Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

DMFFNet: Dual-Mode Multi-Scale Feature Fusion-Based Pedestrian Detection Method

RUIZHE HU, TING RUI (Member, IEEE), YAN OUYANG, JINKANG WANG, QUNYAN JIANG AND YINAN DU

Department of Mechanical Engineering, College of Field Engineering, Army Engineering University, PLA, Nanjing 210007 CHINA

Corresponding author: TING RUI (rtinguu@sohu.com).

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

ABSTRACT Most contemporary pedestrian detection algorithms are based on visible light image detection. However, in environments with dim light, small targets, and easily occluded and cluttered backgrounds, single-mode visible light images relying on color, texture, and other features cannot adequately represent the feature information of targets; as a result, a large number of targets are lost and the algorithm performance is not good. To address this problem, we propose a dual-modal multi-scale feature fusion network (DMFFNet). First, we use the MobileNet v3 backbone network to extract the features of dual-modal images as input for the multi-scale fusion attention (MFA) module, combining the idea of multi-scale feature fusion and attention mechanism. Second, we deeply fuse the multi-scale features output by the MFA with the double deep feature fusion (DDFF) module to enhance the semantic and geometric information of the target. Finally, we optimize the loss function to reflect the distance between the predicted box and the real box more realistically as well as to enhance the ability of the network toward predicting difficult samples. We performed multi-directional evaluations on the KAIST dual-light pedestrian dataset and the visible-thermal infrared pedestrian dataset (VTI) in our laboratory through comparative and ablation experiments. The overall MR^{-2} on the KAIST dual-light pedestrian dataset is 9.26%, and the MR^{-2} in dim light, partial occlusion, and severe occlusion are 5.17%, 23.35%, and 47.31%, respectively. The overall MR^{-2} on the VIT dual-light pedestrian dataset is 9.26%, and the MR^{-2} in dim light, partial occlusion, and severe occlusion are 5.17%, 23.35%, and 47.31%, respectively. The results show that the algorithm performs well on pedestrian detection, especially in dim light and when the target was occluded.

INDEX TERMS Pedestrian detection, Dual-mode, Muti-scale, Attention mechanism, Feature fusion

I. INTRODUCTION

Object detection is an important part of the computer vision field, and it has received considerable research attention, especially for the detection of pedestrians and vehicles [1-4]. With the continuous breakthroughs in target detection technology, the detection of pedestrians is increasingly and more widely used. However, diverse pedestrian poses lead to non-obvious characteristics of pedestrians and the complex background information is easily affected by environmental conditions, making detection challenging. [5-6] Most contemporary pedestrian detection algorithms are based on visible light image detection. Visible light images have high resolution and rich color and texture features; therefore, the use of deep convolutional networks can more accurately extract target features in general scenes and detect and locate the target. [7] However, in environments with dim light, small targets, and easily occluded and cluttered backgrounds, single-mode visible light images relying on color, texture, and other features cannot adequately represent the feature information of targets, and thus, many targets remain undetected, and the algorithm performance is insufficient. [8]

A visible light image contains rich color and texture information of a target, has an obvious structure level, and well-differentiated foreground and background. However, under unfavorable environmental conditions, such as smoke, fog, clouds, dust, and dim light, the image quality is poor, which results in poor target feature representation ability [9]. Due to the principle of radiation imaging, infrared images can reflect the temperature distribution on the surface of an object and can obtain continuous scene image information [10]. In this case, the temperature of the object differs slightly from the surrounding environment, which results in obvious contour shape characteristics.

Herein, to obtain the respective characteristics of visible light images and infrared images, we propose a pedestrian detection algorithm based on MoblileNet v3 [11] with dual-modal multi-scale feature fusion (labeled as DMFFNet). It is used for pedestrian detection in case of dimly lit, small and easily occluded objects, unstable feature representation, and cluttered backgrounds. The four major contributions of this study are listed below:

- Based on the features obtained from visible light and thermal infrared images, we propose a dual-modal multi-scale feature fusion pedestrian detection method that can significantly improve the detection accuracy in case of dim light and occluded targets.
- Multi-scale fusion attention (MFA): MFA can effectively extract more fine-grained multi-scale spatial information and establish longer-distance channel dependencies, resulting in more stable feature representation.
- Double deep feature fusion (DDFF): DDFF can effectively fuse multi-scale feature information on a deeper level, considerably enhancing the representation of semantic information and geometric detail.
- The proposed cross-modal feature fusion algorithm demonstrated considerably improved target detection accuracy in unfavorable environments on KAIST and VTI datasets.

This paper is organized as follows. Section I briefly presents the shortcomings of contemporary target detection algorithms in unfavorable environments and the feature information of visible light and infrared images. Section II reviews the published pedestrian detection methods. Section III introduces the proposed network model and the improved loss function. Section IV evaluates the performance and effectiveness of the various aspects of DMFFNet through comparative and ablation experiments. Section V presents conclusions of this paper and presents future work.

II. Related Work

Pedestrian detection, as a subset of object detection, has received considerable research attention for many years. Early pedestrian detection methods mostly used appearance and motion features. Amnon et al. [12] proposed 13 key features based on nine key parts of the human body and their relative positional relationships. Havasi [13] proposed a cubic symmetry feature based on human legs. With the gradual development of pedestrian feature extraction types, the extraction types are not only appearance and motion features, but can be divided into three categories: low-level features (such as HOG [5] proposed by Dalal), learning-based features (such as Edgelet [6] proposed by Wu et al.), and hybrid features (COV [14] proposed by Tuzel et al.).

Neural networks are widely used in large-scale image classification; therefore, scholars have found that deep learning methods can learn to extract better pedestrian image features—enabling automation of the process—and at the same time, learn better similarity metrics, which can greatly improve the accuracy of pedestrian detection. The regional convolutional neural network (R-CNN) model proposed by Girshick et al. [15] achieved the highest accuracy. Angelova et al. [16] proposed a cascaded CNN-based pedestrian detection algorithm based on the idea of cascaded classifiers in the Adaboost algorithm; it could quickly eliminate most of the background areas in an image. Ouyang et al. [17] proposed the joint deep algorithm that combines HOG features and cascading style sheets features. Liu et al. [18] proposed the SSD model that can directly classify and localize on multi-scale feature maps, which considerably improved the detection speed. In some studies, the YOLO series model [19-22] was applied for pedestrian detection; it divided an image into several squares, and the output was pedestrian detection boxes through non-maximum suppression.

Pedestrian detection algorithms are not ideal for small target detection in dim light and occlusion; to solve this problem, researchers have proposed various improved methods and designed their own frameworks. Zhao et al. [23] performed repeated identification of each target block, followed by non-maximum suppression to finely distinguish the target prediction boxes; however, there were still many missed detections for densely arranged small-scale targets. Sun et al. [24] proposed incorporating local information into contextual information to improve fusion target detection and solve the problems caused by size and shape; the application of shallow image features, however, was insufficient. Fu et al. [25] proposed a feature fusion architecture to generate multi-scale features. The target features were aggregated through a top-to-bottom two-point path, which greatly increased the feature representation ability of the target; however, it had low detection ability for occluded targets. Zhang et al. [26] proposed an end-to-end intensive attention fluid network that uses a global context-aware attention module to capture long-range contextual information and embeds it into a dense attention fluid structure, enabling aggregation of shallow and deep features to guide the generation of high-level feature attention maps, thereby improving detection accuracy. Cheng et al. [27] proposed a cross-scale feature fusion network that uses a squeeze excitation module to simulate the relationship between different channels and obtains feature map information at multiple scales, achieving powerful multi-level feature representation through a fusion mechanism. Hwang et al. [8] proposed the KAIST dual-light pedestrian dataset; it fused visible light and thermal infrared images to considerably reduce the pedestrian missed detection rate in dim light and occlusion scenarios. Several pedestrian detection methods based on visible light and thermal infrared images have been proposed since, such as MBNet [28], Fusion RPN+BF [29], IATDNN+IAMSS [30], IAF R-CNN [31], MSDS-RCNN [32], and CIAN [33]. These improved methods have significantly reduced the pedestrian missed detection rate;

however, there is scope for further improving this rate under unfavorable conditions such as dim light and occlusion.

