFEFM) is proposed. Thirdly, through the feature complementarity of color-thermal image pair and the multi-scale fusion of depth feature layer, the horizontal and vertical multi-dimensional data mining of parallel deep neural network is realized, which provides effective data support for object detection algorithm. Finally, through the reasonable arrangement of proposed modules, a robust multispectral detection framework is proposed. The experimental results on the Korea Advanced Institute of Science and Technology (KAIST) pedestrian benchmark show that the proposed method has the lowest log-average miss rate compared with other state-of-the-art multispectral pedestrian detectors, and has a good balance in speed and accuracy.

INDEX TERMS Attentional mechanism, multispectral pedestrian detection, multimodal feature fusion, one-stage object detection.

I. INTRODUCTION

Pedestrian detection is a key task in various computer vision applications, such as autonomous driving [1], video monitoring [2], etc. In recent years, deep learning has achieved a great improvement in pedestrian detection. However, most of the detectors are based on RGB images, and the detection performance of this kind of detector is often significantly degraded under insufficient lighting conditions and severe weather conditions. Thus, it is difficult to accurately detect pedestrians only by relying on the visible light image. As a promising alternative for overcoming the limited accuracy of a single-visible image approach, multispectral pedestrian detection has been actively studied and attracted more attention, especially in the field of safety.

Different modal information collected by color and thermal sensors provides complementary visual cues for pedestrian detection. RGB images can provide rich color, texture, and other detailed information, while the thermal images can filter out the interference of some environments and provide clearer contour information in case of insufficient lighting [3]. Therefore, multispectral pedestrian detection based on color and thermal images can provide robust detection under different lighting conditions. Integrating color and thermal images in multispectral pedestrian detection is a key task. Most feature fusion methods adopted a simple addition of two modal features [4] or the dimension reduction after feature map concatenation [5]. These methods did not consider

The associate editor coordinating the review of this manuscript and approving it for publication was Gianluigi Ciocca.

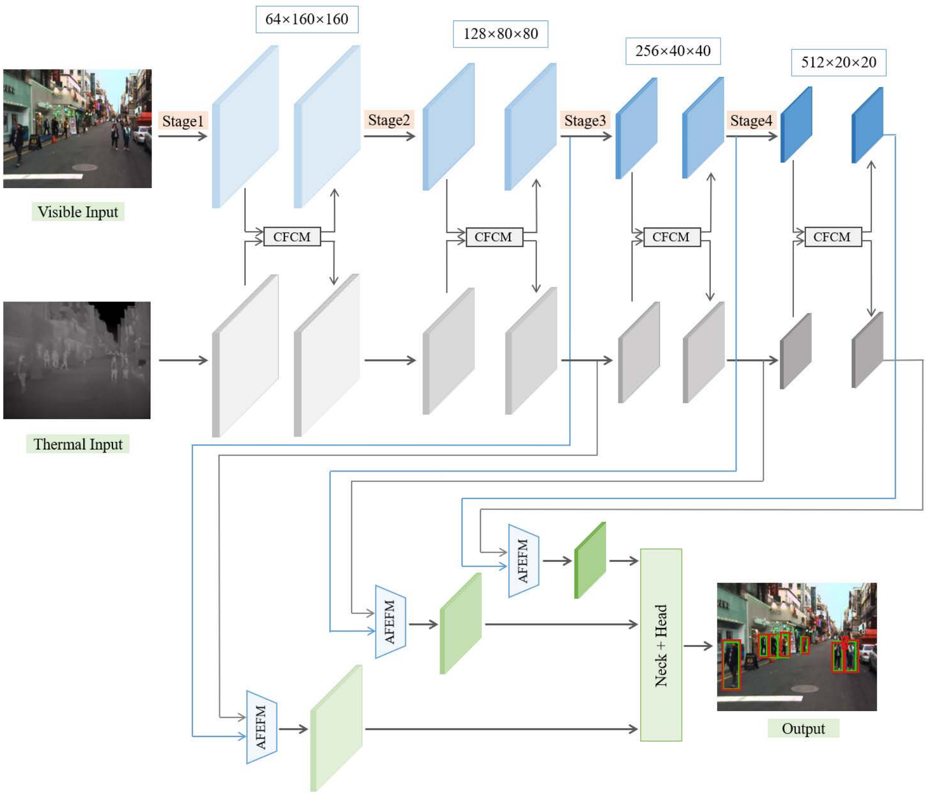


FIGURE 1. Overview of our framework. The feature extraction module adopts CSPDarknet53 [54] and embeds the CFCM module to supplement modality information. The AFEFM module is used to fuse the features of two modalities.

the complementarity between the two modal features and its own importance of each part of the feature map under a single modality [6]. Further, many pedestrian detection methods used a two-stage object detector to ensure detection accuracy. However, such a two-stage object detector sacrificed the inference speed, which is also crucial in practical applications, such as autonomous driving. To address the above problems, this paper proposes a new one-stage detector suitable for multispectral pedestrian detection, as shown in Fig. 1. Firstly, we adopt CSPDarknet53 as the backbone network to extract the multi-scale features under RGB and thermal modalities, respectively. Note that we chose to adopt a one-stage object network to ensure the real-time detection of objects. Also, we propose a CFCM module to realize the complementarity between the information flows of the two modalities in the feature extraction stage to reduce the object loss. Next, the AFEFM module fuses the features of different layers extracted from the two modalities to enhance the feature information of the detected object. Then, we perform a multi-scale fusion of the output features of the AFEFM module to obtain the input of the detection head. Finally, the detection head predicts the object based on the features of the input.

The main contributions of this paper are as follows:

- First, we propose the CFCM module to realize cross-modality complementation of features in the feature extraction stages. It promotes the information interaction between the two modalities, so that the lost object in one modality can be partially recovered through the complementary information provided by the other modality, so as to reduce the object loss and transmit more information.
- Second, we propose the AFEFM module to fuse the same level of depth information in the parallel network of visible and thermal modalities, so that the visual saliency of the two modal images can be mapped to the depth features. This module can suppress interference information and enhance object features.
- Thirdly, through the feature complementarity of color and thermal images and the multi-scale fusion of depth feature layer, we realize the horizontal and vertical multi-dimensional data mining of parallel network, fully enrich the depth semantic information of the object, and provide effective data support for object detection algorithm.
- Finally, a new cross-modal network based on YOLOv5 is proposed through the reasonable arrangement of proposed modules, which is named YOLO_CMN.

The rest of this paper is structured as follows. Section II introduces the related work. Section III describes the proposed method in detail. The experimental results and analysis are given in Section IV. Section V concludes this paper and puts forward the future work.

II. RELATED WORKS

Pedestrian detection has been actively studied due to its significance in many fields. Especially, multispectral pedestrian detection has attracted much attention due to the robust detection performance in different lighting environments. According to the feature extraction method, existing pedestrian detection methods are categorized into traditional and deep learning methods. The traditional way [8]–[11] has the limitations of relying on the hand-crafted design and low detection accuracy. In recent years, inspired by the rapid advance of deep learning in other computer vision tasks, multispectral pedestrian detection networks have also achieved a great improvement.

Since a single spectral object detection is not available for multispectral sensors, a naive approach for multispectral pedestrian detection separately trained and fused different spectral images. In order to improve detection accuracy, four common frameworks [5] were proposed: early fusion, halfway fusion, late fusion, and score fusion. The proposed fusion framework was based on Faster R-CNN [13], a common two-stage object detection algorithm. Many other frameworks were proposed based on Faster R-CNN, such as ConvNets, SAF R-CNN [14], and IAF RCNN [15]. Although two-stage detection has a better detection effect, the detection speed is often low to be used in many practical applications. As an alternative to two-stage detection for real-time pedestrian detection, one-stage detection networks were studied. In [16], an effective one-stage object detection network, YOLO_TLV, was proposed to achieve real-time detection with little accuracy reduction. In order to balance the accuracy and speed of the multispectral pedestrian detection network, [17] proposed an MSFFN fusion network based on the YOLOv3 network. In addition to YOLO, SSD is also one of the typical one-stage object detection networks. GFD-SSD [18] realized the balance between detection accuracy and speed. In addition, MSDS RCNN [12] was proposed, which can be learned by jointly optimizing pedestrian detection and semantic segmentation task.

