Cascaded information enhancement and cross-modal attention feature fusion for multispectral pedestrian detection

Yang Yang[1], Kaixiong Xu[1] and Kaizheng Wang[2]*

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China, [2]Faculty of Electrical Engineering, Kunming University of Science and Technology, Kunming, China

Multispectral pedestrian detection is a technology designed to detect and locate pedestrians in Color and Thermal images, which has been widely used in automatic driving, video surveillance, etc. So far most available multispectral pedestrian detection algorithms only achieved limited success in pedestrian detection because of the lacking take into account the confusion of pedestrian information and background noise in Color and Thermal images. Here we propose a multispectral pedestrian detection algorithm, which mainly consists of a cascaded information enhancement module and a cross-modal attention feature fusion module. On the one hand, the cascaded information enhancement module adopts the channel and spatial attention mechanism to perform attention weighting on the features fused by the cascaded feature fusion block. Moreover, it multiplies the single-modal features with the attention weight element by element to enhance the pedestrian features in the single-modal and thus suppress the interference from the background. On the other hand, the cross-modal attention feature fusion module mines the features of both Color and Thermal modalities to complement each other, then the global features are constructed by adding the cross-modal complemented features element by element, which are attentionally weighted to achieve the effective fusion of the two modal features. Finally, the fused features are input into the detection head to detect and locate pedestrians. Extensive experiments have been performed on two improved versions of annotations (sanitized annotations and paired annotations) of the public dataset KAIST. The experimental results show that our method demonstrates a lower pedestrian miss rate and more accurate pedestrian detection boxes compared to the comparison method. Additionally, the ablation experiment also proved the effectiveness of each module designed in this paper.

KEYWORDS

multispectral pedestrian detection, attention mechanism, feature fusion, convolutional neural network, background noise

1 Introduction

Pedestrian detection, parsing visual content to identify and locate pedestrians on an image/video, has been viewed as an essential and central task within the computer vision field and widely employed in various applications, e.g. autonomous driving, video surveillance and person re-identification [1–7]. The performance of such technology has greatly advanced through the facilitation of convolutional neural networks (CNN). Typically, pedestrian detectors take Color images as input and try to retrieve the pedestrian information from
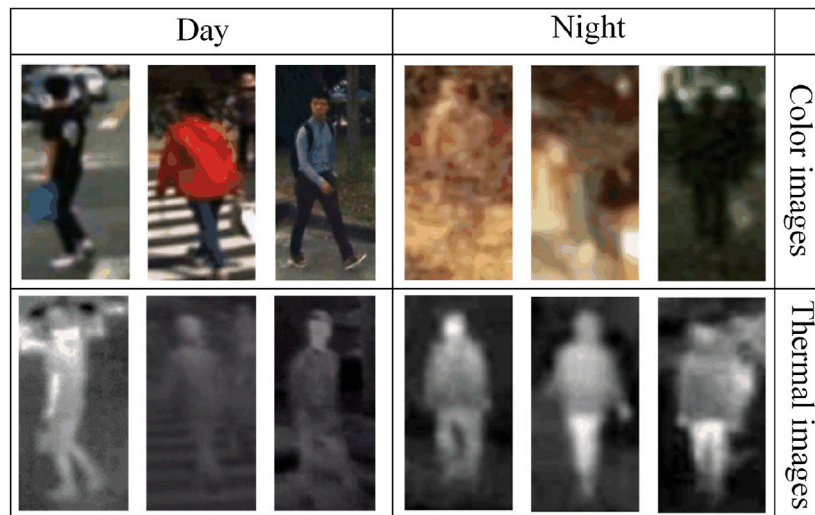
FIGURE 1
Example of color and thermal images of pedestrians in daytime and nighttime scenes.

them. However, the quality of Color images highly depends on the light condition. Missing recognition of pedestrians occurs frequently when pedestrian detectors process Color images with poor resolution and contrast caused by unfavorable lighting. Consequently, the use of such models has been limited for the application of all-weather devices.

Thermal imaging is related to the infrared radiation of pedestrians, barely affected by changes in ambient light. The technique of combining Color and Thermal images has been explored in recent years [8–16]. These methods has been shown to exhibit positive effects on pedestrian detection performance in complex environments as it could retrieve more pedestrian information. However, despite important initial success, there remain two major challenges. First, as shown in Figure 1, the image of pedestrians tends to blend with the background for nighttime Color images resulting from insufficient light [17], and for daytime Thermal images as well due to similar temperatures between the human body and the ambient environment [18]. Second, there is an essential difference between Color images and Thermal images the former displays the color and texture detail information of pedestrians while the latter shows the temperature information. Therefore, solutions needed to be taken to augment the pedestrian features in Color and Thermal modalities in order to suppress background interference, and enable better integration and understanding of both Color and Thermal images to improve the accuracy of pedestrian detection in complex environments.

To address the challenges above, the researches [19,20] designed illumination-aware networks to obtain illumination-measured parameters of Color and Thermal images respectively, which were used as fusion weights for Color and Thermal features in order to realize a self-adaptively fuse of two modal features. However, the acquisition of illumination-measured parameters relied heavily on the classification scores, the accuracy of which was limited by the performance of the classifier. [21] reported confidence-aware networks to predict the confidence of detection boxes for each modal, and then Dempster-Sheffer theory combination rules were employed to fuse the results of different branches based on

uncertainty. Nevertheless, the accuracy of predicting the detection boxes' confidence is also affected by the performance of the confidence-aware network. A cyclic fusion and refinement scheme was introduced by [22] for the sake of gradually improving the quality of Color and Thermal features and automatically adjusting the complementary and consistent information balance of the two modalities to effectively utilize the information of both modalities. However, this method only used a simple feature cascade operation to fuse Color and Thermal features and failed to fully exploit the complementary features of these two modalities.

To tackle the problems aforementioned, we propose a multispectral pedestrian detection algorithm with cascaded information enhancement and cross-modal attention feature fusion. The cascaded information enhancement module (CIEM) is designed to enhance the pedestrian information suppressed by the background in the Color and Thermal images. CIEM uses a cascaded feature fusion block to fuse Color and Thermal features to obtain fused features of both modalities. Since the fused features contain the consistency and complementary information of Color and Thermal modalities, the fused features can be used to enhance Color and Thermal features respectively to reduce the interference of background on pedestrian information. Inspired by the attention mechanism, the attention weights of the fused features are sequentially obtained by channel and spatial attention learning, and the Color and Thermal features are multiplied with the attention weights element by element, respectively. In this way, the single-modal features have the combined information of the two modalities, and the single-modal information is enhanced from the perspective of the fused features. Although CIEM enriches single-modal pedestrian features, simple feature fusion of the enhanced single-modal features is still insufficient for robust multispectral pedestrian detection. Thus, we design the cross-modal attention feature fusion module (CAFFM) to efficiently fuse Color and Thermal features. Cross-modal attention is used in this module to implement the differentiation of pedestrian features between different modalities. In order to supplement the pedestrian information of the other modality to the local modality, the attention of the other

modality is adopted to augment the pedestrian features of the local modality. A global feature is constructed by adding the Color and Thermal features after performing cross-modal feature enhancement, and the global feature is used to guide the fusion of the Color and Thermal features. Overall, the method presented in this paper enables more comprehensive pedestrian features acquisition through cascaded information enhancement and cross-modal attention feature fusion, which effectively enhances the accuracy of multispectral image pedestrian detection. The main contributions of this paper are summarized as follows.

(1) A cascaded information enhancement module is proposed. From the perspective of fused features, it reduces the interference from the background of Color and Thermal modalities on pedestrian detection and augments the pedestrian features of Color and Thermal modalities separately through an attention mechanism.