III. Approach

The proposed network model is illustrated in Figure 1. The network comprises four modules, namely, the MoblieNet v3 [11] feature extraction backbone network, MFA, DDFF, and DFF. The features of the dual-modal images at each scale extracted by MobileNet v3 [11] are first used as the input of the MFA module, which fuses the context information of different scales to produce better pixel-level attention. It can process the spatial information of multi-scale input feature maps and effectively establish long-term dependencies between multi-scale channel attention, reducing the interference of background noise. Thereafter, the DDFF module deeply fuses the multi-scale features output by the MFA module to maximize the correlation between the multi-scale features, which considerably enhances the ability of the features to represent semantic information and geometric details. Finally, the DFF module performs an element-wise add operation on the features of the corresponding scales of the two modalities as the output. The proposed DMFFNet can generate better pixel-level attention, enhance feature representation at different scales, especially for small objects, and greatly improves detection accuracy under unfavorable conditions, such as poor object feature representation, occlusion, and light changes.

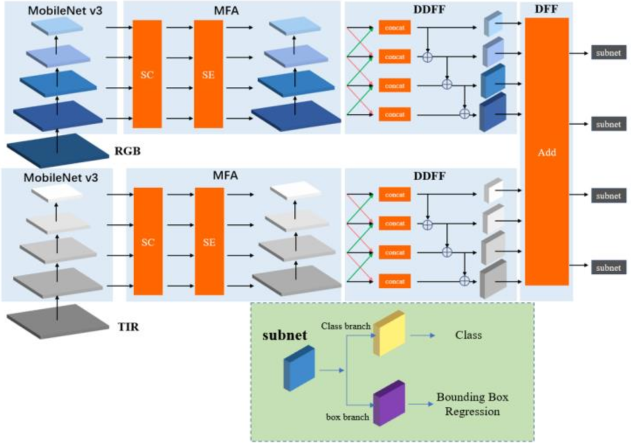


FIGURE 1. Network structure of DMFFNet.

A. Feature extraction

We adopt the MobileNet v3-small [11] backbone network as our feature extraction network. The overall structure is shown in Figure 2. It has a total of 12 layers, and the size of the input image is 224*244. MobileNet v3 [11] inherits the depthwise separable convolution of MobileNet V1 [34], inherits the residual structure with linear bottleneck of MobileNet V2 [35], and introduces the SE channel attention structure. MobileNet v3 uses a new activation function hswish(x) instead of Relu6, the activation function h-swish(x) is expressed in Eq. (1):

$$hswish(x) = x\frac{ReLU6(X+3)}{6} \qquad (1)$$

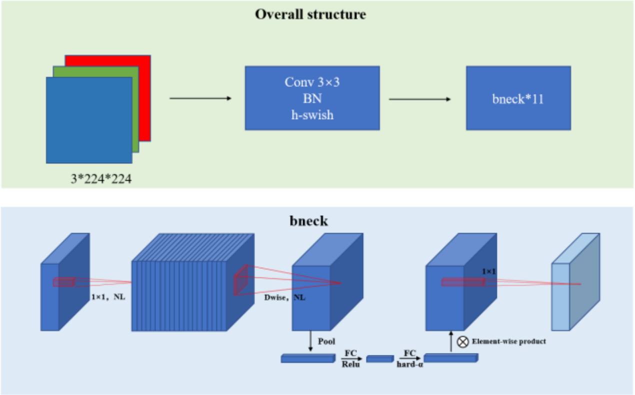


FIGURE 2. Structure of feature extraction.

Compared to the heavyweight network, we use the MobileNet v3 [11] backbone network as our feature extraction network to obtain faster inference speed.

B. MFA module

The attention mechanism is used to emphasize the importance of different features by assigning weights to features. It imitates the law of human brain observation activities, giving more weight (i.e., more attention) to specific important target points on an image; this helps in highlighting features of the target image and suppressing the features of other objects, achieving enhancement of feature information [36]. The MobileNet v3 [11] feature extraction network is used herein to obtain four-scale feature maps of visible light and thermal infrared images, which enriches the extracted image feature and clarifies target location. To better aggregate local and global features of an image so that the extracted bimodal feature information can better characterize the location and category of image objects, while reducing background influence, the MFA module is used to enhance features of the extracted feature maps of two modalities with different scales. The MFA module, which comprises split-and-concat (SC) and squeeze-and-excite (SE) modules, is illustrated in Figure 3.

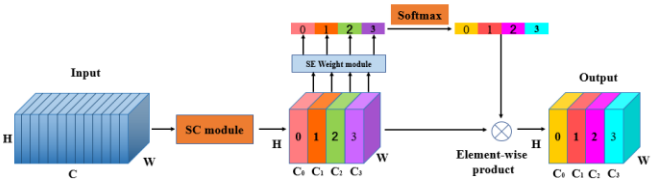


FIGURE 3. Structure of the MFA module.

1) SC MODULE

The SC module is illustrated in Figure 4. A large convolution kernel results in complex and large calculations, which hampers the real-time performance of the algorithm. Through multiple experiments, the missed detection rate and real-time performance are optimized, and the input feature maps are then divided into four groups on the channel. In each group, convolution is carried out with different convolution kernel sizes to obtain respective fields of different scales, extract information of different scales, and obtain different feature maps, F_i, using Eq. (2):

$$F_i = Conv(k_i \times k_i)(x_i) \quad i = 0,1,2,3 \tag{2}$$

where $i = 2 \times (i+1) + 1$, and x_i represent each group of feature maps after segmentation.

Finally, all feature maps, F_i, are fused through the Concat operation to obtain a feature map, F, using Eq. (3):

$$F = Cat([F_0, F_1, F_2, F_3]) \tag{3}$$

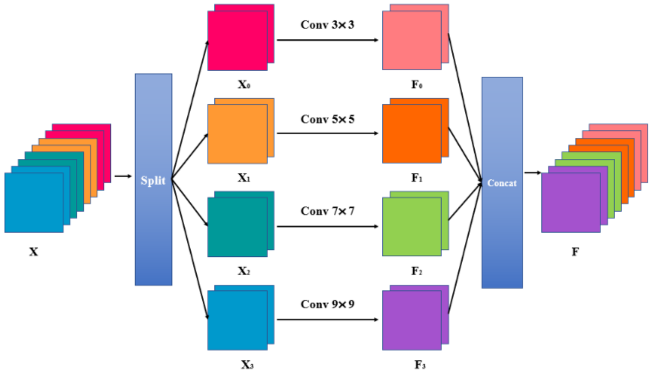


FIGURE 4. SC module.

2) SE MODULE

Channel attention allows the network to selectively measure the importance of each channel, thereby better preserving the correlation of features between each channel. As shown in Figure 5, first, the SE weight module is used to extract the attention of feature maps at different scales, and the channel attention vector is obtained. Second, the channel attention vectors are recalibrated using Softmax to obtain recalibrated multi-scale channel weights. Finally, element-wise product operation is applied to the recalibrated weights and the corresponding feature maps, and a refined feature map with richer multi-scale feature information is obtained as output.

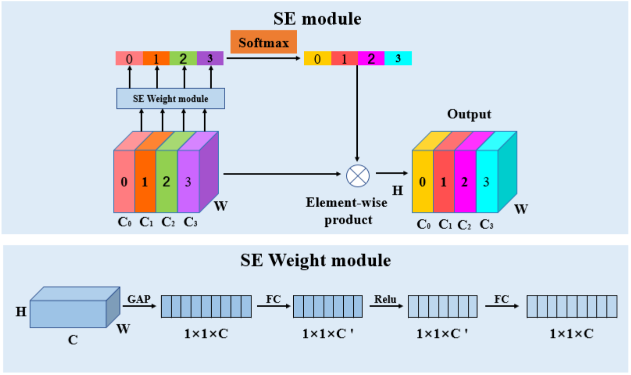


FIGURE 5. SE module and SE weight module

The SE weight module comprises a global average pooling (GAP) layer, an activation layer, and two fully connected (FC) layers. First, the input feature map is subjected to GAP, and then, the dimension is reduced through the FC layer and activated through the rectified linear unit (Relu). Thereafter, the dimension is restored through the FC layer. The formula for GAP is given in Eq. (4):

$$GAP = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x(i,j) \tag{4}$$

The attention weight of the ith channel in the SE weight model can be expressed using Eq. (5):

$$C_i = F_2(R(F_1(G_i))) \tag{5}$$

where G_i represents the GAP of i channels, F_1 represents the first FC layer, which reduces the dimension, R represents the Relu activation function, and F_2 represents the second FC layer, which restores the dimension.