The feature fusion method is a key task to achieve good detection performance in multispectral pedestrian detection. MIN fusion method [5] was proposed to reduce the dimensions of multimodal features with a 1 × 1 convolution layer after concatenation. SUM fusion [4] method adopted element-wise sum, which can be considered as the linear feature fusion with the same weight. According to the experiment, the contribution of different modal features to the detection results is varied under different lights. Thus, the linear feature fusion is not suitable for all cases. Considering this factor, [6], [15], [20] used the illumination-aware fusion methods. Reference [18] introduced gated fusion units in the middle layer of SSD for feature fusion and pedestrian detection. In recent years, the attention mechanism has received extensive attention due to its promising performance in many networks. In [21] and [22], an attention-based feature fusion module was adopted to improve multispectral pedestrian detection.

Previous works showed that color image and thermal image information are complementary, and the deep learning-based neural network integrates multispectral features to improve pedestrian detection performance. Although the existing multispectral pedestrian detection methods have achieved high performance, there is still a big gap between detector performance and human vision. In this paper, a new network architecture is proposed to reduce the log-average miss rate and improve detection speed.

III. OUR APPROACH

A. CROSS-MODALITY FUSION FRAMEWORK

In order to meet the requirements of multispectral pedestrian detection for detection accuracy and speed, we propose a cross-modality fusion framework based on the YOLOv5. YOLOv5 is a one-stage object detection network that consists of four parts: Input, Backbone, Neck, and Output. The backbone network extracts features and comprises three modules: Focus module, CSP module, and CBL module. The neck network is used for multi-scale feature fusion, which enables the network to detect objects with different scales, thus improving detection accuracy. Four variants of YOLOv5 are available according to the depth and width of the network structure. Due to the integration of many excellent algorithms, YOLOv5 can provide high detection accuracy and speed despite being a one-stage detection network.

The proposed network adopts YOLOv5s as the base network to achieve fast and accurate multispectral pedestrian detection. RGB images and thermal images are fed into two sub-networks as the input. The shallow network generates simple geometric features, while the deep network generates rich semantic information. Thus, the feature fusion at different locations has different effects. We first introduce the commonly used multispectral feature fusion framework: input fusion, early fusion, and halfway fusion. Then, we construct a cross-modal network based on YOLOv5s. As depicted in Fig. 2, the fusion framework uses both the AFEFM module and the CFCM module to obtain a better detection effect.

As all modules are integrated into the network and trained end-to-end with the loss function defined as follows:

$$Loss = a \times L_{conf} + b \times L_{local} + c \times L_{cls} \tag{1}$$

where L_{local} represents localization loss, L_{conf} represents confidence loss, and L_{cls} represents classification loss. The *Loss* is the weighted sum of three losses, in which L_{local} is calculated by CIOU_Loss, and L_{conf} and L_{cls} are calculated by BCE_Loss.

Input fusion is the data fusion of the RGB image and the thermal image before feature extraction. Using the AFEFM module, the color and thermal images are fused before being fed into the feature extraction network. Its structural form is the simplest, and the specific structure is shown in Fig. 2 (a).

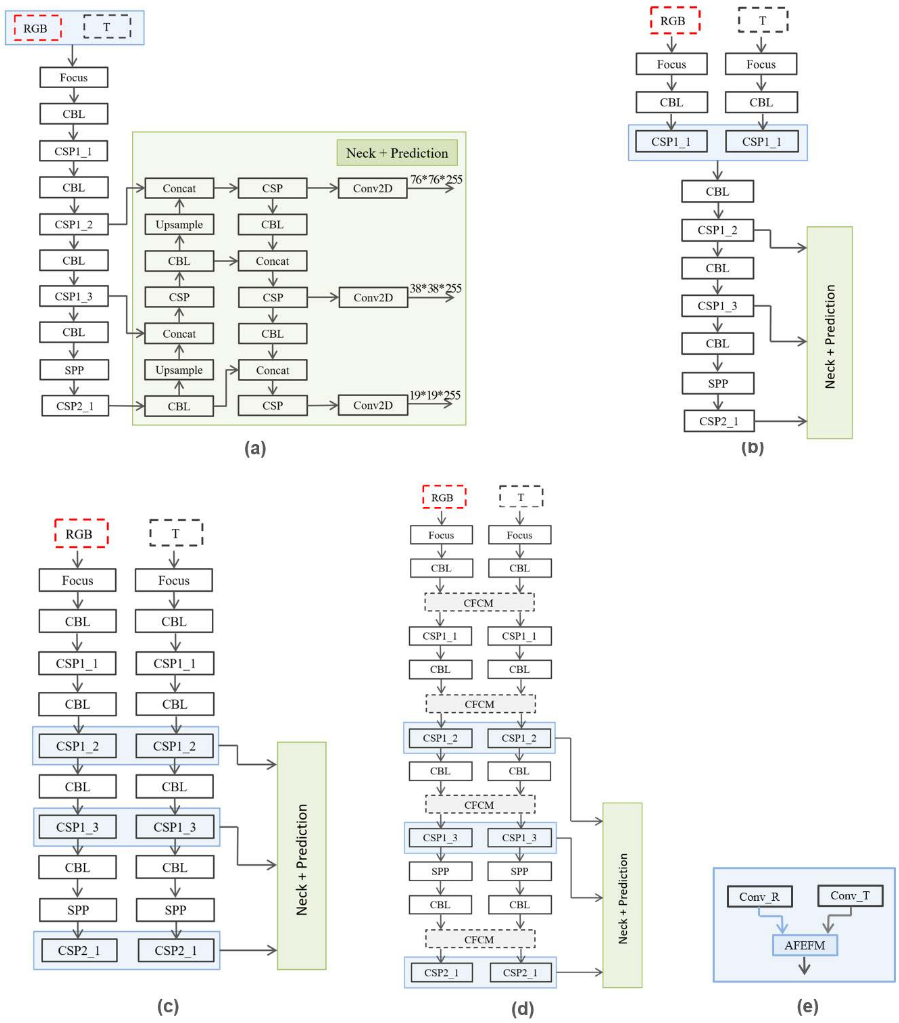


FIGURE 2. Cross-modality fusion framework, CFCM is a cross-modality feature complementary module, and AFEFM is a attention-based feature enhancement fusion module. (a) Input Fusion, (b) Early Fusion, (c) Halfway fusion, (d) YOLO_CMN (our), (e) The blue area represents the feature fusion method of the two modalities.

Early fusion integrates the RGB and thermal information via low-level feature fusion. RGB image and thermal image are respectively fed into two sub-networks for feature extraction. The RGB feature map F_R (128 channels) is extracted after the CSP1_1 module operation. The thermal feature map F_T (128 channels) is also extracted after the CSP1_1 module operation. The AFEFM module fuses those two features and generates the fused feature map whose channel size is also 128. The fused feature map passes through the remainder of the network. The specific network structure is shown in Fig. 2(b).

Halfway fusion deploys the AFEFM modules behind CSP1_2, CSP1_3, and CSP2_1, respectively. The fused features are sent to the Neck network for multi-scale feature fusion. The specific network structure is shown in Fig. 2(c). Since this fusion architecture showed good performance in the experiment, it was selected as the benchmark.

YOLO_CMN refers to the multispectral pedestrian detection architecture proposed in this paper. The TOYO_CMN consists of the halfway fusion structure and the added CFCM modules after the CBL modules. The network structure is depicted in Fig. 1(d). CFCM module makes the two modalities complement each other in the feature extraction stage so that the network can learn more information and reduce the object loss. AFEFM module enhances the important features of the two modalities while suppressing less important features before fusion. Then, they are fed into the Neck network for multi-scale feature fusion to obtain final prediction scores. The two modules are described in the following sections.