(2) The designed cross-modal attention feature fusion module first mines the features of both Color and Thermal modalities separately through a cross-modal attention network and adds them to the other modality for cross-modal feature enhancement. Meanwhile, the cross-modal enhanced Color and Thermal features are used to construct global features to guide the feature fusion of the two modalities.

(3) Numerous experiments are conducted on the public dataset KAIST to demonstrate the effectiveness and superiority of the proposed method. In addition, the ablation experiments also demonstrate the effectiveness of the proposed modules.

2 Related works

2.1 Multispectral pedestrian detection

Multispectral sensors can obtain paired Color-Thermal images to provide complementary information about pedestrian targets. A large multispectral pedestrian detection (KAIST) dataset was constructed by [8]. Meanwhile, by combining the traditional aggregated channel feature (ACF) pedestrian detector [23] with the HOG algorithm [24], an extended ACF (ACF + T + THOG) method was proposed to fuse Color and Thermal features. In 2016, [9] proposed four fusion modalities of low-layer feature, middle-layer feature, high-layer feature, and confidence fraction fusion with VGG16 as the backbone network, and the middle-layer feature fusion was proved to offer the maximum integration capability of Color and Thermal features. Inspired by this, [25] developed a multispectral region candidate network with Faster RCNN (Region with CNN features, RCNN) [26] as the architecture and replaced the original classifier in Faster RCNN with an enhanced decision tree classifier to reduce the missed and false detection of pedestrians. Recently,[27] deployed the EfficientDet as the backbone network and proposed an EfficientDet-based fusion framework for multispectral pedestrian detection to improve the detection accuracy of pedestrians in Color and Thermal images by adding and cascading the Color and Thermal features. Although the studies [8,9,25,27] fused Color and Thermal features for pedestrian detection, they mainly focused on exploring the impact of different stages of fusion on pedestrian detection, and only adopted simple feature fusion and not focusing on the case of pedestrian and background confusion.

In 2019, [28] observed a weak alignment problem of pedestrian position between Color and Thermal images, for which the KAIST dataset was re-annotated and Aligned Region CNN (AR-CNN) was proposed to handle weakly aligned multispectral pedestrian detection data in an end-to-end manner. But the deployment of this algorithm requires pairs of annotations, and the annotation of the dataset is a time-consuming and labor-intensive task, which makes the algorithm difficult to be applied in realistic scenes. [29] proposed a new single-stage multispectral pedestrian detection framework. This framework used multi-label learning to learn input state-aware features based on the state of the input image pair by assigning an individual label (if the pedestrian is visible in only one image of the image pair, the label vector is assigned as $y_1 \in [0, 1]$ or $y_2 \in [1, 0]$; if the pedestrian is visible in both images of the image pair, the label vector is assigned as $y_3 \in [1, 1]$) to solve the problem of weak alignment of pedestrian locations between Color and Thermal images, but the model still requires pairs of annotations during training. [19] designed illumination-aware networks to obtain illumination-measured parameters for Color and Thermal images separately and used them as the fusion weights for Color and Thermal features. [20] designed a differential modality perception fusion module to guide the features of the two modalities to become similar, and then used the illumination perception network to assign fusion weights to the Color and Thermal features. [30] reported an uncertainty-aware cross-modal guidance (UCG) module to guide the distribution of modal features with high prediction uncertainty to align with the distribution of modal features with low prediction uncertainty. The researches [19,20] noticed that the pedestrians in Color and Thermal images are easily confused with the background and used illumination-aware networks to assign fusion weights to Color and Thermal features. However, the acquisition of illumination-measured parameters relied heavily on the classification scores, whose accuracy was limited by the performance of the classifier. In contrast, the method proposed in this paper not only considers the confusion of pedestrians and background in Color and Thermal images but also effectively fuses the two modal features.

2.2 Attention mechanisms

Attention mechanisms [31] utilized in computer vision are aimed to perform the processing of visual information. Currently, attention mechanisms have been widely used in semantic segmentation [32], image captioning [33], image fusion [34,35], image dehazing [36], saliency target detection [37], person re-identification [38–40], etc. [41] introduced the idea of a squeeze and excitation network (SENet) to simulate the interdependence between feature channels in order to generate channel attention to recalibrate the feature mapping of channel directions. [42] employed the use of a selective kernel unit (SKNet) to adaptively fuse branches with different kernel sizes based on input information. A work inspired by this was from [43]. They designed a multi-scale channel attention feature fusion network that used channel attention mechanisms to replace simple fusion operations such as feature cascades or summations in feature fusion to produce richer feature representations. However, this recent progress in multispectral pedestrian detection has also been limited to two main challenges the interference caused by background and the difference of fundamental characteristics in Color and Thermal images. Therefore, we propose a multispectral pedestrian detection algorithm with cascaded information enhancement and

FIGURE 2
Overall framework of the proposed algorithm.
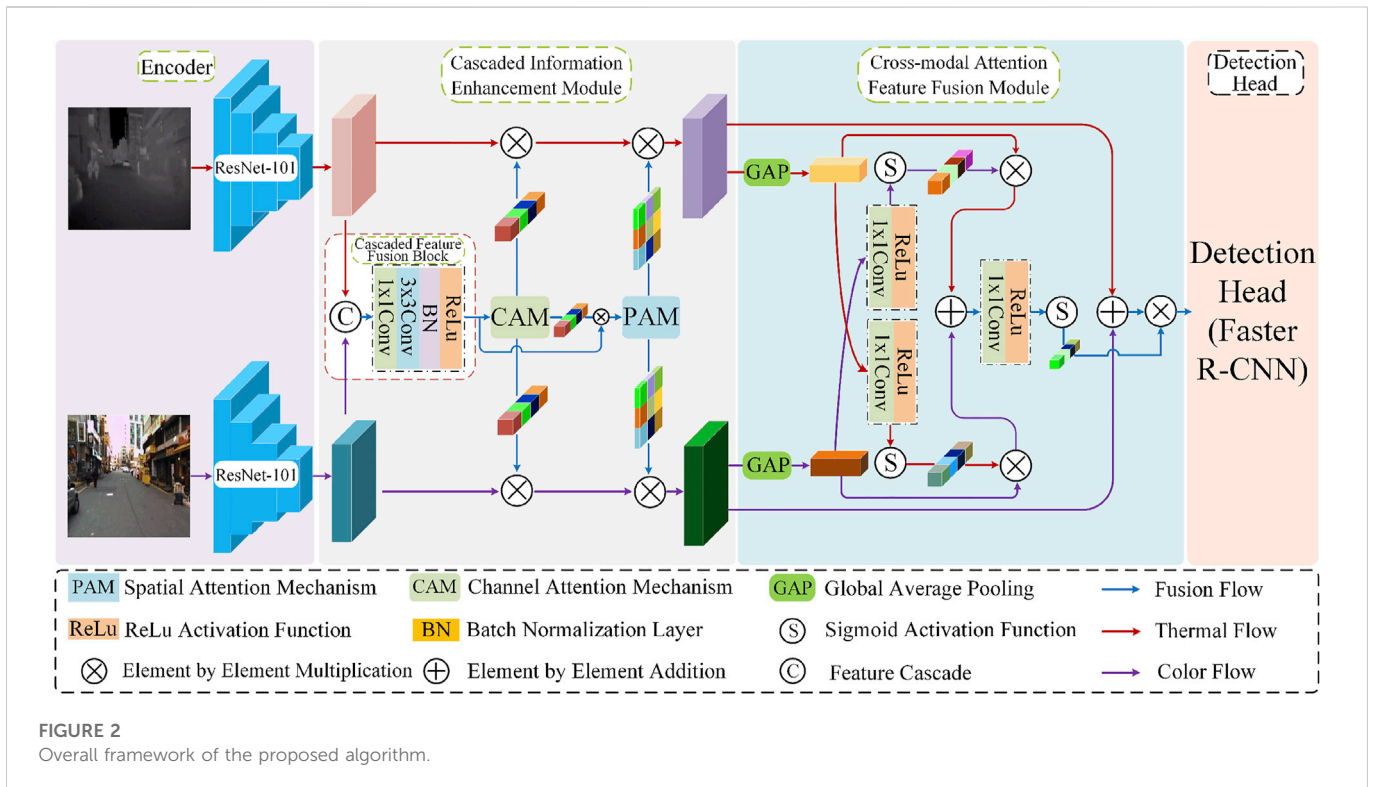
cross-modal attention feature fusion based on the attention mechanism.