C. DDFF module

The receptive field of a high-level network is relatively large, and the representation ability of semantic information is strong; however, the feature map is low-resolution, and the representation ability of geometric information is weak. For a low-level network, the receptive field is relatively small, and the representation ability of geometric details is strong. Although the resolution is high, the representation ability for semantic information is weak. For such situations, we use the DDFF of high- and low-level features, which results in feature representation with richer semantic and geometric information.

By combining the extracted global context features in two different ways, information loss can be greatly reduced. As shown in Figure 6, the first feature fusion fuses adjacent features of different scales. The scale is first unified through convolution and upsampling with different steps, and then, the channel numbers of adjacent features are merged through the Concat operation, labeled as B_i; the formula is expressed in Eq. (6):

$$B_i = \begin{cases} Cat(Conv(A_i, S=1), Conv(A_{i+1}, S=2)) & i = 0 \\ Cat(Up(A_{i-1}), Conv(A_i, S=1), Conv(A_{i+1}, S=2)) & i = 1,2 \\ Cat(Up(A_{i-1}), Conv(A_i, S=1)) & i = 3 \end{cases} \tag{6}$$

where $Conv$ represents the convolution operation, S represents the stride during the convolution operation, A_i represents the input feature maps of different sizes, Up represents the upsampling operation, and Cat represents the Concat operation.

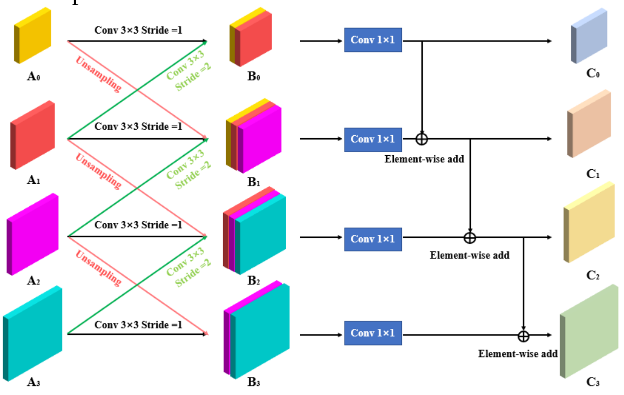


FIGURE 6. DDFF module.

The first feature fusion uses the semantic information of different scales to obtain a feature map with strong semantic and geometric detail representation abilities, and it reduces the interference of background noise. Considering that the number of feature channels after the Concat operation is relatively large, we use a 1×1 convolution kernel to adjust the number of channels, which does not consider the huge semantic gap between these features. The second feature fusion uses a parallel strategy to perform element-wise add fusion operation on the feature map fused by Concat, and it combines two adjacent feature vectors into a complex vector. The add fusion operation does not increase the dimension of the image, but it increases the amount of information in each dimension, thus increasing the perception ability of contextual information.

DDFF enhances the semantic and geometric information of a target through two feature fusions, increases the perception ability of context information, and further eliminates the influence of noise and complex background.

D. DFF module

To improve the detection accuracy in unfavorable environments such as dim light and occlusion, we fuse the dual-modal information to complement the image information and greatly enrich the image feature information. In contemporary methods, image fusion is mainly divided into three levels, namely, pixel, feature, and decision levels [37]. Decision-level fusion is the highest-level processing method for image fusion [38]. It is based on the preliminary judgment of an algorithm about information such as the location and category of the specific target; it uses information synthesis to ensure judgment accuracy; however, it relies too much on the previous feature processing results. Pixel-level feature fusion calculates each pixel of each modal image one by one [39]. It is the lowest-level fusion operation, targeting each pixel, which results in considerably complex computation. Feature-level fusion involves feature integration for a certain area of an image object or a certain feature such as color or edge [40]; the favorable factors of each feature are integrated for the classifier to use. Therefore, feature-level fusion can highlight certain types of targets in a targeted manner, which is beneficial for highlighting image details.

After the visible light and infrared images have passed through the MFA and DDFF modules, the perception ability of spatial position information, channel semantic information, and context information is enhanced, and correlation dependencies are generated on local regions and global features. Then, the feature-level image fusion strategy is adopted for the dual-modal information, and the information is integrated using the element-wise add function. The integration method is presented in Figure 7.

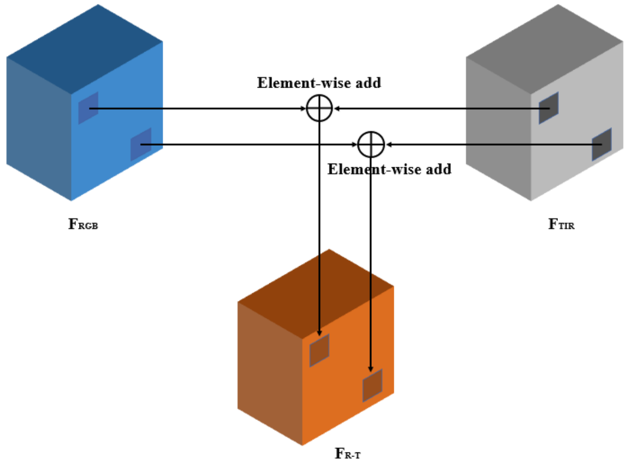


FIGURE 7. DFF module.

Elements in the same spatial position of the enhanced visible light feature map (F_{RGB}) and the thermal infrared feature map (F_{TIR}) are added to realize information fusion of the dual-modal images at the feature level; then, a fusion feature map (F_{R-T}) with high representation ability of target features is generated. The fusion calculation method is expressed using Eq. (7).

$$F_{cij} = R_{cij} + T_{cij} \tag{7}$$

where F_{cij}, R_{cij}, and T_{cij} represent the enhanced visible light, thermal infrared, and the position of (i, j) in channel c of the fused image, respectively. By using two different information representation methods of visible light and infrared images, information such as color texture in visible light is fused with the highlight position and contour information in infrared images, thereby improving target feature representation in unfavorable environments.

E. Model output and loss function improvement

We optimize the loss function based on YOLO v4 [22]. First, the anchor box is reset; then, we conduct object-oriented representations, and finally, the position loss function is redefined.

1) ANCHOR BOX SETTING

For the case when the output of pedestrian detection is mainly small objects, we redefine the anchor box. Anchor boxes are a series of prior boxes with specific size and scale. In the process of target detection, the size of the prior bounding box influences the detection accuracy. In the training phase, it is essential to design appropriate anchor boxes according to the characteristics of specific objects. In YOLO v4 [22], the anchor boxes are mainly used for general target detection and cannot meet the detection requirements of small targets; thus, it is necessary to readjust the anchor boxes. Moreover, the cluster center selected by YOLO v4 [22] using the K-means algorithm has a large randomness, which may easily lead to local optimization of the cluster and tends to generate a large-sized cluster box, decreasing the accuracy of small target positioning. Based on this, in this study, the K-means++ [41]

clustering algorithm with less randomness is used to cluster the target annotation boxes, effectively reducing the clustering deviation caused by the initial clustering. The processing of the K-means++ algorithm is described below.

First, according to the data distribution density, a reasonable annotation box is selected from the dataset as the initial cluster center. Then, the intersection ratio (IOU) of the remaining samples and the cluster center is calculated, sample data with a smaller distance are selected as the new cluster center according to the distance degree, $D(b,a)$, and finally, k cluster centers are obtained through iterative calculation. The distance $D(b,a)$ is calculated using Eq. (8):

$$\begin{cases} IOU(b,a)=\frac{b\cap a}{b\cup a} \\ D(b,a)=1-IOU(b,a) \end{cases} \tag{8}$$

where a and b represent the area of the labeling box and the clustering box, respectively. The value of $D(b,a)$ is between 0 and 1. Smaller value of *D(b,a)* implies smaller distance between the cluster center and the label box and a better clustering effect. Thus, this method can effectively reduce the clustering deviation caused by the original algorithm at the initial clustering points and determine a more suitable anchor frame size.