B. CROSS-MODAL FEATURE COMPLEMENTARY MODULE

RGB images have detailed information and contour information under good illumination conditions. On the other hand, the thermal image is less disturbed by illumination conditions and has clear pedestrian contour information. Inspired by the thought of differential modality information in [6], we proposed the CFCM module, which suppressed and enhanced its own modal information while integrating

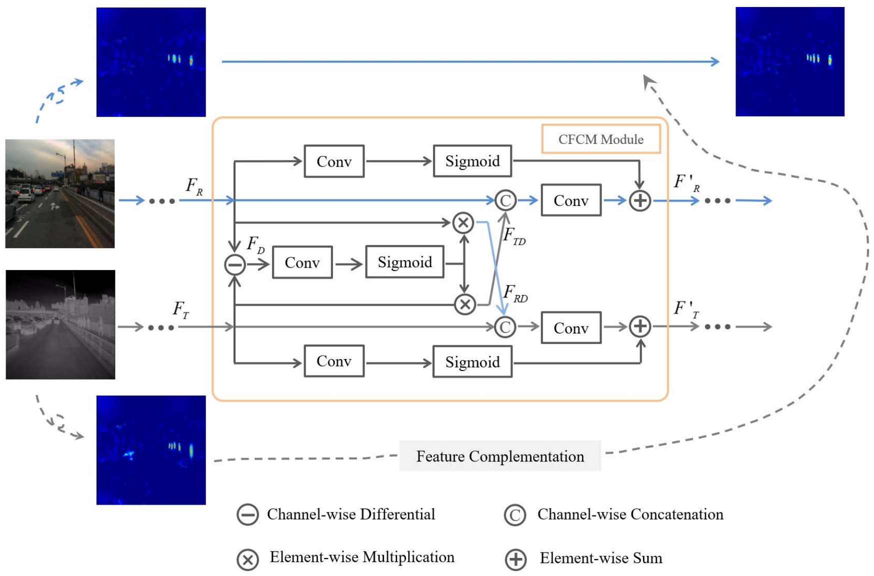


FIGURE 3. **CFCM module. CFCM first complements the features of the two modalities, and then uses the attentional mechanism to enhance the features of the different modalities.**

other modal features. In order to enable the detection network to realize all-day pedestrian detection, the CFCM module makes the features of two modalities interact, so that the lost object in one modality can be partially recovered through the complementary information provided by the other modality, so as to reduce the object loss and transmit more information. The CFCM module takes the different modalities into another modality to obtain complementary information of the other modality during feature extraction. In this way, the two modalities can learn more complementary features.

The CFCM module operates as follows. First, channel-wise differential weighting is used to get the difference-features of the two modal feature maps. Second, the different features of different modalities are amplified and fused with the features of another modality. Finally, in order to make the network pay attention to important features, channel attention operation is performed on the feature maps of the two modalities, and features are fused.

$$F_D = F_R - F_T \tag{2}$$

where $F_R \in \mathbb{R}^{C\times H\times W}$ is RGB convolution feature map, and $F_T \in \mathbb{R}^{C\times H\times W}$ is thermal convolution feature map. $F_D \in \mathbb{R}^{C\times H\times W}$ is obtained by channel-wise differential weighting for the two features.

$$F_{TD} = \sigma\left(GAP\left(F_D\right)\right) \odot F_T \tag{3}$$

$$F_{RD} = \sigma\left(GAP\left(F_D\right)\right) \odot F_R \tag{4}$$

where GAP refers to global average pooling, σ is the sigmoid activation function, and $\odot$ represents element-wise multiplication. $F_{TD} \in \mathbb{R}^{C\times H\times W}$ and $F_{RD} \in \mathbb{R}^{C\times H\times W}$ are obtained after feature enhancement, suppression, and fusion with F_R.

$$F'_T = \mathcal{F}\left(F_T||F_{RD}\right) \oplus \sigma\left(GAP\left(F_T\right)\right) \tag{5}$$

$$F'_R = \mathcal{F}\left(F_R||F_{TD}\right) \oplus \sigma\left(GAP\left(F_R\right)\right) \tag{6}$$

where || denotes the channel-wise concatenation operation, $\oplus$ represents element-wise sum, $\mathcal{F}(\cdot)$ is the residual function. $F'_T \in \mathbb{R}^{C\times H\times W}$ is the fusion of $\mathcal{F}\left(F_T||F_{RD}\right) \in \mathbb{R}^{2C\times H\times W}$ and F_T after feature enhancement. $F'_R \in \mathbb{R}^{C\times H\times W}$ is the fusion of $\mathcal{F}\left(F_R||F_{TD}\right) \in \mathbb{R}^{2C\times H\times W}$ and F_R after feature enhancement. After the above operation, the information of the two modalities is complementary and fused, and the obtained feature map containing more information is sent to the network for further feature extraction.

C. ATTENTION-BASED FEATURE ENHANCEMENT FUSION MODULE

Two modal features extracted via two sub-networks are fused with the AFEFM module, where important features are enhanced while suppressing noise interference based on the attention mechanism.

Global average pooling is often used to encode global spatial information in channel attention, which compresses features from 3-dimensions to 1-dimension, losing spatial information. We decompose global average pooling along with horizontal and vertical coordinates to pay attention to spatial information when using the channel attention module. Further, max-pooling can gather another important clue about the features of different objects to infer finer channel attention. Thus, both the average pooling and max pooling are employed in the proposed network.

$$V_c^h(h) = \frac{1}{W}\sum_{0\leq i<W} x_c\left(h, i\right) + Max\left(x_c\left(h\right)\right) \tag{7}$$

$$V_c^w(h) = \frac{1}{W}\sum_{0\leq j<H} x_c(j, w) + Max\left(x_c\left(w\right)\right) \tag{8}$$

where $V_c^h(h) \in \mathbb{R}^{C\times W\times 1}$ and $V_c^w(h) \in \mathbb{R}^{C\times 1\times H}$ are feature maps generated along with the horizontal and vertical directions, respectively. $\frac{1}{W}\sum_{0\leq i<W} x_c(h,i)$ refers to the average pooling of the h-th row of the c-th channel. $Max\left(x_c(h)\right)$ represents the max-pooling of the h-th row of the c-th channel. $\frac{1}{W}\sum_{0\leq j<H} x_c(j,w)$ is the average pooling of the w-th column of the c-th channel. $Max\left(x_c(w)\right)$ refers to the max-pooling of the w-th column of the c-th channel. A pair of direction-aware feature maps are generated by aggregating the features of the above two transformations and the two spatial directions, respectively.

$$f=\delta\left(F_1\left(\left[V^h,V^w\right]\right)\right) \tag{9}$$

$$R^h=\sigma\left(F_h\left(f^h\right)\right) \tag{10}$$

$$R^w=\sigma\left(F_w\left(f^w\right)\right) \tag{11}$$

where $[\cdot,\cdot]$ represents the join operations along spatial dimensions, δ is a non-linear activation function, $F_1(\cdot)$, $F_h(\cdot)$, and $F_w(\cdot)$ are 1×1 convolutional function. $f\in\mathbb{R}^{C/r\times(H+W)}$ is intermediate features map. $f^h\in\mathbb{R}^{C/r\times H\times 1}$ and $f^w\in\mathbb{R}^{C/r\times 1\times W}$ divide f into two independent tensors along with the horizontal and vertical directions. The outputs R^h and R^w are expanded and used as attention weights, respectively. Each output element is as follows:

$$y_c(i,j)=x_c(i,j)\times R_c^h(i)\times R_c^w(j) \tag{12}$$

F_R and F_T are fed into the attention module to obtain y_{TR}. F_{MLP}^R and F_{MLP}^T are obtained by two multi-layer perceptron (MLP) networks. The feature fusion of the two modalities is conducted as follows:

$$F'_{RT}=\mathcal{F}\left(\left(F_R\oplus y_R\right)||\left(F_T\oplus y_T\right)\right) \tag{13}$$

IV. EXPERIMENTAL RESULTS

This section describes the dataset, evaluation indicators, and implementation details of the experiment. Then, the performance of the proposed method is evaluated in comparison with other state-of-the-art methods. Lastly, the ablation studies are given on the two different modules and model architectures.

A. DATASET

The evaluation was conducted on the commonly used multispectral pedestrian dataset, the KAIST dataset. KAIST dataset collected 95,328 color-thermal image pairs using visible light camera and infrared thermal imaging camera, including 50,172 pairs in the training set and 45,156 pairs in the test set. According to the sampling principle in [7] and [12], every two frames were sampled from the training video, and heavy occluded and small person instances were removed (<50 pixels). The resulting training set contains 7,601 color-thermal image pairs. The test set was sampled every 20 frames, resulting in 2,252 color-thermal image pairs. Annotations of the KAIST dataset have been improved for the training set and test set. The dataset contains objects with different light conditions, different scales, and different degrees of occlusion, which is difficult to detect.