3 Methods

The overall network framework of the proposed algorithm is shown in Figure 2. The network consists of an encoder, a cascaded information enhancement module (CIEM), a cross-modal attentional feature fusion module (CAFFM) and a detection head. Specifically, ResNet-101 [44] is used as the backbone network of the encoder to encode the features of the input Color images X_c and Thermal images X_t to obtain the corresponding feature maps $F_c \in \mathrm{R}^{W \times H \times C}$ and $F_t \in \mathrm{R}^{W \times H \times C}$ (W, H, C represent the width, height and the number of channels of the feature maps, respectively). CIEM enhances single-modal information from the perspective of fused features by cascading feature fusion blocks to fuse F_c and F_t, and attention weighting the fused features to enrich pedestrian features. CAFFM complements the features of different modalities by mining the complementary features between the two modalities and constructs global features to guide the effective fusion of the two modal features. The detection head is employed for pedestrian recognition and localization of the final fused features.

3.1 Cascaded information enhancement module

Considering the confusion of pedestrians with the backgrounds in Color and Thermal images, we design a cascaded information enhancement module (CIEM) to augment

the pedestrian features of both modalities to mitigate the effect of background interference on pedestrian detection. Specifically, a cascaded feature fusion block is used to fuse the Color features F_c and Thermal features F_t. The cascaded feature fusion block consists of feature cascade, 1×1 convolution, 3×3 convolution, BN layer, and $ReLu$ activation function. The feature cascade operation splice F_c and F_t along the direction of channels. 1×1 convolution is conducive to cross-channel feature interaction in the channel dimension and reducing the number of channels in the splice feature map, while 3×3 convolution expands the field of perception and makes a more comprehensive fusion of features for generating fusion features F_{ct}:

$$F_{ct} = ReLu\left(BN\left(Conv_3\left(Conv_1\left[F_c, F_t\right]\right)\right)\right) \tag{1}$$

where BN denotes batch normalization, $Conv_n(\cdot)$ denotes a convolution kernel with kernel size $n \times n$, $[\cdot, \cdot]$ denotes the cascade of features along the channel direction, $ReLu(\cdot)$ represents $ReLu$ activation function. Fusion feature F_{ct} is used to enhance the single-modal information because F_{ct} combines the consistency and complementarity of the Color features F_c and Thermal features F_t. The use of F_{ct} for enhancing the single-modal feature can reduce the interference of the noise in the single-modal features (for example, it is difficult to distinguish between the pedestrian information and the background noise).

In order to effectively enhance pedestrian features, the fusion feature F_{ct} is sent into the channel attention module (CAM) and spatial attention module (PAM) [45] to make the network pay attention to pedestrian features. The network structure of CAM and PAM is shown in Figure 3. F_{ct} first learns the channel attention weight $w_{ca} \in \mathrm{R}^{1 \times 1 \times C}$ by CAM, then uses w_{ca} to weight F_{ct}, and the spatial attention weight $w_{pa} \in \mathrm{R}^{W \times H \times 1}$ is obtained from the weighted features by PAM.
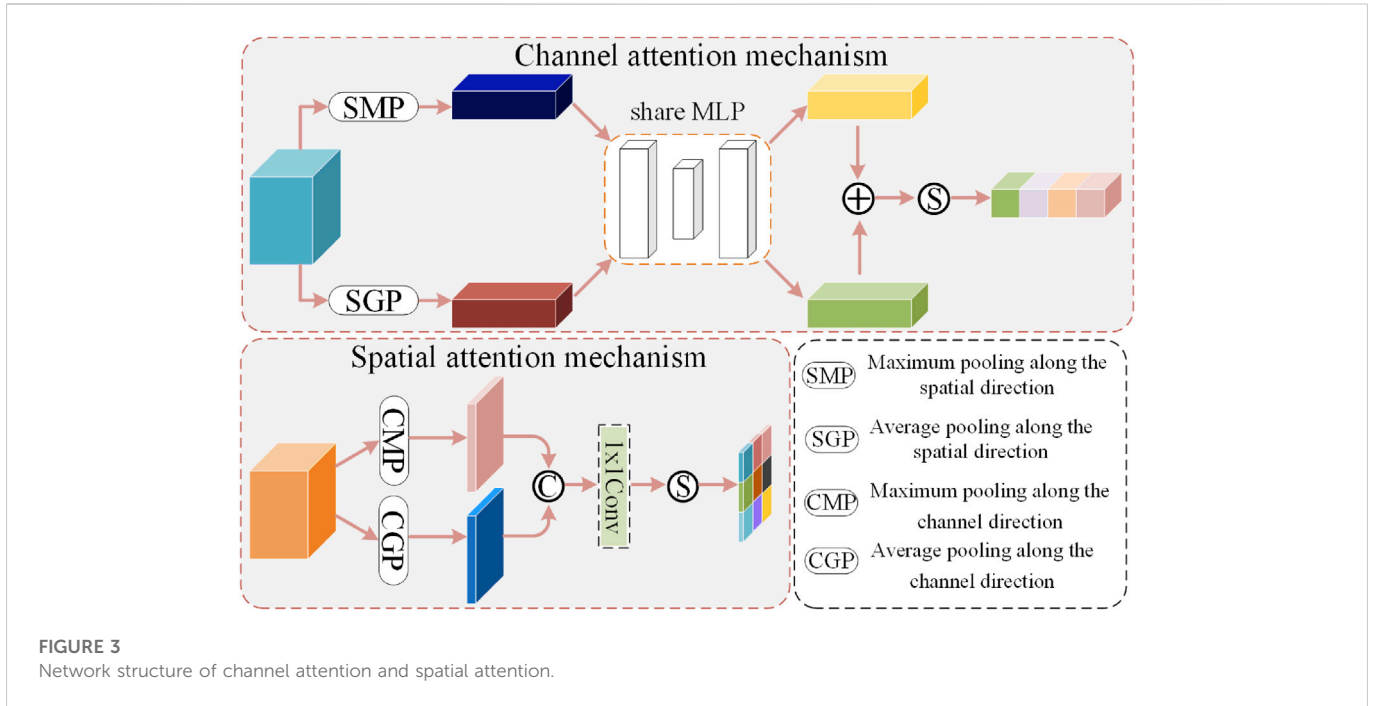
FIGURE 3
Network structure of channel attention and spatial attention.

The single-modal Color features F_c and Thermal features F_t are multiplied element by element with the attention weights w_{ca} and w_{pa} to enhance the single-modal features from the perspective of fused features. The whole process can be described as follows:

$$F_t' = (F_t \otimes w_{ca}) \otimes w_{pa} \qquad (2)$$

$$F_c' = (F_c \otimes w_{ca}) \otimes w_{pa} \qquad (3)$$

where F_t' and F_c' denote the Color features and Thermal features obtained by the cascaded information enhancement module, respectively. $\otimes$ represents the element by element multiplication.