2) OBJECT-ORIENTED REPRESENTATION

The proposed method, denoted by $(x,y,w,h,\alpha_1,\alpha_2,\alpha_3,\alpha_4)$ object-oriented, is intuitively illustrated in Figure 8. For an object O(red box) and the corresponding horizontal bounding box B_h(blue box), $P_i(i\in\{1,2,3,4\})$ represents the four points of its horizontal bounding box, and P_i' ($i\in\{1,2,3,4\}$) represents the four points of the bounding box. This can also be represented by (x,y,w,h), where (x,y) is the center point, and w and h are the width and height of the box, respectively. The additional variables are calculated using Eqs. (9)-(10).

$$\alpha_i=\begin{cases} \frac{\|s_i\|}{w} & i=1,3 \\ \frac{\|s_i\|}{h} & i=2,4 \end{cases} \tag{9}$$

$$\|s_i\|=\|p_i-p_i{}'\| \tag{10}$$

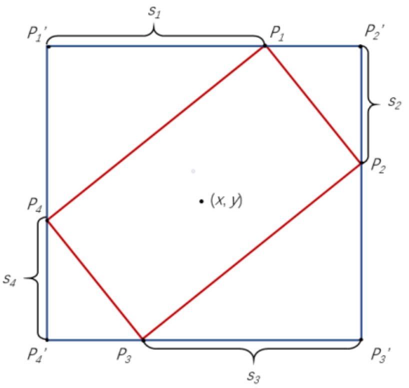


FIGURE 8. Representation for an oriented object

where s_i represents the distance between p_i and p_i'. By regressing the offset of the horizontal bounding box (B_h) on the four sides, DMFFNet achieves better learning for accurate object-oriented localization.

3) LOSS FUNCTION

The loss function comprises classification loss, L_{cls}, location loss, L_{loc}, and confidence loss, L_{cfi}, and it is calculated using Eq. (11).

$$L=L_{cls}+L_{loc}+L_{cfi} \tag{11}$$

The classification loss, L_{cls}, and confidence loss, L_{cfi}, are the same as those in YOLO v4 [14]. The location loss, L_{loc}, has two components, the horizontal bounding box and the length ratio $(\alpha_1,\alpha_2,\alpha_3,\alpha_4)$, and it is expressed using Eqs. (12)-(14).

$$L_{loc}=\lambda_1\times L_h+\lambda_2\times L_\alpha \tag{12}$$

$$L_h=1-CIoU \tag{13}$$

$$L_\alpha=\sum_{i=1}^{4}\text{smooth}_{L_1}(\alpha_i-\tilde{\alpha}_i) \tag{14}$$

where λ_1 and λ_2 are hyperparameters that measure the importance of each loss, α_i is the length ratio of the corresponding horizontal bounding box (B_h), and $\tilde{\alpha}_i$ is the length ratio of the ground truth box.

IV. Experiment Evaluation

The most authoritative public dataset of visible light–thermal infrared (VTI) dual-modality used for pedestrian detection is the KAIST dual-light pedestrian dataset [8]. Herein, to verify the effectiveness and robustness of the algorithm, we evaluate the performance of the various aspects of DMFFNet by comparing and ablation experiments on the KAIST dual-light pedestrian dataset [8]. To further verify the performance of the proposed method for smaller pedestrian scale and less favorable environment, we conduct supplementary experiments on the VTI dataset using a drone made in our laboratory.

A. Dataset introduction

The KAIST dual-light pedestrian dataset [8] includes VTI images in conventional scenes captured using a visible light camera, a thermal infrared camera, a beam splitter, and a three-axis fixture to form a hardware imaging acquisition system. Using the acquisition system to collect target data under different conditions of light and time periods, in the acquired images of all frames, Soonmin Hwang and team manually annotated 95382 VTI image pairs, including 103128 dense annotations and 1182 unique target annotations. In this study, the training and test sets comprised 7373 and 2252 images, respectively, which also included a reasonable proportion of occlusion of the target to different degrees and different situations in day and night scenes.

The VTI images dataset collected a total of 80 segments of pedestrian target data from the UAV's perspective in different scenarios. Furthermore, 32615 pairs of visible light and thermal infrared data were obtained by frame-by-frame extraction; the final dataset contained a total of 15628 target data in interlaced or independent backgrounds, such as rivers, buildings, and vegetation, and in different light environments; cases of when the target was occluded to different degrees were screened out. The number of visible light and thermal infrared data were equal.

B. Experimental setting

Experiments on the DMFFNet were carried out with fixed parameters. The PyTorch-1.4.0 deep learning framework was used, and a GeForce RTX 2080Ti GPU was configured to train the network using stochastic gradient descent (SGD). The momentum parameter was 0.9, the decay value was 0.005, the batch size was set to 8, a total of 100 epochs were trained, the baseline learning rate was 0.001, and the confidence threshold was 0.5.

C. Test results and analysis

1) RESULTS ON KAIST DUAL-LIGHT PEDESTRIAN DATASET

We trained the network using the KAIST dual-light pedestrian training set [8]; the subsets Reasonable (Day, Night, All) and Occlusion (Bare, Partial, Heavy) were used to evaluate the network performance. The results were compared with those of other mainstream VTI fusion algorithms; the miss rate-false positive per image (MR-FPPI) curve is shown in Figure 9, and the comparison results are shown in Table 1. We used MR^{-2} as the evaluation index, as it is used in most contemporary algorithms. In the MR-FPPI curve, 9 FPPIs within the range of [0.01,1] are uniformly selected, their corresponding nine log (MR) values are averaged, and the final value (MR^{-2}) is obtained by exponentially calculating the above average values. The formulas are given in Eqs. (15)-(17).

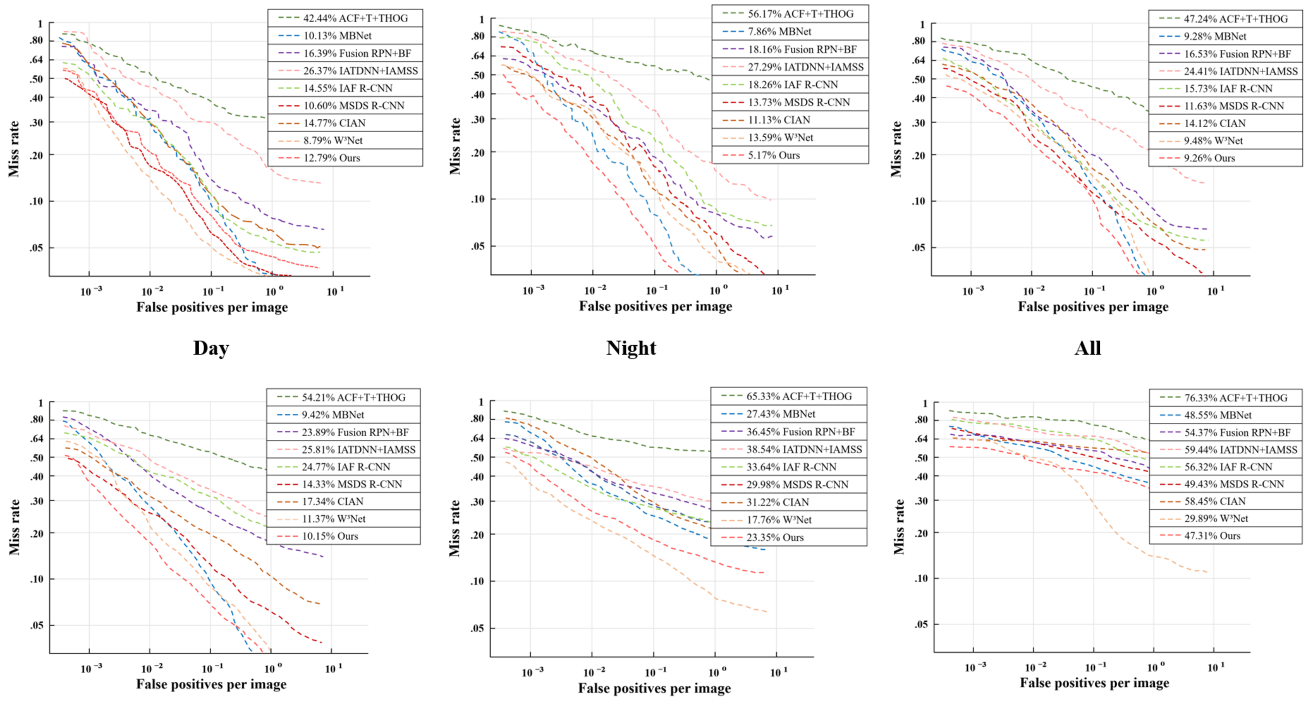
$$MR = \frac{FN}{TP+FN} \tag{15}$$

$$FPPI = \frac{FP}{N} \tag{16}$$

$$log(MR^{-2}) = \frac{1}{9}\sum_{i=1}^{9} log(MR_i) \tag{17}$$

where MR refers to the missed detection rate in the test results, FPPI refers to the average false positive rate of each image, and N refers to the number of images. True positive (TP) indicates the case when the predicted samples are positive and the predicted results are correct, false positive (FP) refers to the case when the predicted samples are positive and the predicted results are incorrect, and false negative (FN) refers to the case when the predicted samples are negative and the predicted results are incorrect.