B. EVALUATION METRICS

The evaluation standard proposed in [28] is widely used in pedestrian detection–log-average miss rate (MR). Specifically, the generated detection bounding box (BB_d) by the model is compared to the ground truth bounding box (BB_g). A greater IOU than the threshold indicates that BB_d and BB_g match. IoU is defined as follows:

$$IoU=\frac{area\left(BB_d\cap BB_g\right)}{area\left(BB_d\cup BB_g\right)} \tag{14}$$

Unmatched BB_d is marked as false positives (FP), unmatched BB_g is marked as False Negatives (FN), and matched BB_d and BB_g as True Positives (TP). Miss rate is defined as the ratio of the total missed samples to all positive samples.

$$Miss\ Rate==\frac{FN}{FN+TP} \tag{15}$$

Let denote Num(img) be the number of pictures in the test set. Then, False Positives Per Image (FPPI) is expressed by the following formula:

$$\text{FPPI}=\frac{FP}{Num(img)} \tag{16}$$

According to the formula, the smaller the MR of the algorithm, the better the network detection performance.

C. IMPLEMENTATION DETAILS

Our method uses the same environment configuration as YOLOv5. Before training the model on the KAIST dataset, K-means clustering is applied to obtain anchors. The obtained anchor boxes after clustering are (18,42), (23,55), (34,84), (48,114), (59,141), (84,205), (110,259), (142,367), and (207,498). The proposed model was trained using a random gradient descent optimizer. The learning rate at the beginning of training was 0.001. When the training loss was not reduced, and the validation recall was not improved, the learning rate was reduced by ten times. After reducing the learning rate twice, the training stopped. All models were trained on GeForce RTX 2080Ti GPU with batch size = 8. We kept all hyperparameters the same as the original settings of the YOLOv5 model. In order to make a fair comparison, the MR is used to evaluate the performance of different models.

D. COMPARISON WITH THE STATE-OF-THE-ART METHODS

The proposed network is compared with state-of-the-art methods on the KAIST test dataset, including ACF+T+THOG, Halfway Fusion, YOLO_TLV, Fusion RPN+BDT, IAF R-CNN, IATDNN+IASS, MSDS-RCNN, AR-CNN, and MBNet. Among these detection methods, YOLO_TLV, MBNet, and our method are the one-stage methods, and the rest are the two-stage methods. The experimental results are depicted in Fig. 7, which shows that our detection method is the best and has the lowest MR. The MR values are 7.85%, 8.03% and 7.82%, respectively. Due to the

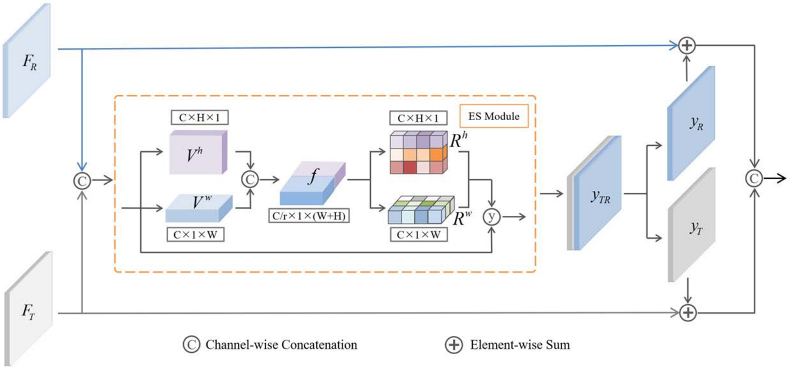


FIGURE 4. AFEFM fusion module. AFEFM sends the features of the two modalities into the ES module after channel-wise concatenation operation, and captures the horizontal and vertical attention features at the same time. Then, the output features after feature enhancement and suppression are separated by multi-layer perceptron network, and the two output features are obtained by element-wise sum operation with F_R and F_T respectively. Finally, the two output features are connected and convoluted to complete the feature fusion of the two modalities.

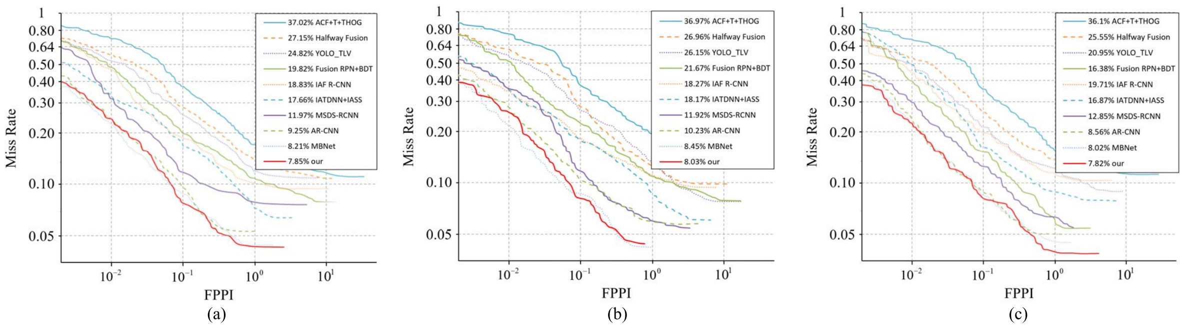
base network YOLOv5 and the proposed fusion method, the proposed network can provide good detection performance despite being a one-stage detection model. Although the proposed network is a one-stage detection model, it can still provide good detection performance due to the effectiveness of our proposed methods.

Compared with YOLO_TLV, the MR values are reduced by 16.97%, 18.12%, and 13.13%, respectively. Compared with MBNet, the MR values are reduced by 0.36%, 0.42%, and 0.2%, respectively. The proposed method shows the outperforming detection performance over the compared networks. In Fig. 5 (b)(c), we can observe that our method achieves good performance under reasonable settings during the day and night, and our method works better at night than during the day. This shows that the proposed detection method is more suitable for pedestrian detection at night.

Fig. 5 shows the experimental results tested on a reasonable subset, reasonable subset is to select a part of image data that can reflect the performance of the algorithm from the complete dataset based on the demand for the algorithm. This is because public datasets are often designed for applications rather than for an algorithm. Therefore, it is very necessary to filter the image data according to the specific needs of the algorithm. The experimental strategy based on Fig. 5 is the common strategy used in the methods we compared. In the experiment, they eliminate extremely difficult objects such as less than 55 pixels and heavy occlusion. In order to verify the effectiveness of our method for difficult samples, we conducted a global experiment, tested all samples of the test set in the database, and counted the experimental results. The experimental results are shown in Fig. 6, in which figures (a) (b) (c) are the experimental results under different lighting conditions in all, daytime and nighttime respectively. The experimental results show that the proposed method is the best. It should be noted that through horizontal comparison, it is not difficult to find that the performance improvement effect of our method is more obvious at night. There is no doubt that this is the result of our network fusing thermal images. Through vertical comparison, Fig. 6 is significantly better than Fig. 5 in performance improvement. The only difference between the two groups of experiments is that the experimental dataset is different. It can be concluded that compared with other methods, our method has certain advantages in detecting difficult samples, which shows that the deep information fusion strategy in our method has better performance in dealing with difficult samples.

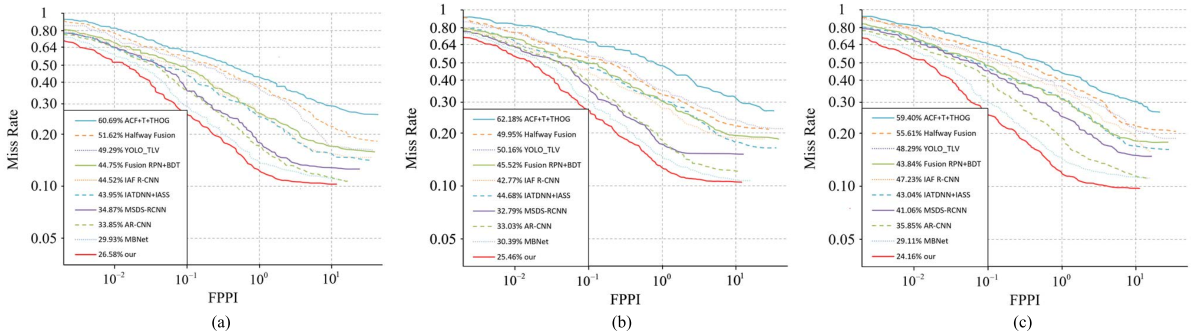
For quantitative analysis, we analyze the experimental results from two perspectives. From the perspective of scale, the dataset divides the object into near, medium and far according to the height of pedestrians in pixel units: size < 45, $45 \leq$ size ≤ 115 and $115 <$ size. In order to further verify the performance of the proposed method against small objects, we compare different methods under these three subsets. As can be seen from Table 1, compared with MBNet, our method reduces the MR by 0%, 2.56% and 3.79%, respectively. In order to verify the ability of our method to detect occluded objects, we verify the performance through three subsets: no occlusion, partial occlusion and heavy occlusion. It can be seen from the statistical results that our method has a strong ability in detecting occluded objects. These further verify the ability of our method to deal with difficult samples. Through the analysis of some samples of the experimental results, we find that our method is also difficult to work for the very small and heavy occluded objects, which shows that for the deep network, the completeness of the object's own information expression is an extremely important factor affecting the object detection results. Therefore, a good visual sensor and the image and video acquisition strategy are the key factors for the final application of artificial intelligence.