3.2 Cross-modal attention feature fusion module

There is an essential difference between Color and Thermal images, Color images reflect the color and texture detail information of pedestrians while Thermal images contain the temperature information of pedestrians, however, they also have some complementary information. In order to explore the complementary features of different image modalities and fuse them effectively, we design a cross-modal attention feature fusion module.

Specifically, the Color features F_c' and Thermal features F_t' enhanced by CIEM are first mapped into feature vectors $v_c \in R^{1 \times 1 \times C}$ and $v_t \in R^{1 \times 1 \times C}$, respectively, by using global average pooling operation. The cross-modal attention network consists of a set of symmetric 1×1 convolutions, $ReLu$ activation functions, and $Sigmoid$ activation functions. In order to obtain the complementary features of the two modalities, more pedestrian features need to be mined from the single-modal. The feature vectors v_t and v_c are learned to the respective modal attention weights $w_t \in R^{1 \times 1 \times C}$ and $w_c \in R^{1 \times 1 \times C}$ by a

cross-modal attention network, and then the Color features F_c' are multiplied element by element with the attention weights w_t of the Thermal modality, and the Thermal features F_t' are multiplied element by element with the attention weights w_c of the Color modality to complement the features of the other modality into the present modality. The specific process can be expressed as follows.

$$w_t = Sigmoid\left(ReLu\left(Conv_1\left(GAP\left(F_t'\right)\right)\right)\right) \qquad (4)$$

$$F_{ct}' = w_t \otimes GAP\left(F_c'\right) \qquad (5)$$

$$w_c = Sigmoid\left(ReLu\left(Conv_1\left(GAP\left(F_c'\right)\right)\right)\right) \qquad (6)$$

$$F_{tc}' = w_c \otimes GAP\left(F_t'\right) \qquad (7)$$

where F_{ct}' denotes Color features after supplementation with Thermal features, F_{tc}' denotes Thermal features after supplementation with Color features, $GAP(\cdot)$ denotes global average pooling operation, $Conv_1(\cdot)$ denotes convolution with convolution kernel size 1×1, $ReLu(\cdot)$ denotes $ReLu$ activation operation, and $Sigmoid\ (\cdot)$ denotes $Sigmoid$ activation operation.

In order to efficiently fuse the two modal features, the features F_{ct}' and F_{tc}' are subjected to an element by element addition operation to obtain a global feature vector containing Color and Thermal features. Then, the features F_t' and F_c' are added element by element and multiplied with the attention weight w_{ct} of the global feature vector element by element to guide the fusion of Color and Thermal features from the perspective of global features to obtain the final fused feature F. The fused feature F is input to the detection head to obtain the pedestrian detection results. The feature fusion process can be expressed as follows:

$$w_{ct} = Sigmoid\left(ReLu\left(Conv_1\left(F_{ct}' \oplus F_{tc}'\right)\right)\right) \qquad (8)$$

$$F = w_{ct} \otimes \left(F_t' \oplus F_c'\right) \qquad (9)$$

where $\oplus$ denotes element by element addition.

3.3 Loss function

The loss function in this paper is consistent with the literature [26] and uses the Region Proposal Network (RPN) loss function L_{RPN} and Fast RCNN [46] loss function L_{FR} to jointly optimize the network:

$$L = L_{RPN} + L_{FR} \qquad (10)$$

Both L_{RPN} and L_{FR} consist of classification loss L_{cls} and bounding box regression loss L_{reg}:

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \qquad (11)$$

Where, N_{cls} is the number of anchors, N_{reg} is the sum of positive and negative sample number, p_i is the probability that the i-th anchor is predicted to be the target, p_i^* is 1 when the anchor is a positive sample, otherwise it is 0, t_i denotes the bounding box regression parameter predicting the i-th anchor, and t_i^* denotes the GT bounding box parameter of the i-th anchor, $\lambda = 1$.

The difference between the classification loss of RPN network and Fast RCNN network is that the RPN network focuses only on the foreground and background when classifying, so its loss is a binary cross-entropy loss, while the Fast RCNN classification is focused to the target category and is a multi-category cross-entropy loss:

$$L_{cls}(p_i, p_i^*) = -\log[p_i^* p_i + (1 - p_i^*)(1 - p_i)] \qquad (12)$$

The bounding box regression loss of RPN network and Fast RCNN network uses Smooth$_{L_1}$ loss:

$$L_{reg}(t_i, t_i^*) = R(t_i - t_i^*) \qquad (13)$$

Where, R denotes Smooth$_{L_1}$ function,

$$\text{Smooth}_{L_1}(x) = \begin{cases} \dfrac{x^2}{2\sigma^2} & \text{if } |x| < \dfrac{1}{\sigma^2} \\ |x| - 0.5 & \text{otherwise} \end{cases} \qquad (14)$$

The difference between the bounding box regression loss of RPN loss and the regression loss of Fast RCNN loss is that the RPN network is trained when $\sigma = 3$ and the Fast RCNN network is trained when $\sigma = 1$.

4 Experimental results and analysis

4.1 Datasets

This paper evaluates the algorithm performance on the KAIST pedestrian dataset [8], which is composed of 95,328 pairs of Color and Thermal images captured during daytime and nighttime. It is the most widely used multispectral pedestrian detection dataset at present. The dataset is labeled with four categories including person, people, person?, and cyclist. Considering the application areas of multispectral pedestrian detection (e.g., automatic driving), all four categories are treated as positive examples for detection in this paper. To address the problem of the annotation errors and missing annotations in the original annotation of the KAIST dataset, studies [9,28,47] performed data cleaning and re-annotation of the original data. Given that the annotations used in various studies are not consistent, we use 7601 pairs of Color and Thermal images from synthetic annotation (SA) [47] and 8892 pairs of Color and Thermal

images from paired annotation (PA) [28] for model training. The test set consists of 2252 pairs of Color and Thermal images, of which 1455 pairs are from the daytime and 797 pairs are from the nighttime. For a fair comparison with other methods, the test experiments were performed according to the reasonable settings proposed in the literature [8].

4.2 Evaluation indexes

In this paper, Log-average Miss Rate (MR) proposed by [48] is employed as an evaluation index and combined with the plotting of the Miss Rate-FPPI curve to assess the effectiveness of the algorithm. The horizontal coordinate of the Miss Rate-FPPI curve indicates the average number of False Positives Per Image (FPPI), and the vertical coordinate represents the Miss Rate (MR), which is expressed as:

$$\text{MissRate} = \frac{FN}{TP + FN} \qquad (15)$$

$$FPPI = \frac{FP}{\text{Total (images)}} \qquad (16)$$

where FN denotes False Negative, TP denotes True Positive, FP denotes False Positive, the sum of TP and FN is the number of all positive samples, and Total (images) denotes the total number of predicted images. It is worth noting that the lower the Miss Rate-FPPI curve trend, the better the detection performance; the smaller the MR value, the better the detection performance. In order to calculate MR, in logarithmic space, nine points are taken from the horizontal coordinate (limited value range is $[10^{-2}, 10^0]$) of Miss Rate-FPPI curve, and then there are nine corresponding vertical coordinates $m_1, m_2, \ldots m_9$. By averaging these values, MR can be obtained as follows:

$$MR = \exp\left[\frac{1}{n} \sum_{i=1}^{n} \ln(m_i)\right] \qquad (17)$$

where n is 9.