Figure 9 shows the MR-FPPI curves of DMFFNet and other mainstream VTI fusion algorithms on the KAIST dual-light pedestrian dataset [8].



Day

Night

All

Bare

Partial

Heavy

FIGURE 9. MR-FPPI curves of ACF+T+THOG, MBNet, Fusion RPN+BF, IATDNN+IAMSS, IAF R-CNN, MSDS R-CNN, CIAN, W^{3}Net and ours on KASIT dataset

TABLE 1
COMPARATIVE TEST RESULTS OF THE KAIST DATASET

Method	Speed/s	MR^{-2} (%)					
		Reasonable			Occlusion		
		Day	Night	All	Bare	Partial	Heavy
ACF+T+THOG [24]	0.032	42.44	56.17	47.24	54.21	65.33	76.33
MBNet [25]	0.074	10.13	7.86	9.28	**9.42**	27.43	48.55
Fusion RPN+BF [26]	0.802	16.39	18.16	16.53	23.89	36.45	54.37
IATDNN+IAMSS [27]	0.254	26.37	27.29	24.41	25.81	38.54	59.44
IAF R-CNN [28]	0.216	14.55	18.26	15.73	24.77	33.64	56.32
MSDS R-CNN [29]	0.228	10.60	13.73	11.63	14.33	29.98	49.43
CIAN [30]	0.067	14.77	11.13	14.12	17.34	31.22	58.45
W^{3}Net [42]	0.274	**8.79**	13.59	9.48	11.37	**17.76**	**29.89**
Ours	**0.021**	12.79	**5.17**	**9.26**	10.15	23.35	47.31

Table 1 shows that the MR^{-2} of the proposed network in the Reasonable range is 12.79%, 5.17%, and 9.26% during the day, night, and all-day, respectively. In night scenes with poor lighting conditions, the algorithm outperforms other algorithms and achieves optimal results. DMFFNet is also the fastest, as seen in column 2. It enhances the target feature information of infrared and visible light features using the MFA module, and enhances context awareness through the DDFF module. The pixel-wise add operation of the two modal feature maps is performed at the feature level, and the complementary advantages of the dual-feature information are realized at the target level, which greatly enhances the feature expression ability of the image target. Thus, DMFFNet outperforms other networks in detection results on the night test subset within a reasonable range, proving its effectiveness with a lower MR^{-2}. Some resulting plots are shown in Figure 10.

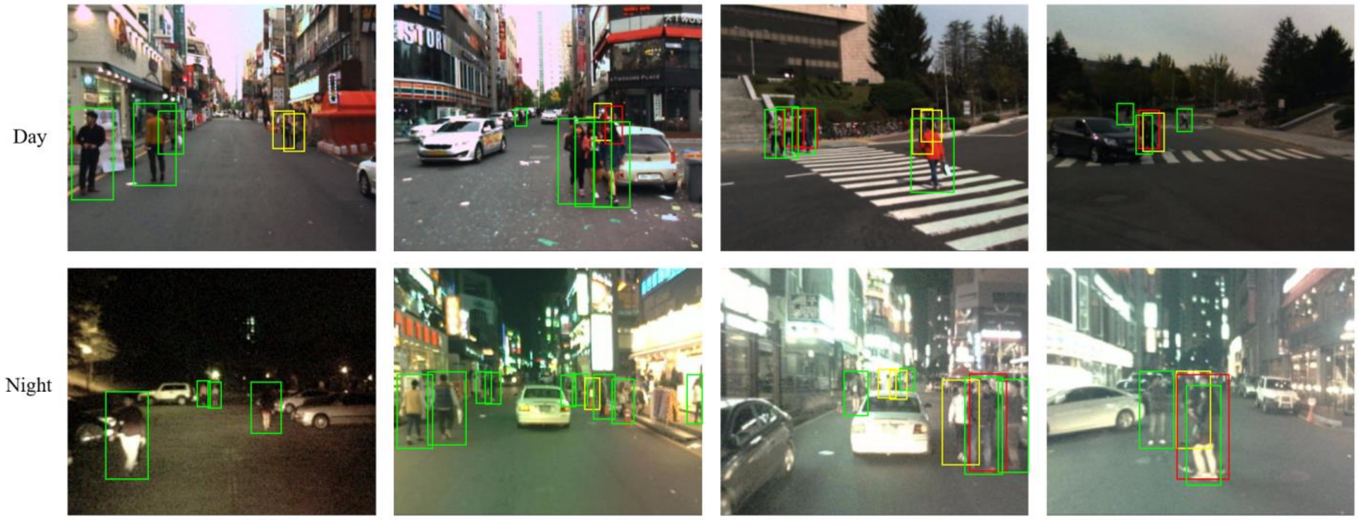


FIGURE 10. Some detection results of our method on KASIT dataset. The dense annotations provided with the dataset such as green, yellow, and red boxes indicate no-occlusion, partial occlusion, and heavy occlusion respectively.

The detection results show that when the target is occluded by ambient light and objects, in a separate visible light image or infrared image, the target feature expression ability is weak, and the coexistence ability of category features and location features is insufficient. However, after the enhancement and fusion of the dual-modal image features, the location and category feature information of the target are determined, which considerably improves the detection effect of the algorithm.

Further, DMFFNet is compared with mainstream visible light target detection algorithms. In the test set, subsets Reasonable (Day, Night) and Occlusion (Bare, Partial) were selected in the KAIST dual-light pedestrian dataset [8]. The comparison results are shown in Table 2. We used average precision (AP) as the evaluation index, as it is used in mainstream visible light target detection algorithms for performance evaluation. Recall is the number of ground truths divided by TP, and it indicates the proportion of correct detection results of an actual target; it is calculated using the formula given in Eq. (18). Precision is the sum of correct detection results divided by TP, and it indicates the accuracy of the detected results; it is calculated using the formula given in Eq. (19). AP represents the average detection accuracy of the model for a certain category of objects, and it is calculated using the formula given in Eq. (20).

$$Recall = \frac{TP}{TP+FN} \tag{18}$$

$$Precision = \frac{TP}{TP+FP} \tag{19}$$

$$AP = \int_0^1 P(R)dR = \sum_{k=0}^{n} P(k)R(k) \tag{20}$$

where $P(R)$ represents the precision of the recall rate at R point, k represents the precision truncation point, the corresponding $P(k)$ and $R(k)$ represent the precision and range of k points, respectively, and n represents the number of precision truncation points.

As seen in Table 2, dual-modal YOLO v5s and dual-modal Faster R-CNN use the original network to extract light features and thermal infrared features, respectively; they perform feature-level addition and fusion after processing the two modal features in the network neck. Among the test data types shown in column 2, RGB represents the visible light image, TIR represents the thermal infrared image, and RGB+TIR represents the registered VTI dual-modal image. The middle four columns represent the test accuracy under different conditions, and the last column represents the inference time.