The performance of the deep learning algorithm is often strongly related to the scale of the deep network and the scale of the training dataset. The detection performance of networks and data scales with different volumes is not comparable. When the dataset is determined, the network scale is often



(a) (b) (c)

FIGURE 5. Comparison of the detection results on the KAIST dataset in reasonable settings. (a) Reasonable all, (b) Reasonable daytime, (c) Reasonable nighttime.



(a) (b) (c)

FIGURE 6. Comparison of the detection results on the KAIST dataset in all settings. (a) All all, (b) All daytime, (c) All nighttime.

TABLE 1. Comparisons with other multispectral detectors on the KAIST dataset.

Methods	Scale			Occlusion		
	Near	Medium	Far	None	Partial	Heavy
ACF+T+THOG[7]	28. 63	51. 71	85. 57	54. 69	70. 49	81. 50
Halfway Fusion[5]	8. 01	30. 69	74. 57	47. 44	61. 17	72. 04
YOLO_TLV[16]	3. 20	31. 80	86. 93	44. 76	55. 92	72. 73
Fusion RPN+BDT[3]	0. 12	30. 87	84. 42	40. 57	46. 66	70. 46
IAF R-CNN[15]	1. 02	23. 12	71. 16	41. 65	46. 48	63. 03
IATDNN+IASS[20]	0. 03	27. 01	80. 10	40. 85	47. 36	62. 13
MSDS-RCNN[12]	1. 26	16. 13	67. 36	31. 22	37. 67	60. 62
AR-CNN[19]	0. 02	15. 97	69. 23	30. 67	36. 59	56. 74
MBNet[6]	0. 00	15. 31	56. 46	26. 69	35. 4	58. 00
YOLO_CMN(our)	0. 00	12. 75	52. 67	24. 82	31. 94	53. 74

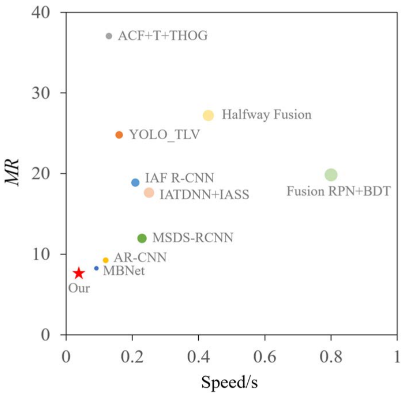


FIGURE 7. Log-average miss rate versus the running time of each detector in reasonable settings.

positively correlated with the workload of calculation. The size of the network will determine the demand for computing resources, and the demand for computing resources often determines the application field and scope of the algorithm. Therefore, comparing the computing resource requirements of algorithms is very important to evaluate the performance of algorithms, and the usual strategy is to compare the processing speed of algorithms on the same computing platform. In order to achieve this goal, Fig. 7 shows the comparison of the calculation speed between our method and several other methods mentioned in this paper. It can be seen from the results that our method performs best in terms of speed, and the result is 50 FPS.

In order to further verify the effectiveness of our proposed algorithm, the proposed algorithm is qualitatively analyzed with other advanced algorithms, and the detection results of objects under different environmental conditions are visualized. Fig. 8 and 9 show the detection results during the day and night respectively.

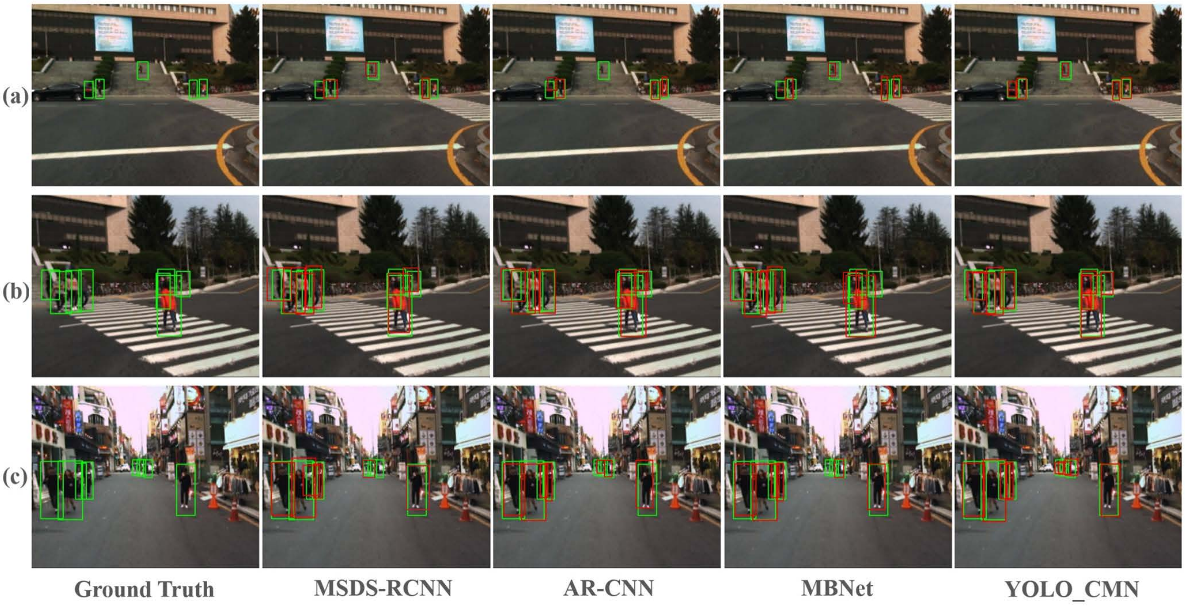


FIGURE 8. Comparison of detection results of four fusion frameworks in daytime scenarios. (a) Object occluded by the background, (b) object occluded by the pedestrian, (c) objects with different scales.

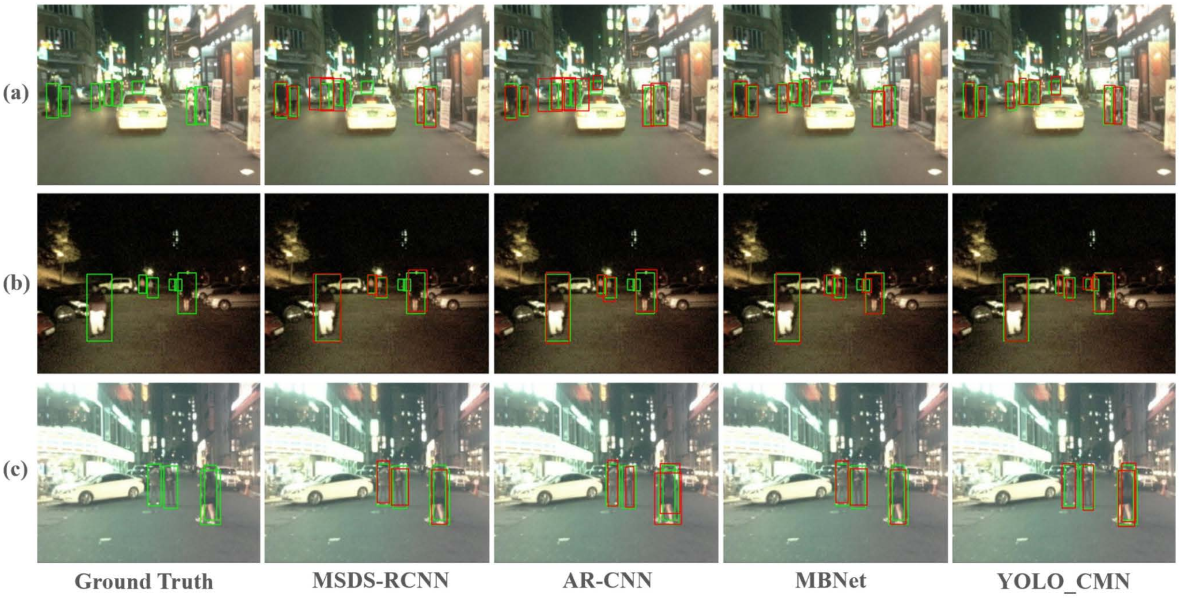


FIGURE 9. Comparison of detection results of four fusion frameworks in nighttime scenarios. (a) Object occluded by the background, (b) objects with different scales, (c) object occluded by the pedestrian.