4.3 Implementation details

In this paper, the deep learning framework pytorch1.7 is adopted. The experimental platform is the ubuntu18.04 operating system and a single NVIDIA GeForce RTX 2080Ti GPU. Stochastic Gradient Descent (SGD) algorithm is used to optimize the network during model training, with momentum value of 0.9, weight attenuation value 5×10^{-4}, and initial learning rate is 1×10^{-3}. The model is iterated for five epochs with the batch size of 4, and the learning rate decay to 1×10^{-4} after the 3rd epoch.

4.4 Experimental results and analysis

4.4.1 Construction of the baseline

This work constructs a baseline algorithm architecture based on ResNet-101 backbone network and Faster RCNN detection head. Simple characteristic fusion (feature cascade, element by element addition and element by element multiplication) of the Color and Thermal features output by the backbone network is carried out in

TABLE 1 Experimental results of baseline under different fusion modes.

Fusion modes	All-weather
feature cascade	14.62
element by element multiplication	14.26
element by element addition	**13.84**

The bold values in highlight the optimal results for this column.

TABLE 3 MRs of different modal inputs.

Input	All-weather	Day	Night
dual-stream Color images	25.37	19.31	31.18
dual-stream Thermal images	17.55	22.81	12.61
Color images + Thermal images	**13.84**	**15.35**	**12.48**

The bold values in highlight the optimal results for this column.

TABLE 2 MRs of different methods on KAIST datasets.

Methods	SA			PA(Color)			PA(Thermal)		
	All-weather	Day	Night	All-weather	Day	Night	All-weather	Day	Night
ACF + T + THOG	41.65	39.18	48.29	41.74	39.30	49.52	41.36	38.74	48.30
Halfway Fusion	25.75	24.88	26.59	25.10	24.29	26.12	25.51	25.20	24.90
CMT_CNN	36.83	34.56	41.82	36.25	34.12	41.21	–	–	–
IAF R-CNN	15.73	14.55	18.26	15.65	14.95	18.11	16.00	15.22	17.56
IATDNN + IAMSS	14.95	14.67	15.72	15.14	14.82	15.87	15.08	15.02	15.20
CIAN	14.12	14.77	11.13	14.64	15.13	12.43	14.68	16.21	9.88
CS-RCNN	11.43	11.86	8.82	–	–	–	–	–	–
IT-MN	14.19	14.30	13.98	–	–	–	–	–	–
DCRD	12.58	13.12	11.65	13.64	13.15	13.98	–	–	–
Ours	10.71	13.09	8.45	11.11	12.85	8.77	10.98	13.07	8.53

three sets of experiments. The fused feature is used as the input of the detection head. In order to ensure the high efficiency of the build baseline algorithm, synthesis annotation is employed to train and test the baseline. The test results are shown in Table 1. The MR values using feature cascade, element by element addition and element by element multiplication in the all-weather scene are 14.62%, 13.84% and 14.26%, respectively. By comparing these three results, it can be seen that the feature element by element addition demonstrates the best performance. Therefore, we adopt the method of adding features element by element as the baseline integration method.

4.4.2 Performance comparison of different methods

The performance of this method is compared with several other state-of-the-art methods. The compared methods include hand-represented methods, e.g., ACT + T + THOG [8] and deep learning-based methods, e.g., Halfway Fusion [9], CMT_CNN[49], CIAN[50], IAF R-CNN[51], IATDNN + IAMSS[19], CS-RCNN [52], IT-MN [53], and DCRD [54]. Here, the model is trained using 7601 pairs of Color and Thermal images from SA and 8892 pairs of Color and Thermal images from PA, respectively. Besides, 2252 pairs of Color and Thermal images from the test set are used for model testing. Table 2 lists the experimental results.

Table 2 shows that when the model is trained with SA, the MRs of the method proposed in this paper are 10.71%, 13.09% and 8.45% for all-weather, daytime and nighttime scenes, respectively, which are 0.72%, −1.23% and 0.37% lower than the compared method CS-RCNN with the best performance, respectively. The PA (Color) and PA (Thermal) in Table 2 represent the Color annotation and Thermal

annotation in the pairwise annotation PA, respectively, for the purpose of training the model. It can be seen from two that the MRs of the method in this paper are 11.11% and 10.98% when using Color annotation and Thermal annotation in the all-weather scene, which are 2.53% and 3.70%, respectively, lower than those of compared method with the best performance. In addition, by analyzing the experimental results of two improved versions of annotations, it can be found that pedestrian detection results are different when using different annotations, indicating the importance of annotations.

4.4.3 Analysis of ablation experiments
4.4.3.1 Complementarity and importance of color and thermal features

This section compares the effect of different input sources on pedestrian detection performance. In order to eliminate the impact of the proposed module on detection performance, three sets of experiments are conducted on baseline: 1) the combination of Color and Thermal images as the input source (the input of the two branches of the backbone network are respectively Color and Thermal images); 2) dual-stream Color image as the input source (use Color images to replace Thermal images, that is, the backbone network input source is Color images); 3) dual-stream Thermal images as the input source (use Thermal images to replace Color images, that is, the backbone network input source is Thermal images).The training set of the model here is 7061 pairs of images of SA, and the test set is 2252 pairs of Color and Thermal images. Table 3 shows the MRs of these three input sources for the all-weather, daytime, and nighttime scenes. It can be seen from Table 3 that the MRs obtained using Color and Thermal images as input to the network are
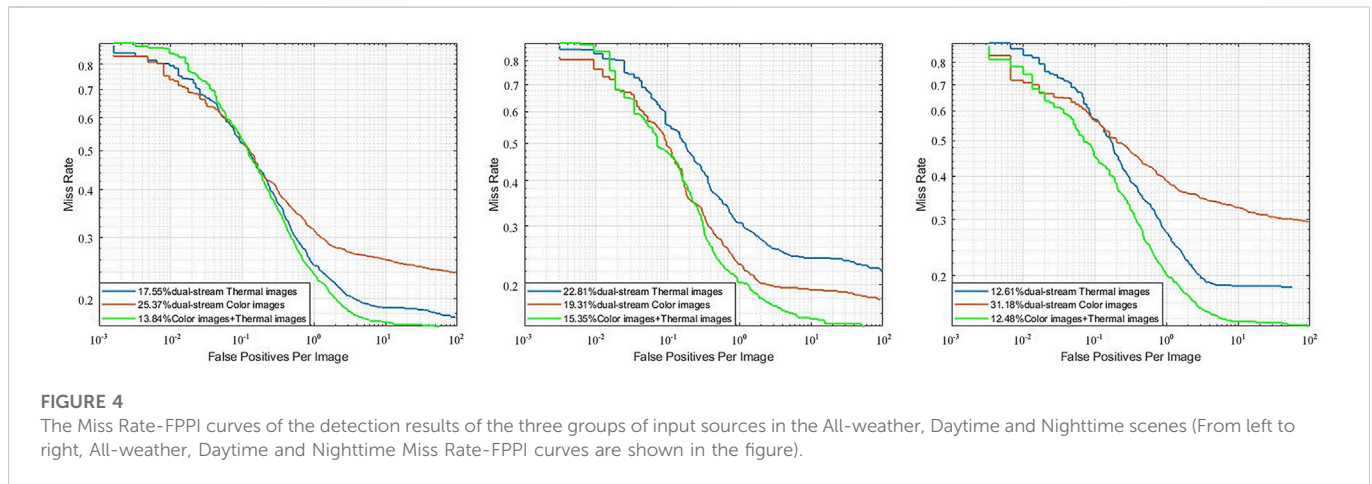
FIGURE 4
The Miss Rate–FPPI curves of the detection results of the three groups of input sources in the All-weather, Daytime and Nighttime scenes (From left to right, All-weather, Daytime and Nighttime Miss Rate-FPPI curves are shown in the figure).