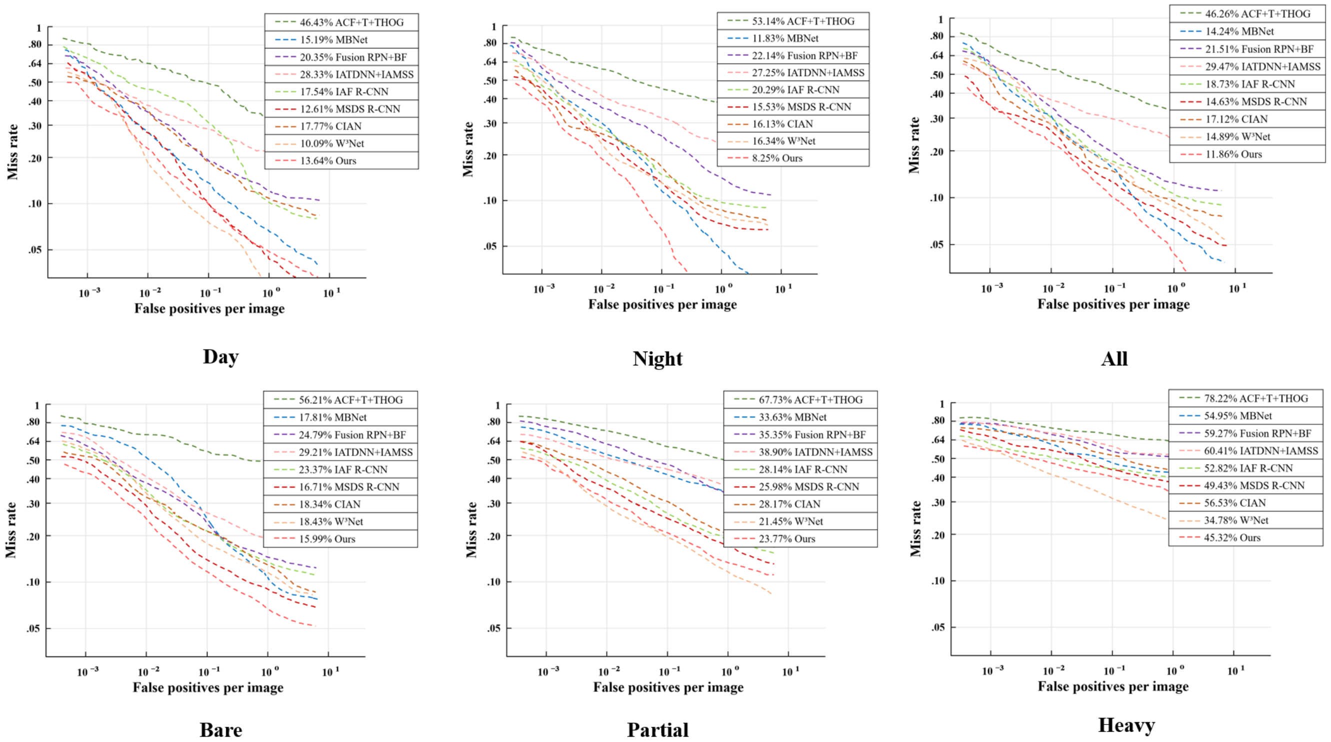
TABLE 2
COMPARATIVE EXPERIMENTAL RESULTS OF YOLOV5S, DUAL-MODAL YOLOV5S, FASTER R-CNN, DUAL-MODAL FASTER R-CNN, AND OURS ON THE KASIT DATASET.

Method	Data	AP (%)				Speed/s
		Day	Night	Bare	Partial	
YOLO v5s	RGB	92.35	85.79	91.43	87.64	0.017
	TIR	91.21	88.16	90.25	89.68	**0.016**
Dual-modal YOLO v5s	RGB+TIR	95.42	93.86	95.22	92.90	0.019
Faster R-CNN	RGB	95.48	89.32	95.67	90.41	0.214
	TIR	93.78	90.21	93.44	90.56	0.217
Dual-modal Faster R-CNN	RGB+TIR	**97.62**	95.29	**97.42**	95.21	0.532
DMFFNet	RGB+TIR	96.39	**95.38**	96.08	**95.87**	0.021

Test results in Table 2 show that for DMFFNet, the accuracy is 96.39% and 96.08% in good light and no occlusion, respectively. Even in the case of poor lighting or when the target is occluded, high accuracies of 95.38% and 95.87% are achieved, respectively. This may be because DMFFNet combines the advantages of VTI modal information and optimally uses image color, texture and location semantics, and other information, because of which, feature information of the target is more fully expressed, and the detection effect of bad light and occluded targets is significantly improved. Although the speed is lower, the accuracy is 1.52% and 2.97% higher than the dual-modal YOLO v5s, and 9.59% and 8.23% higher than the YOLO v5s single visible light, in cases of night and partial occlusion, respectively. In these cases, while the accuracy is similar to the dual-modal Faster R-CNN [43], the speed is much faster.

2) RESULTS ON THE VTI DATASET

To verify the performance of DMFFNet under smaller pedestrian scales and less favorable environments, the same evaluation was carried out on the VTI dataset. The results were compared with other mainstream VTI fusion algorithms; the MR-FPPI curve is shown in Figure 11, and the comparison results are shown in Table 3.



Day

Night

All

Bare

Partial

Heavy

FIGURE 11. MR-FPPI curves of ACF+T+THOG, MBNet, Fusion RPN+BF, IATDNN+IAMSS, IAF R-CNN, MSDS R-CNN, CIAN, W³Net and ours on the VTI dataset

TABLE 3
COMPARISON TEST RESULTS OF THE VTI DATASET.

Method	Speed (s)	MR^{-2} (%)					
		Reasonable			Occlusion		
		Day	Night	All	Bare	Partial	Heavy
ACF+T+THOG [24]	0.038	46.43	53.14	46.26	56.21	67.73	78.22
MBNet [25]	0.082	15.19	11.83	14.24	17.81	33.63	54.95
Fusion RPN+BF [26]	0.915	20.35	22.14	21.51	24.79	35.35	59.27
IATDNN+IAMSS [27]	0.353	28.33	27.25	29.47	29.21	38.90	60.41
IAF R-CNN [28]	0.315	17.54	20.29	18.73	23.37	28.14	52.82
MSDS-RCNN [29]	0.328	12.61	15.53	14.63	16.71	25.98	49.43
CIAN [30]	0.071	17.77	16.13	17.12	18.34	28.17	56.53
W^{3}Net [42]	0.0318	**10.09**	16.34	14.89	18.43	**21.45**	**34.78**
DMFFNet	**0.025**	13.64	**8.25**	**11.86**	**15.99**	23.77	45.32

Figure 11 shows the MR-FPPI curves of DMFFNet and other mainstream VTI fusion algorithms in the VTI human dataset.

Table 3 shows that the detection error rates of DMFFNet in the cases of day, night, and all-day are 13.64%, 8.25%, and 11.86%, respectively. DMFFNet exhibits better performance than other methods because we added the MFA and DDFF modules, which enhanced the expression of small targets and eliminated the noise of complex backgrounds under UAV vision. Some resulting plots are presented in Figure 12.

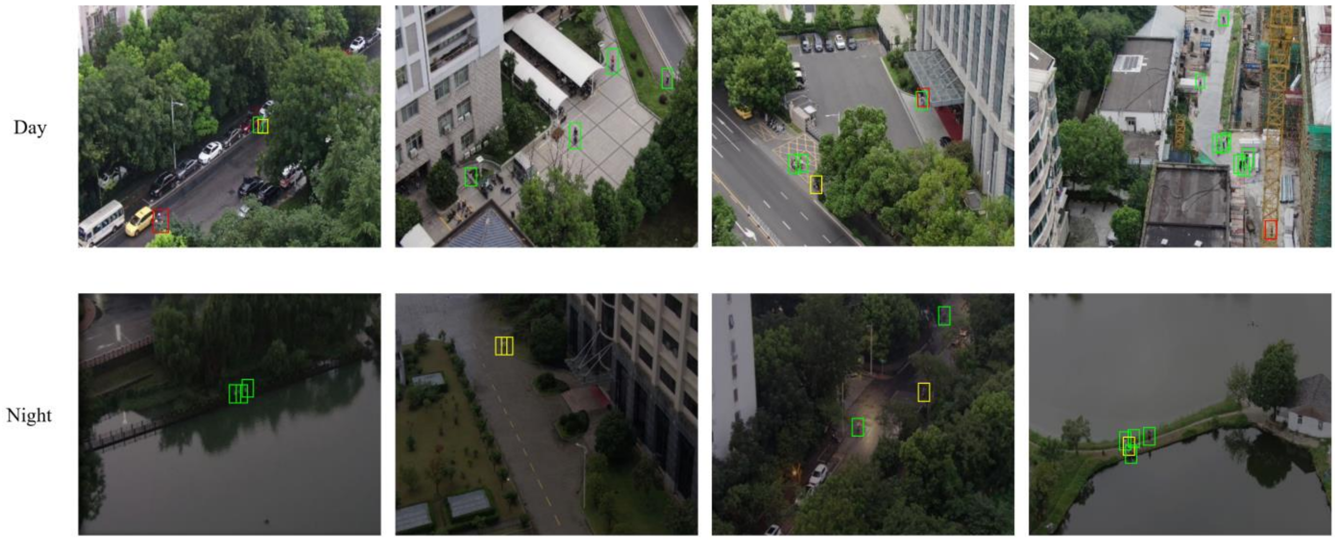


FIGURE 12. **Some detection results of our method on the VTI dataset. The dense annotations provided with the dataset such as green, yellow, and red boxes indicate no-occlusion, partial occlusion, and heavy occlusion respectively.**

Further, DMFFNet is compared with mainstream visible light target detection algorithms, and the comparison results are presented in Table 4. AP is adopted as the performance evaluation index.