Fig. 8 (a), (b) and (c) show the pedestrian detection results under different illumination conditions during the day, respectively. Fig. 9 (a), (b) and (c) show the pedestrian detection results under different illumination conditions at night, respectively. From the visualization results, we can see that under different lighting conditions, the number of missed objects in our algorithm is the least and has the best detection performance. In Fig. 8 (c) and Fig. 9 (b), there are objects with different scales. The visualization results verify that our algorithm also has good detection performance for objects with different scales. Fig. 8 (a) (b) and Fig. 9 (a) contain the object occluded by the background, and Fig. 8 (b) and Fig. 9 (c) contain the object occluded by the pedestrian. Under these different occlusions, our algorithm has better visualization results. The visual experimental results show a good detection effect for multi-scale and dense pedestrians under different background environments and different illumination conditions. Compared with other advanced multispectral pedestrian detection frameworks, YOLO_CMN has less missed detection for different forms of pedestrians, and the prediction bounding box is closer to the ground truth bounding box, which further verifies the effectiveness of the proposed algorithm. It can be seen from the detection results that after the enhancement and fusion of the two modal image

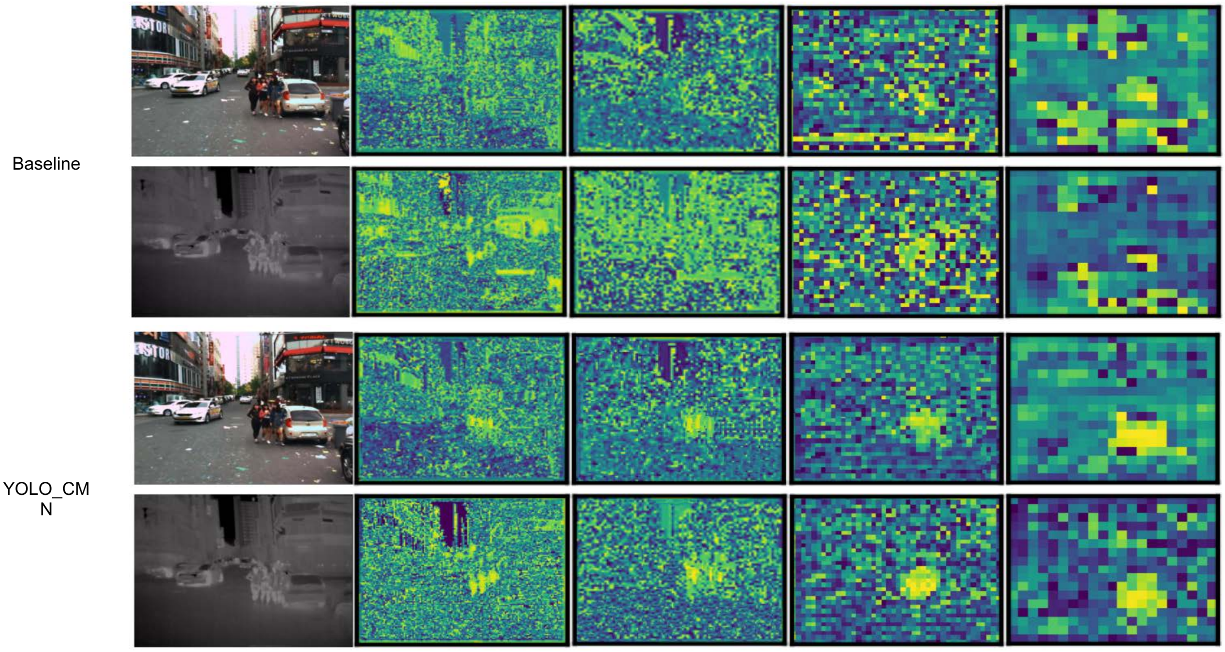


FIGURE 10. The overall perspective of the baseline and YOLO_CMN feature maps.

features, the position and category feature information of the object are integrated to enhance the feature expression ability of the object, so that the algorithm can better detect the object and make the detection effect remarkable.

In conclusion, we believe that the reason why our proposed algorithm has the better performance under different lighting conditions is inseparable from the improvement of the detection performance of difficult samples. Through comparative experiments, it is verified that the proposed algorithm has achieved a good balance in speed and accuracy.

E. ABLATION STUDIES

CFCM module can help one modality obtain complimentary feature information from another modality during feature extraction. In order to verify the effectiveness of the proposed module, we conducted ablation studies on the module. CFCM module is applied to the feature extraction stage of the two sub-networks. The specific deployment is shown in Fig. 2 (d). This part of the experiment takes the architecture in Fig. 2 (c) as the baseline, which does not use the CFCM module. On the basis of this architecture, we use different numbers of CFCM modules to fuse and complement the feature maps of different modalities. As summarized in Table 2, the more CFCM modules are used in the feature extraction network, the lower the MR and the better the performance of the detection network. The baseline does not use CFCM module, and the MR values are 11.23%, 10.84%, and 11.47% on three subsets for reasonable settings, reasonable daytime, and reasonable nighttime, respectively. In YOLO_CMN architecture, the MR values are 7.85%, 8.03%, and 7.82%, respectively. Compared with baseline, the missed detection rates were reduced by 3.38%, 2.81%, and 3.65%, respectively.

TABLE 2. Comparison of MR using different numbers of CFCM.

Number of CFCM				MR (Reasonable)		
0	1	2	3	All	Day	Night
√	×	×	×	11.23	10.84	11.47
×	√	×	×	9.11	9.52	9.03
×	×	√	×	7.98	8.35	8.25
×	×	×	√	7.85	8.03	7.82

The feature map is obtained after the operation of the CSP module of the backbone network. The visualization results of the feature map are shown in Fig. 10. By comparing the feature maps of baseline and YOLO_CMN, it can be seen that the addition of the CFCM module enhances pedestrian features and suppresses background features. With the deepening of the network, the CFCM module is conducive to refining and integrating the background features of the pictures, reducing background noise interference, and enhancing the features of pedestrian areas. Especially in the case of insufficient illumination conditions, it is difficult to extract pedestrian features on RGB images, while the pedestrian features contained in thermal images are relatively prominent. CFCM module enables the feature information in thermal images to be learned simultaneously when extracting pedestrian features in the RGB modality. Due to the fusion and complementarity of the two modal features, the network can learn richer feature information and improve the detection performance. In general, the CFCM module promotes modal interaction in the network, reduces target loss, highlights pedestrian features, reduces redundant learning, transmits more information, and improves the detection effect under different illumination conditions.