TABLE 4 MRs for ablation studies of the proposed method on SA.

Methods	All-weather	Day	Night
Baseline	13.84	15.35	12.48
baseline + CIEM	11.21	13.15	9.07
baseline + CAFFM	11.68	13.81	9.50
Overall model	**10.71**	**13.09**	**8.45**

The bold values in highlight the optimal results for this column.

13.84%, 15.35% and 12.48% for the all-weather, daytime and nighttime scenes, respectively, which are 11.53%, 3.96%, 18.70% and 3.71%, 7.46%, 0.13% lower than using Color images and Thermal images as input alone. The experimental results prove that the detection network combining Color and Thermal features delivers better performance, indicating that Color and Thermal features are important for pedestrian detection.

Figure 4 shows the Miss Rate-FPPI curves of the detection results for these three input sources in the all-weather, daytime, and nighttime scenes (blue, red and green curves indicate dual-stream Thermal images, dual-stream Color images, and Color and Thermal images, respectively). By analyzing the Miss Rate-FPPI curve trend and combining with the experimental data in Table 3, it can be seen that the detection effect of Color images as the input source is better than that of Thermal images in the daytime scene while the result is the opposite for the night scene, and the detection effect of Color and Thermal images combined as the input source is better than that of single-modal input in both daytime and nighttime. It shows that there are complementary features between Color and Thermal modalities, and the fusion of the two modal features can improve the pedestrian detection performance.

4.4.3.2 Ablation experiments

In this section, ablation experiments are conducted to demonstrate the effectiveness of the proposed cascaded information enhancement module (CIEM) and cross-modal attentional feature fusion module (CAFFM). Here, 7061 pairs of SA images are used to train the model, and 2252 pairs of Color and Thermal images in the test set are used to test the model.

Effectiveness of CIEM: CIEM is used to enhance the pedestrian features in Color and Thermal images to reduce the interference from the background. The experimental results are shown in Table 4. The MRs of baseline on SA are 13.84%, 15.35% and 12.48% for all-weather, daytime and nighttime scenes, respectively. When CIEM is additionally employed, the MRs are 11.21%, 13.15% and 9.07% for all-weather, daytime and nighttime scenes, respectively, which are reduced by 2.63%, 2.20% and 3.41% compared to the baseline, respectively. It is shown that the proposed CIEM effectively enhances the pedestrian features in both modalities, reduces the interference of background, and improves the pedestrian detection performance.

Validity of CAFFM: CAFFM is used to effectively fuse Color and Thermal features. The experimental results are shown in Table 4. On the SA, when the baseline is used with CAFFM, the MRs are 11.68%, 13.81% and 9.50% in all-weather, daytime and nighttime scenes, respectively, which are reduced by 2.16%, 1.54% and 2.98% compared baseline, respectively. It shows that the proposed CAFFM effectively fuses the two modal features to achieve robust multispectral pedestrian detection.

Overall effectiveness: The proposed CIEM and CAFFM are additionally used on the basis of baseline. Experimental results show a reduction of 3.13%, 2.26% and 4.03% in MRs for all-weather, daytime and nighttime scenes, respectively, compared to the baseline, indicating the overall effectiveness of the proposed method. A closer look reveals that with additional employment of CIEM and CAFFM alone, MRs are decreased by 2.63% and 2.16%, respectively, in the all-weather scene, but the MR of the overall model is reduced by 3.13%. It demonstrates that there is some orthogonal complementarity in the role of the proposed two modules.

Figure 5 shows the Miss Rate-FPPI curves for CIEM and CAFFM ablation studies in all-weather, daytime and nighttime scenes (blue, red, orange and green curves represent baseline, baseline + CIEM, baseline + CAFFM and overall model, respectively). It is clear that the curve trends of each module and the overall model are both lower than that of the baseline, which further proves the effectiveness of the method presented in this work.

Furthermore, in order to qualitatively analyze the effectiveness of the proposed CIEM and CAFFM, four pairs of Color and Thermal images (two pairs of images are taken from daytime and two pairs of images are taken from nighttime) are selected from the test set for testing. The pedestrian detection results of the baseline and each proposed module are shown in Figure 6. The first row is the
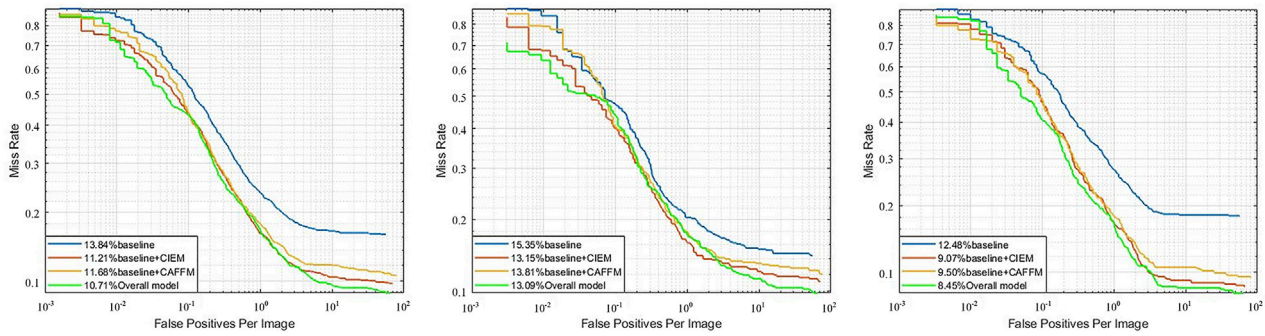
FIGURE 5
The Miss Rate-FPPI curves of CIEM and CAFFM ablation studies in All-weather, Daytime and Nighttime scenes (From left to right, All-weather, Daytime and Nighttime Miss Rate-FPPI curves are shown in the figure).
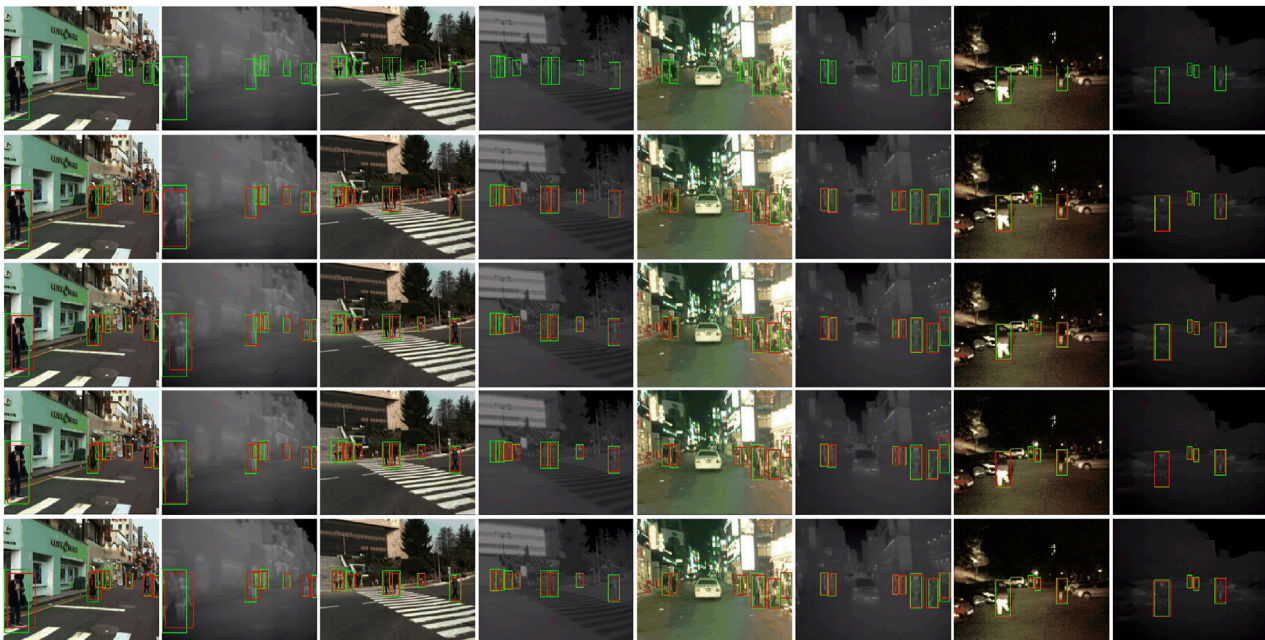