TABLE 4
COMPARATIVE EXPERIMENTAL RESULTS OF YOLOV5S, DUAL-MODAL YOLOV5S, FASTER R-CNN, DUAL-MODAL FASTER R-CNN, AND OURS ON THE VTI DATASET.

Method	Data	AP (%)				Speed /s
		Day	Night	Bare	Partial	
YOLO v5s	RGB	87.45	81.79	86.42	84.71	0.019
	TIR	89.51	88.14	89.47	89.25	**0.018**
Dual-modal YOLO v5s	RGB+TIR	93.56	91.87	93.22	92.91	0.023
Faster R-CNN	RGB	90.58	89.72	90.77	88.61	0.277
	TIR	91.68	90.28	91.44	90.76	0.298
Dual-modal Faster R-CNN	RGB+TIR	95.82	94.37	**95.56**	94.33	0.617
DMFFNet	RGB+TIR	**95.91**	**94.88**	95.18	**94.87**	0.025

Test results in Table 2 show that for DMFFNet, the accuracy is 95.91% and 95.18% in good light and no occlusion, respectively. Even in the case of poor lighting or when the target is occluded, high accuracies of 94.88% and 94.87% are achieved, respectively. This may be because DMFFNet combines the advantages of VTI modal information and optimally uses information such as image color, texture, and location semantics, because of which, feature information of the target is more fully expressed, and the detection effect of bad light and occluded targets is significantly improved. Although it is slower, the accuracy is 2.35% and 3.01% higher than the dual-modal YOLO v5s, and 13.09% and 9.16% higher than the YOLO v5s single visible light, in cases of night and partial occlusion, respectively. In the case of night and partial occlusion, while the accuracy is similar to the dual-modal Faster R-CNN [37], the speed is much faster.

3) RESULTS COMBINED DATASET

To test the generalization and robustness of our model. We did comparative experiments on a single dataset (KAIST, VIT) and a mixed dataset (KAIST+VIT). Combining datasets is simply a fusion of the KAIST training set and the VIT training set. And test the performance of DMFFNet trained on a single dataset and a combined dataset on the KAIST test set and the VIT test set, respectively. The results are shown in Table 5.

TABLE 5
COMPARATIVE EXPERIMENTS ON SINGLE DATASETS (KAIST, VIT) AND COMBINED DATASETS (KAIST+VIT).

Training set	MR^{-2} (%)	
	Test set	
	KAIST	VIT
KAIST	9.26	21.79
VIT	17.76	11.86
KAIST+VIT	9.87	12.98

It can be seen from Table 5 that our model has certain generalization and robustness, and through the training of mixed datasets, the generalization and randomness of the model are better.

D. Ablation experiments

In this section, we validate the impact of each module in DMFFNet on the object detection performance on a subset (Reasonable and Occlusion (Bare, Partial, Heavy)) of the KAIST dual-light pedestrian dataset. Table 6 shows the ablation results of adding each module (MFA, DDFF, DFF) on MobileNet v3.

TABLE 6
EXPERIMENTAL RESULTS OF ABLATION ON THE KAIST DATASET

Baseline	MFA	DDFF	DFF	MR^{-2} (%)						Speed/s
				Reasonable		Occlusion				
				Day	Night	All	Bare	Partial	Heavy	
✓				16.69	43.48	35.44	32.93	40.12	60.23	0.012

✓	✓			14.59	37.25	28.16	26.73	34.30	57.26	0.014
✓		✓		15.88	39.66	31.02	29.80	37.52	58.86	0.014
✓			✓	14.22	9.66	13.77	14.44	28.43	51.43	0.014
✓	✓	✓		15.71	33.05	26.76	25.30	31.52	54.65	0.018
✓	✓		✓	13.60	7.33	11.86	12.85	25.29	49.50	0.020
✓		✓	✓	13.87	7.38	11.50	13.79	29.94	53.71	0.019
✓	✓	✓	✓	**12.79**	**5.17**	**9.26**	**10.15**	**23.35**	**47.31**	**0.021**

Table 5 shows the MR^{-2} and inference speed of different modules on Reasonable and Occlusion images; without adding any module, the values of MR^{-2} of day, night, and all-day are high, 16.69%, 43.48%, and 35.44%, respectively. In the case of almost no occlusion, partial occlusion, and severe occlusion, the value of MR^{-2} are high, 32.93%, 40.12%, and 60.23%, respectively. After adding the MFA and DDFF modules, the values of MR^{-2} decreased slightly, and after adding the DFF module, the values of MR^{-2} decreased considerably. The values of MR^{-2} of day, night, and all-day were 14.22%, 9.66%, and 13.77%, respectively, and for almost no occlusion, partial occlusion, and severe occlusion were 14.44%, 28.43%, and 51.43%, respectively. When all modules are added, the values of MR^{-2} of day, night, and all-day are only 12.79%, 5.17%, and 9.26%, respectively. With almost no occlusion, partial occlusion, and severe occlusion, the values are only 10.15%, 23.35%, and 47.31%, respectively, and the detection speed is slightly increased to 0.021 s.

The experimental results show that under dim light conditions, the visible feature information of the target in visible light gradually decreases, which makes target recognition difficult. By extracting the feature information of the infrared image target, the highlighted contour information displayed by the target can be effectively used, and the visible light image information can be effectively supplemented. Moreover, the MFA module can effectively extract more fine-grained multi-scale spatial information and establish longer-distance channel dependencies. The DDFF module can effectively fuse multi-scale feature information on a deeper level, because of which, the semantic information and geometric detail representation abilities of the feature map are significantly enhanced. Finally, the values of MR^{-2} for day, night, and all-day decreased by 3.9%, 38.31%, and 26.18%, respectively, compared to the case without any module. The values of MR^{-2} with almost no occlusion, partial occlusion, and severe occlusion decreased by 22.78%, 16.77%, and 12.92%, respectively, with no module added. The feature map comparison of the ablation experiment is shown in Figure 13.

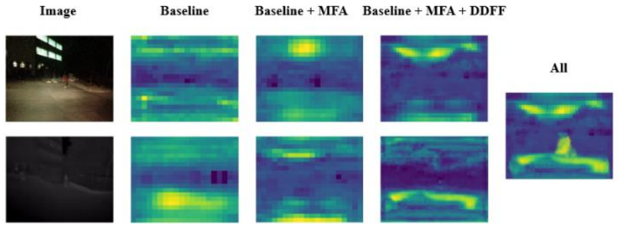


FIGURE 13. **The feature map comparison of Baseline, Baseline+MFA, Baseline+MFA+DDFF, and All module.**

V. Conclusion and Future Work

This paper presents a pedestrian detection network for scenarios with dim lighting, occluded objects, and cluttered backgrounds. It comprises a MobileNet v3 backbone network, an MFA module, a DDFF module, and a DFF module. Our proposed method significantly reduces the missed detection rate of pedestrian detection in unfavorable environments such as dim light and exhibits excellent real-time detection performance.

In the future, we will optimize our network structure to further reduce our missed detection rate and improve real-time detection performance, in addition to applying our method for other object detection tasks.

REFERENCES

[1] P. Zheng, H. Bai, and W. Li, "Small target detection algorithm in complex background," *Journal of ZheJiang University (Engineering Science)*, vol. 54, no. 9, pp. 1777-1784, 2020.

[2] D.-H. Chen, Y.-D. Cao, and J. Yan, "Towards Pedestrian Target Detection with Optimized Mask R-CNN," *Complexity*, vol. 2020, pp. 1-8, 2020/12/22 2020, doi: 10.1155/2020/6662603.

[3] Q.-C. Mao, H.-M. Sun, L.-Q. Zuo, and R.-S. Jia, "Finding every car: a traffic surveillance multi-scale vehicle object detection method," *Applied Intelligence*, vol. 50, no. 10, pp. 3125-3136, 2020/05/05 2020, doi: 10.1007/s10489-020-01704-5.