AFEFM module first enhances and suppresses the features of the two modalities and then fuses them to obtain

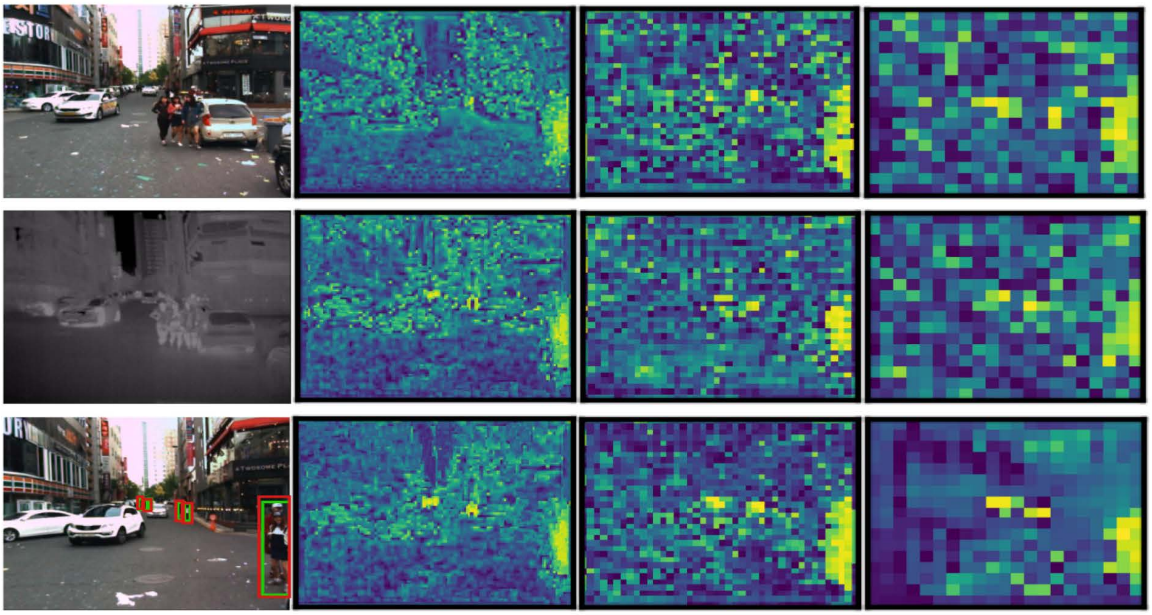


FIGURE 11. Feature map visualization at different stages. Columns 2 to 4 are the visualization results of feature maps from stages 2 to 4, respectively. The first and second rows are the feature images in the RGB and thermal modalities, respectively. The third row is the feature graph after AFEFM module fusion.

TABLE 3. Comparative analysis of different cross-modality fusion methods.

Method	MR (Reasonable)		
	All	Day	Night
SUM	9.01	9.54	9.87
MIN	8.12	8.54	8.97
AFEFM	7.85	8.03	7.82

more abundant information. In order to evaluate the effectiveness of the AFEFM module, we conducted the following experiments. The feature fusion strategy of the AFEFM module is shown in Fig. 2. The classical SUM and MIN fusion methods are compared to the AFEFM module for ablation experiments. The experimental results are shown in Table 3. Compared with the SUM fusion method, the MR values are reduced by 1.16%, 1.51%, and 2.05%, respectively. Compared with the MIN fusion method, the MR values are reduced by 0.27%, 0.51%, and 1.15%, respectively. The fusion features obtained by the AFEFM module have stronger semantic expression ability. As seen from the feature map in Fig. 11, the AFEFM module can integrate features of different modalities to further highlight pedestrian features. In general, the proposed AFEFM module fully integrates color flow and thermal flow, so that the information of the two modal feature maps can be further complementary.

In order to evaluate our dual-modality feature fusion architecture, we compared YOLO_CMN with some classical architectures, such as input fusion architecture, early fusion architecture, and halfway fusion architecture. The experimental results are shown in Table 4. Further, the proposed network is compared with methods using only color images and using only thermal images. Color Only and Thermal Only in Table 4 are the test results of single spectral images directly trained on YOLOv5s. The MR values of Color Only are 27.67%, 26.45%, and 30.15% on three subsets for all, daytime, and nighttime, respectively. The MR values of Thermal Only are 23.79%, 25.16%, and 22.98%, respectively. The MR of single-modality is significantly higher than that of dual-modalities, which proves that the detection performance of dual-modality is better than that of single-modality. Among the dual-modalities algorithm, YOLO_CMN has the lowest MR, while Input Fusion has the highest MR. Compared with Input Fusion, the MR values are reduced by 7.06%, 6.81%, and 7.3%, respectively. The results show that the proposed network can effectively fuse features and improve detection performance.

TABLE 4. Comparison between different architectures.

Architecture	MR (Reasonable)		
	All	Day	Night
Color Only	27.67	26.45	30.15
Thermal Only	23.79	25.16	22.98
Input Fusion	14.91	14.84	15.12
Early Fusion	13.87	13.46	13.91
Halfway Fusion	11.23	10.84	11.47
YOLO_CMN	7.85	8.03	7.82

These ablation studies show that the proposed architecture has good detection performance and fast detection speed. Overall, the network achieves a good balance between detection accuracy and speed, which can be applied in practical engineering.

V. CONCLUSION AND FUTURE WORKS

This paper proposes a cross-modal detection network for all-day pedestrian detection. A low-cost CFCM module is added to the feature extraction stage of the lightweight feature extraction network (CSPDarknet53). It promotes the information interaction between different modalities during the feature extraction stage. Accordingly, the network can realize the complementarity between the information flows of the two modalities in the feature extraction stage to reduce the object loss. And we propose the AFEFM module to fuse the color and thermal streams, further enhancing features and reducing the error detection of the object. In addition, essential features of the two modalities are learned through the enhancement and suppression processes. And through the feature complementarity of color and thermal images and the multi-scale fusion of depth feature layer, we realize the horizontal and vertical multi-dimensional data mining of parallel depth network, fully enrich the depth semantic information of the object, to improve the detection performance of the detector. The experimental results show that the proposed model can effectively integrate the visible and infrared features, and can effectively detect pedestrians of different scales in various illumination variants and occlusions. Further, the proposed model is applicable in real-time applications.

Future works will include exploring a more reasonable attention mechanism for a more effective fusion of dual-modality features to achieve better detection performance and a lighter module to improve the detection speed of the network.

REFERENCES

[1] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1239–1258, Jul. 2010.

[2] M. Bilal, A. Khan, M. U. K. Khan, and C.-M. Kyung, "A low-complexity pedestrian detection framework for smart video surveillance systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 10, pp. 2260–2273, Oct. 2017.

[3] D. Kônig, M. Adam, C. Jarvers, G. Layher, H. Neumann, M. T. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. Workshops (CVPRW)*, Jul. 2017, pp. 243–250.

[4] Y. Chen, H. Xie, and H. Shin, "Multi-layer fusion techniques using a CNN for multispectral pedestrian detection," *IET Comput. Vis.*, vol. 12, no. 8, pp. 1179–1187, 2018.

[5] J. Liu *et al.*, "Multispectral deep neural networks for pedestrian detection," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2016, pp. 731–733.

[6] K. Zhou, L. Chen, and X. Cao, *Improving Multispectral Pedestrian Detection by Addressing Modality Imbalance Problems*. Cham, Switzerland: Springer, 2020, pp. 787–803.

[7] 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. (CVPR)*, Jun. 2015, pp. 1037–1045.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 886–893.

[9] S. Kim and K. Cho, "Trade-off between accuracy and speed for pedestrian detection using HOG feature," in *Proc. IEEE 3rd Int. Confer. Consum. Electr. Berlin (ICCE-Berlin)*, 2013, pp. 207–209.

[10] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, Jan. 1996.

[11] D. M. Gavrila, "A Bayesian, exemplar-based approach to hierarchical shape matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 8, pp. 1408–1421, Aug. 2007.

[12] C. Li *et al.*, "Multispectral pedestrian detection via simultaneous detection and segmentation," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[14] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Apr. 2018, doi: 10.1109/TMM.2017.2759508.

[15] 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, Jan. 2019.

[16] M. Vandersteegen, K. Van Beeck, and T. Goedeme, "Real-time multispectral pedestrian detection with a single-pass deep neural network," in *Proc. 15th Int. Conf. Image Anal. Recognit. (ICIAR)*, Jun. 2018, pp. 27–29.

[17] X. Song, S. Gao, and C. Chen, "A multispectral feature fusion network for robust pedestrian detection," *Alexandria Eng. J.*, vol. 60, no. 1, pp. 73–85, Feb. 2021.

[18] Y. Zheng, I. H. Izzat, and S. Ziaee, "GFD-SSD: Gated fusion double SSD for multispectral pedestrian detection," 2019, *arXiv:1903.06999*.

[19] L. Zhang *et al.*, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5126–5136.

[20] 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, Oct. 2019.

[21] 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. (WACV)*, Jan. 2021, pp. 72–80.