FIGURE 6
In this paper, each module and baseline pedestrian detection results (The first row is the visualization results of labeled boxes for Color and Thermal images, and the second to the fifth rows are the visualization results of the labeled and prediction boxes for baseline, baseline + CIEM, baseline + CAFFM and the overall model pedestrian detection with the green and red boxes representing the labeled and prediction boxes, respectively).

visualization results of labeled boxes for Color and Thermal images, and the second to the fifth rows are the visualization results of the labeled and prediction boxes for baseline, baseline + CIEM, baseline + CAFFM, and the overall model pedestrian detection with the green and red boxes representing the labeled and prediction boxes, respectively. It can be seen that the proposed method successfully addresses the problem of pedestrian missing detection in complex environments and achieves more accurate detection boxes. For example, the second row, pedestrian detection missing happens in the first, third, and fourth pairs of images in the baseline detection result, however, the pedestrian miss detection problem is properly solved with CIEM and CAFFM added to the baseline and the overall model produces more accurate pedestrian detection boxes.

5 Conclusion

In this paper, we propose a multispectral pedestrian detection algorithm including cascaded information enhancement module and cross-modal attention feature fusion module. The proposed method improves the accuracy of pedestrian detection in multispectral images (Color and Thermal images) by effectively fusing the features from the two modules and augmenting the pedestrian features. Specifically, on the one hand, a cascaded information enhancement module (CIEM) is designed to enhance single-modal features to enrich the pedestrian features and suppress interference from the background noise. On the other hand, unlike previous methods that simply splice Color and Thermal features directly, a cross-modal attention feature fusion

module (CAFFM) is introduced to mine the features of both Color and Thermal modalities and to complement each other, then complementary enhanced modal features are used to construct global features. Extensive experiments have been conducted on two improved annotations of the public dataset KAIST. The experimental results show that the proposed method is conducive to obtain more comprehensive pedestrian features and improve the accuracy of multispectral image pedestrian detection.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://gitcode.net/mirrors/soonminhwang/rgbt-ped-detection?utm_source=csdn_github_accelerator.

Author contributions

YY responsible for scheme design, experiment and writing of the paper. KX guide the scheme design and experiment of the paper. KW guide experimental data analysis, paper writing and modification.

Funding

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Jeong M, Ko BC, Nam J-Y. Early detection of sudden pedestrian crossing for safe driving during summer nights. *IEEE Trans Circuits Syst Video Technol* (2017) 27:1368–80. doi:10.1109/TCSVT.2016.2539684

2. Zhang S, Cheng D, Gong Y, Shi D, Qiu X, Xia Y, et al. Pedestrian search in surveillance videos by learning discriminative deep features. *Neurocomputing* (2018) 283:120–8. doi:10.1016/j.neucom.2017.12.042

3. Li L, Xie M, Li F, Zhang Y, Li H, Tan T. Unsupervised domain adaptive person re-identification guided by low-rank priori. *J Chongqing Univ* (2021) 44:57–70. doi:10.11835/j.issn.1000-582X.2021.11.008

4. Li H, Chen Y, Tao D, Yu Z, Qi G. Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification. *IEEE Trans Inf Forensics Security* (2021) 16:1480–94. doi:10.1109/TIFS.2020.3036800

5. Li H, Dong N, Yu Z, Tao D, Qi G. Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification. *IEEE Trans Circuits Syst Video Technol* (2022) 32:2814–30. doi:10.1109/TCSVT.2021.3099943

6. Li S, Li F, Wang K, Qi G, Li H. Mutual prediction learning and mixed viewpoints for unsupervised-domain adaptation person re-identification on blockchain. *Simulation Model Pract Theor* (2022) 119:102568. doi:10.1016/j.simpat.2022.102568

7. Wang S, Liu R, Li H, Qi G, Yu Z. Occluded person re-identification via defending against attacks from obstacles. *IEEE Trans Inf Forensics Security* (2023) 18:147–61. doi:10.1109/TIFS.2022.3218449

8. Hwang S, Park J, Kim N, Choi Y, Kweon IS. Multispectral pedestrian detection: Benchmark dataset and baseline. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 07-12 June 2015; Boston, MA, USA (2015). p. 1037–45. doi:10.1109/CVPR.2015.7298706

9. Liu J, Zhang S, Wang S, Metaxas DN. Multispectral deep neural networks for pedestrian detection. In: Proceedings of the British Machine Vision Conference 2016; 19-22 September 2016; York, UK (2016).

10. González A, Fang Z, Socarras Y, Serrat J, Vázquez D, Xu J, et al. Pedestrian detection at day/night time with visible and fir cameras: A comparison. *Sensors* (2016) 16:820. doi:10.3390/s16060820

11. Zhang Y, Yang M, Li N, Yu Z. Analysis-synthesis dictionary pair learning and patch saliency measure for image fusion. *Signal Process.* (2020) 167:107327. doi:10.1016/j.sigpro.2019.107327

12. Liu Y, Wang L, Cheng J, Li C, Chen X. Multi-focus image fusion: A survey of the state of the art. *Inf Fusion* (2020) 64:71–91. doi:10.1016/j.inffus.2020.06.013

13. Li H, He X, Tao D, Tang Y, Wang R. Joint medical image fusion, denoising and enhancement via discriminative low-rank sparse dictionaries learning. *Pattern Recognition* (2018) 79:130–46. doi:10.1016/j.patcog.2018.02.005

14. Li H, Wang Y, Yang Z, Wang R, Li X, Tao D. Discriminative dictionary learning-based multiple component decomposition for detail-preserving noisy image fusion. *IEEE Trans Instrumentation Meas* (2020) 69:1082–102. doi:10.1109/tim.2019.2912239

15. Xie M, Wang J, Zhang Y. A unified framework for damaged image fusion and completion based on low-rank and sparse decomposition. *Signal Processing: Image Commun* (2021) 29:116400. doi:10.1016/j.image.2021.116400

16. Wang S, Huang B, Li H, Qi G, Tao D, Yu Z. Key point-aware occlusion suppression and semantic alignment for occluded person re-identification. *Inf Sci* (2022) 606:669–87. doi:10.1016/j.ins.2022.05.077

17. Zhu Z, Luo Y, Chen S, Qi G, Mazur N, Zhong C, et al. Camera style transformation with preserved self-similarity and domain-dissimilarity in unsupervised person re-identification. *J Vis Commun Image Representation* (2021) 80:103303. doi:10.1016/j.jvcir.2021.103303

18. Yang X, Qian Y, Zhu H, Wang C, Yang M. Baanet: Learning bi-directional adaptive attention gates for multispectral pedestrian detection. In: 2022 International Conference on Robotics and Automation (ICRA); 23-27 May 2022; Philadelphia, PA, USA (2022). p. 2920–6. doi:10.1109/ICRA46639.2022.9811999

19. Guan D, Cao Y, Yang J, Cao Y, Yang MY. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Inf Fusion* (2019) 50:148–57. doi:10.1016/j.inffus.2018.11.017

20. Zhou K, Chen L, Cao X. Improving multispectral pedestrian detection by addressing modality imbalance problems. In: *European conference on computer vision*. Berlin, Germany: Springer (2020). p. 787–803.