[4] DPSNet: MultiTask Learning Using Geometry Reasoning for Scene Depth and Semantics [J]. IEEE Transactions on Neural Networks and Learning Systems，2021.

[5] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," presented at the *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2005.177.

[6] W. Bo and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors," presented at the *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, 2005. [Online]. Available: http://dx.doi.org/10.1109/iccv.2005.74.

[7] Zhang, S., Yang, J., Schiele, B.: Occluded pedestrian detection through guided attention in cnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6995–7003 (2018).

[8] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1037-1045.

[9] Thuong L T, Quang T H D, Vy H Y, et al. GAN-based Thermal Infrared Image Colorization for Enhancing Object Identification[C]//2021 International Symposium on Electrical and Electronics Engineering (ISEE). IEEE, 2021:90-94.

[10] T. Le-Tien, T. H. Duy Quang, H. Y. Vy, T. Nguyen-Thanh, and H. Phan-Xuan, "GAN-based Thermal Infrared Image Colorization for Enhancing Object Identification," presented at the *2021 International Symposium on Electrical and Electronics Engineering (ISEE)*, 2021/04/15, 2021. [Online]. Available: http://dx.doi.org/10.1109/isee51682.2021.9418801.

[11] A. Howard *et al.*, "Searching for MobileNetV3," presented at the *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*,

2019/10, 2019. [Online]. Available: http://dx.doi.org/10.1109/iccv.2019.00140.

[12] A. Shashua, Y. Gdalyahu, and G. Hayun, "Pedestrian detection for driving assistance systems: single-frame classification and system level performance," presented at the *IEEE Intelligent Vehicles Symposium, 2004*. [Online]. Available: http://dx.doi.org/10.1109/ivs.2004.1336346.

[13] L. Havasi, Z. Szlávik, and T. Szirányi, "PEDESTRIAN DETECTION USING DERIVED THIRD-ORDER SYMMETRY OF LEGS A novel method of motion-based information extraction from video image-sequences," in *Computational Imaging and Vision*, ed: Kluwer Academic Publishers, pp. 733-739.

[14] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian Detection via Classification on Riemannian Manifolds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1713-1727, 2008/10 2008, doi: 10.1109/tpami.2008.75.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," presented at the *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014/06, 2014. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2014.81.

[16] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. Ogale, and D. Ferguson, "Real-Time Pedestrian Detection with Deep Network Cascades," presented at the *Procedings of the British Machine Vision Conference 2015*, 2015. [Online]. Available: http://dx.doi.org/10.5244/c.29.32.

[17] W. Ouyang and X. Wang, "Joint Deep Learning for Pedestrian Detection," presented at the *2013 IEEE International Conference on Computer Vision*, 2013/12, 2013. [Online]. Available: http://dx.doi.org/10.1109/iccv.2013.257.

[18] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," in *Computer Vision – ECCV 2016*, ed: Springer International Publishing, 2016, pp. 21-37.

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," presented at the *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016/06, 2016. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2016.91.

[20] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," presented at the *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017/07, 2017. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2017.690.

[21] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[22] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[23] W. Zhao, W. Ma, L. Jiao, P. Chen, S. Yang, and B. Hou, "Multi-Scale Image Block-Level F-CNN for Remote Sensing Images Object Detection," *IEEE Access*, vol. 7, pp. 43607-43621, 2019, doi: 10.1109/access.2019.2908016.

[24] X. Sun, P. Wang, C. Wang, Y. Liu, and K. Fu, "PBNet: Part-based convolutional neural network for complex composite object detection in remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 50-65, 2021/03 2021, doi: 10.1016/j.isprsjprs.2020.12.015.

[25] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, "Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 161, pp. 294-308, 2020/03 2020, doi: 10.1016/j.isprsjprs.2020.01.025.

[26] Q. Zhang *et al.*, "Dense Attention Fluid Network for Salient Object Detection in Optical Remote Sensing Images," *IEEE Transactions on Image Processing*, vol. 30, pp. 1305-1317, 2021, doi: 10.1109/tip.2020.3042084.

[27] G. Cheng, Y. Si, H. Hong, X. Yao, and L. Guo, "Cross-Scale Feature Fusion for Object Detection in Optical Remote Sensing Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 3, pp. 431-435, 2021/03 2021, doi: 10.1109/lgrs.2020.2975541.

[28] Zhou K, Chen L, Cao X. Improving multispectral pedestrian detection by addressing modality imbalance problems[C]//European Conference on Computer Vision. Springer, Cham, 2020: 787-803.

[29] D. Konig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully Convolutional Region Proposal Networks for Multispectral Person Detection," presented at the *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017/07, 2017. [Online]. Available: http://dx.doi.org/10.1109/cvprw.2017.36.

[30] 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," *Information Fusion*, vol. 50, pp. 148-157, 2019/10 2019, doi: 10.1016/j.inffus.2018.11.017.

[31] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *Pattern Recognition*, vol. 85, pp. 161-171, 2019/01 2019, doi: 10.1016/j.patcog.2018.08.005.

[32] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," *arXiv preprint arXiv:1808.04818*, 2018.

[33] L. Zhang *et al.*, "Cross-modality interactive attention network for multispectral pedestrian detection," *Information Fusion*, vol. 50, pp. 20-29, 2019/10 2019, doi: 10.1016/j.inffus.2018.09.015.

[34] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[35] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.

[36] F. Li, R. Feng, W. Han, and L. Wang, "Ensemble model with cascade attention mechanism for high-resolution remote sensing image scene classification," *Optics Express*, vol. 28, no. 15, p. 22358, 2020/07/14 2020, doi: 10.1364/oe.395866.

[37] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Information Fusion*, vol. 76, pp. 323-336, 2021/12 2021, doi: 10.1016/j.inffus.2021.06.008.

[38] J. Shen, C. Zhang, Y. Zheng, and R. Wang, "Decision-Level Fusion with a Pluginable Importance Factor Generator for Remote Sensing Image Scene Classification," *Remote Sensing*, vol. 13, no. 18, p. 3579, 2021/09/08 2021, doi: 10.3390/rs13183579.

[39] L. Fu, W.-B. Gu, Y.-B. Ai, W. Li, and D. Wang, "Adaptive spatial pixel-level feature fusion network for multispectral pedestrian detection," *Infrared Physics & Technology*, vol. 116, p. 103770, 2021/08 2021, doi: 10.1016/j.infrared.2021.103770.

[40] S. S. Malik, S. P. P. Kumar, and G. B. Maruthi, "DT-CWT: Feature level image fusion based on dual-tree complex wavelet transform," presented at the *International Conference on Information Communication and Embedded Systems (ICICES2014)*, 2014/02, 2014. [Online]. Available: http://dx.doi.org/10.1109/icices.2014.7033982.

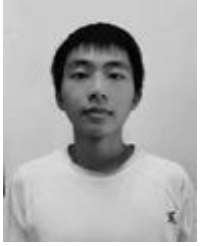
[41] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," Stanford, 2006.

[42] Luo Y, Zhang C, Zhao M, et al. Where, What, Whether: Multi-modal learning meets pedestrian detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 14065-14073.

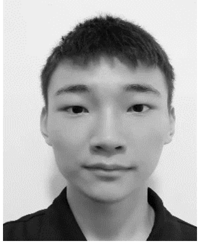
[43] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017/06/01 2017, doi: 10.1109/tpami.2016.2577031.

RUIZHE HU was born in 1999, he graduated from Wuhan University of Technology in 2020 with a bachelor's degree in mechanical engineering. He is currently studying for a master's degree in the Field Engineering College of PLA Army Engineering University. His research interests include computer vision and adversarial example.

Ting Rui received the M.S. degree and Ph.D. from PLA University of Science and Technology, Nanjing, China in1998 and 2001, respectively. Ting RUI is Professor of Army engineering University of PLA. He mainly applies computer vision, machine learning, multimedia, and video surveillance. He has authored and co-authored more than 80 scientific articles.



YAN OUYANG was born in Hengyang, Hunan, China, in 1998. He is currently pursuing a master’s degree in mechanical engineering from the Army Engineering University. His research interests include deep learning, reinforcement learning and computer vision.



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

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

Yinan Du received a bachelor's degree in process equipment and control engineering from Hefei University of technology in 2020. He is currently studying for a master's degree in mechanical engineering at the Army Engineering University. At present, the research field are fine-grained recognition and target detection.