[22] T. Liu, R. Zhao, and K. M. Lam, "Attention-based cross-modality interaction for multispectral pedestrian detection," in *Proc. Int. Workshop Adv. Imag. Technol.*, vol. 11766, M. Nakajima, J. G. Kim, W. N. Lie, and Q. Kemao, Eds. Bellingham, WA, USA: SPIE, 2021, doi: 10.1117/12.2590661.

[23] M. Kieu, L. Berlincioni, L. Galteri, M. Bertini, A. D. Bagdanov, and A. del Bimbo, "Robust pedestrian detection in thermal imagery using synthesized images," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 8804–8811, doi: 10.1109/ICPR48806.2021.9412764.

[24] K. Dasgupta *et al.*, "Spatio-contextual deep network based multimodal pedestrian detection for autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, to be published, doi: 10.1109/TITS.2022.3146575.

[25] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Multispectral fusion for object detection with cyclic Fuse-and-Refine blocks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 276–280.

[26] M. Kieu, A. D. Bagdanov, M. Bertini, and A. Bimbo, "Task-conditioned domain adaptation for pedestrian detection in thermal imagery," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 546–562.

[27] Z. Cao, H. Yang, J. Zhao, S. Guo, and L. Li, "Attention fusion for one-stage multispectral pedestrian detection," *Sensors*, vol. 21, no. 12, Jun. 2021, Art. no. 4184.

[28] 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.

[29] L. Fu, W.-B. Gu, Y.-B. Ai, W. Li, and D. Wang, "Adaptive spatial pixel-level feature fusion network for multispectral pedestrian detection," *Infr. Phys. Technol.*, vol. 116, Aug. 2021, Art. no. 103770.

[30] L. Ding, Y. Wang, R. Laganière, D. Huang, X. Luo, and H. Zhang, "A robust and fast multispectral pedestrian detection deep network," *Knowl.-Based Syst.*, vol. 227, Sep. 2021, Art. no. 106990.

[31] R. Liu, Y. Ruichek, and M. El Bagdouri, "Multispectral background subtraction with deep learning," *J. Vis. Commun. Image Represent.*, vol. 80, Oct. 2021, Art. no. 103267.

[32] R. Lu, B. Chen, Z. Cheng, and P. Wang, "RAFnet: Recurrent attention fusion network of hyperspectral and multispectral images," *Signal Process.*, vol. 177, Dec. 2020, Art. no. 107737.

[33] R. Dian, S. Li, B. Sun, and A. Guo, "Recent advances and new guidelines on hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 69, pp. 40–51, May 2021.

[34] M. Li, A. D. Bagdanov, M. Bertini, and A. Bimbo, "Deep Visible and thermal image fusion with cross-modality feature selection for pedestrian detection," in *Proc. IFIP Int. Conf. Netw. Parallel Comput.* Cham, Switzerland: Springer, Sep. 2021, pp. 117–127, doi: 10.1007/978-3-030-79478-1_10.

[35] N. Pourmomtaz and M. Nahvi, "Multispectral particle filter tracking using adaptive decision-based fusion of visible and thermal sequences," *Multimedia Tools Appl.*, vol. 79, nos. 25–26, pp. 18405–18434, Jul. 2020.

[36] J. Sun, Y. Li, H. Chen, J. Li, and F. Li, "Pedestrian detection based on depth information," in *Proc. 12th Int. Conf. Mach. Learn. Comput.*, Feb. 2020, pp. 249–553.

[37] Q. Hou, D. Zhou, and J. Feng. "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 19–25.

[38] N. Guo, K. Gu, J. Qiao, and J. Bi, "Improved deep CNNs based on nonlinear hybrid attention module for image classification," *Neural Netw.*, vol. 140, pp. 158–166, Aug. 2021, doi: 10.1016/j.neunet.2021.01.005.

[39] Y. Zhang, Z. Yin, L. Nie, and S. Huang, "Attention based multi-layer fusion of multispectral images for pedestrian detection," *IEEE Access*, vol. 8, pp. 165071–165084, 2020.

[40] T. T. Feng and H. Y. Ge, "Pedestrian detection based on attention mechanism and feature enhancement with SSD," in *Proc. 5th Int. Conf. Commun., Image Signal Process. (CCISP)*, Nov. 2020, pp. 145–148.

[41] W. Li, "Infrared image pedestrian detection via YOLO-V3," in *Proc. IEEE 5th Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Mar. 2021, pp. 1052–1055, doi: 10.1109/IAEAC50856.2021.9390896.

[42] H. Zhang, D. O. Pop, A. Rogozan, and A. Bensrhair, "Accelerate high resolution image pedestrian detection with non-pedestrian area estimation," *IEEE Access*, vol. 9, pp. 8625–8636, 2021.

[43] W. Boyuan and W. Muqing, "Study on pedestrian detection based on an improved YOLOv4 algorithm," in *Proc. IEEE 6th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2020, pp. 1198–1202.

[44] B. Han, Y. Wang, Z. Yang, and X. Gao, "Small-scale pedestrian detection based on deep neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 7, pp. 3046–3055, Jul. 2020.

[45] E. Dong, C. Jing, and Z. Zhang, "A multi-feature fusion based pedestrian detection method," in *Proc. IEEE Int. Conf. Mechatronics Autom. (ICMA)*, Oct. 2020, pp. 176–180.

[46] X. Zhang, L. Cheng, B. Li, and H.-M. Hu, "Too far to see? Not really!—Pedestrian detection with scale-aware localization policy," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3703–3715, Aug. 2018.

[47] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, "Generalizable pedestrian detection: The elephant in the room," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11328–11337.

[48] Y. Zhang, H. He, J. Li, Y. Li, J. See, and W. Lin, "Variational pedestrian detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11622–11631.

[49] A. Angelova, A. Krizhevsky, and V. Vanhoucke, "Pedestrian detection with a large-field-of-view deep network," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 704–711.

[50] H. Choi, S. Kim, K. Park, and K. Sohn, "Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 621–626.

[51] G. Brazil and X. Liu, "Pedestrian detection with autoregressive network phases," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7231–7240.

[52] X. Chu, A. Zheng, X. Zhang, and J. Sun, "Detection in crowded scenes: One proposal, multiple predictions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12214–12223.

[53] Y. Dai, J. Liu, Y. Sun, Z. Tong, C. Zhang, and L.-Y. Duan, "IDM: An intermediate domain module for domain adaptive person re-ID," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11864–11874.

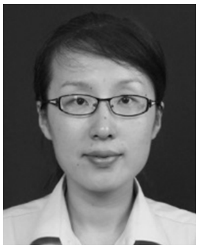
[54] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[55] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13024–13029.

[56] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5187–5196.

[57] X. Zhang, S. Cao, and C. Chen, "Scale-aware hierarchical detection network for pedestrian detection," *IEEE Access*, vol. 8, pp. 94429–94439, 2020.

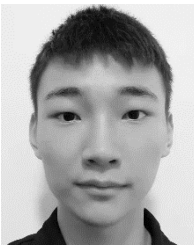
QUNYAN JIANG born in 1998. She received the bachelor's degree in automotive service engineering from Tongji Zhejiang College, China, in 2020. She is currently pursuing the master's degree with the College of Field Engineering, Army Engineering University of PLA. Her research interests include computer vision and model compression.



JUYING DAI born in 1982. She received the bachelor's and master's degrees from the Nanjing University of Aeronautics and Astronautics, China, and the Ph.D. degree from the Army Engineering University of PLA, China. She is an Associate Professor with the Army Engineering University of PLA. Her research interests include signal processing and diagnosis.

TING RUI received the M.S. and Ph.D. degrees from the PLA University of Science and Technology, Nanjing, China, in1998 and 2001, respectively. He is a Professor with the Army engineering University of PLA. He has authored and coauthored more than 80 scientific articles. His research interests include computer vision, machine learning, multimedia, and video surveillance.

FAMING SHAO born in 1978. He received the Ph.D. degree from the Army Engineering University of PLA, China. He is an Associate Professor with the Army Engineering University of PLA. His research interests include signal processing, deep learning, and software engineering.



JINKANG WANG received the bachelor's degree in mechanical engineering from the Army Engineering University of PLA, China, in 2020. He is currently pursuing the master's degree in mechanical engineering from the Army Engineering University. His current research interests include mechanics, machine learning, and computer vision.

GUANLIN LU is currently pursuing the master's degree with the College of Field Engineering, Army Engineering University of PLA. His research interest includes machine learning.