21. Li Q, Zhang C, Hu Q, Fu H, Zhu P. Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection. *IEEE Trans Multimedia* (2022) 1. doi:10.1109/tmm.2022.3160589

22. Zhang H, Fromont E, Lefevre S, Avignon B. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In: 2020 IEEE International Conference on Image Processing (ICIP); 25-28 October 2020; Abu Dhabi, United Arab Emirates (2020). p. 276–80. doi:10.1109/ICIP40778.2020.9191080

23. Dollár P, Appel R, Belongie S, Perona P. Fast feature pyramids for object detection. *IEEE Trans Pattern Anal Machine Intelligence* (2014) 36:1532–45. doi:10.1109/TPAMI.2014.2300479

24. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); 20-25 June 2005; San Diego, CA, USA (2005). p. 886–93. doi:10.1109/CVPR.2005.177

25. König D, Adam M, Jarvers C, Layher G, Neumann H, Teutsch M. Fully convolutional region proposal networks for multispectral person detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 21-26 July 2017; Honolulu, HI, USA (2017). p. 243–50. doi:10.1109/CVPRW.2017.36

26. Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Machine Intelligence* (2017) 39:1137–49. doi:10.1109/TPAMI.2016.2577031

27. Kim J, Park I, Kim S. A fusion framework for multi-spectral pedestrian detection using efficientdet. In: 2021 21st International Conference on Control, Automation and Systems (ICCAS); 12-15 October 2021; Jeju, Korea (2021). p. 1111–3. doi:10.23919/ICCAS52745.2021.9650057

28. Zhang L, Zhu X, Chen X, Yang X, Lei Z, Liu Z. Weakly aligned cross-modal learning for multispectral pedestrian detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 27 October 2019 - 02 November 2019; Seoul, Korea (South) (2019). p. 5126–36. doi:10.1109/ICCV.2019.00523

29. Kim J, Kim H, Kim T, Kim N, Choi Y. Mlpd: Multi-label pedestrian detector in multispectral domain. IEEE Robotics Automation Lett (2021) 6:7846–53. doi:10.1109/LRA.2021.3099870

30. Kim JU, Park S, Ro YM. Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. IEEE Trans Circuits Syst Video Technol (2022) 32:1510–23. doi:10.1109/TCSVT.2021.3076466

31. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach California USA. Curran Associates Inc. (2017).

32. Li S, Zou C, Li Y, Zhao X, Gao Y. Attention-based multi-modal fusion network for semantic scene completion. Proc AAAI Conf Artif Intelligence (2020) 34:11402–9. doi:10.1609/aaai.v34i07.6803

33. Li B, Zhou Y, Ren H. Image emotion caption based on visual attention mechanisms. In: 2020 IEEE 6th International Conference on Computer and Communications (ICCC); 11-14 December 2020; Chengdu, China. IEEE (2020). p. 1456–60.

34. Xiao W, Zhang Y, Wang H, Li F, Jin H. Heterogeneous knowledge distillation for simultaneous infrared-visible image fusion and super-resolution. IEEE Trans Instrumentation Meas (2022) 71:1–15. doi:10.1109/tim.2022.3149101

35. Li H, Cen Y, Liu Y, Chen X, Yu Z. Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion. IEEE Trans Image Process (2021) 30:4070–83. doi:10.1109/tip.2021.3069339

36. Li H, Gao J, Zhang Y, Xie M, Yu Z. Haze transfer and feature aggregation network for real-world single image dehazing. Knowledge-Based Syst (2022) 251:109309. doi:10.1016/j.knosys.2022.109309

37. Xu M, Fu P, Liu B, Li J. Multi-stream attention-aware graph convolution network for video salient object detection. IEEE Trans Image Process (2021) 30:4183–97. doi:10.1109/TIP.2021.3070200

38. Li H, Xu K, Li J, Yu Z. Dual-stream reciprocal disentanglement learning for domain adaptation person re-identification. Knowledge-Based Syst (2022) 251:109315. doi:10.1016/j.knosys.2022.109315

39. Zhang Y, Wang Y, Li H, Li S. Cross-compatible embedding and semantic consistent feature construction for sketch re-identification. In: MM '22. Proceedings of the 30th ACM International Conference on Multimedia; 10 October 2022; New York, NY, USA. Association for Computing Machinery (2022). p. 3347–55. doi:10.1145/3503161.3548224

40. Wang Y, Qi G, Li S, Chai Y, Li H. Body part-level domain alignment for domain-adaptive person re-identification with transformer framework. IEEE Trans Inf Forensics Security (2022) 17:3321–34. doi:10.1109/TIFS.2022.3207893

41. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-excitation networks. IEEE Trans Pattern Anal Machine Intelligence (2020) 42:2011–23. doi:10.1109/TPAMI.2019.2913372

42. Li X, Wang W, Hu X, Yang J. Selective kernel networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 15-20 June 2019; Long Beach, CA, USA (2019). p. 510–9. doi:10.1109/CVPR.2019.00060

43. Dai Y, Gieseke F, Oehmcke S, Wu Y, Barnard K. Attentional feature fusion. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV); 03-08 January 2021; Waikoloa, HI, USA (2021). p. 3559–68. doi:10.1109/WACV48630.2021.00360

44. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 27-30 June 2016; NV, USA (2016). p. 770–8. doi:10.1109/CVPR.2016.90

45. Woo S, Park J, Lee J-Y, Kweon IS. Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV) (2018). p. 3–19.

46. Girshick R. Fast r-cnn. In: 2015 IEEE International Conference on Computer Vision (ICCV); 07-13 December 2015; Santiago, Chile (2015). p. 1440–8. doi:10.1109/ICCV.2015.169

47. Li C, Song D, Tong R, Tang M (2018). Multispectral pedestrian detection via simultaneous detection and segmentation. arXiv preprint arXiv:1808.04818

48. Dollar P, Wojek C, Schiele B, Perona P. Pedestrian detection: An evaluation of the state of the art. IEEE Trans Pattern Anal Machine Intelligence (2012) 34:743–61. doi:10.1109/TPAMI.2011.155

49. Xu D, Ouyang W, Ricci E, Wang X, Sebe N. Learning cross-modal deep representations for robust pedestrian detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 26 Jul 2017; Hawaii (2017). p. 4236–44. doi:10.1109/CVPR.2017.451

50. Zhang L, Liu Z, Zhang S, Yang X, Qiao H, Huang K, et al. Cross-modality interactive attention network for multispectral pedestrian detection. Inf Fusion (2019) 50:20–9. doi:10.1016/j.inffus.2018.09.015

51. Li C, Song D, Tong R, Tang M. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. Pattern Recognition (2019) 85:161–71. doi:10.1016/j.patcog.2018.08.005

52. Zhang Y, Yin Z, Nie L, Huang S. Attention based multi-layer fusion of multispectral images for pedestrian detection. IEEE Access (2020) 8:165071–84. doi:10.1109/ACCESS.2020.3022623

53. Zhuang Y, Pu Z, Hu J, Wang Y. Illumination and temperature-aware multispectral networks for edge-computing-enabled pedestrian detection. IEEE Trans Netw Sci Eng (2022) 9:1282–95. doi:10.1109/TNSE.2021.3139335

54. Liu T, Lam K-M, Zhao R, Qiu G. Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection. IEEE Trans Circuits Syst Video Technol (2022) 32:315–29. doi:10.1109/TCSVT.2021.3060162