Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.048200

ARTICLE

Lightweight Cross-Modal Multispectral Pedestrian Detection Based on Spatial Reweighted Attention Mechanism

Lujuan Deng, Ruochong Fu*, Zuhe Li, Boyi Liu, Mengze Xue and Yuhao Cui

School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, 450002, China

*Corresponding Author: Ruochong Fu. Email: 332107040635@email.zzuli.edu.cn

Received: 30 November 2023 Accepted: 02 February 2024 Published: 26 March 2024

ABSTRACT

Multispectral pedestrian detection technology leverages infrared images to provide reliable information for visible light images, demonstrating significant advantages in low-light conditions and background occlusion scenarios. However, while continuously improving cross-modal feature extraction and fusion, ensuring the model's detection speed is also a challenging issue. We have devised a deep learning network model for cross-modal pedestrian detection based on Resnet50, aiming to focus on more reliable features and enhance the model's detection efficiency. This model employs a spatial attention mechanism to reweight the input visible light and infrared image data, enhancing the model's focus on different spatial positions and sharing the weighted feature data across different modalities, thereby reducing the interference of multi-modal features. Subsequently, lightweight modules with depthwise separable convolution are incorporated to reduce the model's parameter count and computational load through channel-wise and point-wise convolutions. The network model algorithm proposed in this paper was experimentally validated on the publicly available KAIST dataset and compared with other existing methods. The experimental results demonstrate that our approach achieves favorable performance in various complex environments, affirming the effectiveness of the multispectral pedestrian detection technology proposed in this paper.

KEYWORDS

Multispectral pedestrian detection; convolutional neural networks; depth separable convolution; spatially reweighted attention mechanism

1 Introduction

Multispectral pedestrian detection is a technology that utilizes multispectral image data for pedestrian detection [1–3]. Its development stemmed from the need to address the performance degradation of traditional pedestrian detection methods in poor lighting and night scenes, leading to the incorporation of multispectral information to enhance detection accuracy [4,5]. However, when relying solely on one type of visual sensor, establishing an efficient and accurate pedestrian detection system that operates effectively under various weather conditions remains challenging. Therefore, the fusion of visible light and infrared image-based pedestrian detection technology is gradually becoming one of the key technologies for applications [6] that require round-the-clock operation and monitoring, such

as autonomous driving and security defense [7,8]. Visible light cameras can generate high-resolution images with rich color and texture information, making them the primary component in current pedestrian detection systems. However, in nighttime or low-light environments, relying solely on visible light images for pedestrian detection may lead to issues such as false positives or false negatives, posing potential safety risks. With the continuous improvement of infrared detection methods and the reduction in hardware costs, infrared cameras have gained widespread popularity in applications such as image capture and intelligent monitoring systems [9–11]. This is attributed to the characteristic of infrared cameras being insensitive to changes in lighting conditions while highly sensitive to temperature changes, making them particularly suitable for pedestrian detection tasks in nighttime traffic scenarios [12]. Integrating infrared cameras into practical applications allows for pedestrian detection in different time periods and scenarios, addressing the instability of feature extraction in low-light conditions that visible light cameras may exhibit [13]. This holds significant importance in establishing all-weather, multi-scenario intelligent monitoring systems. However, compared to visible light images, infrared images typically have lower resolution and lack fine details such as texture and color. In conditions with relatively good lighting, the visual information provided by infrared images is relatively limited. Therefore, by simultaneously utilizing multispectral images acquired from both visible light and infrared sensors as input, it is possible to fully leverage their complementarity in pedestrian detection [14]. This approach enables better differentiation between pedestrian targets and backgrounds, thereby achieving more precise pedestrian target localization. This approach addresses the limitations of relying solely on visible light images to handle challenges such as lighting variations and adverse weather conditions, while also overcoming issues associated with using thermal infrared images that may be affected by other heat sources or light interference. Therefore, the fusion of visible light and infrared image information in cross-modal pedestrian detection systems enhances robustness and accuracy, thereby elevating the overall performance of pedestrian detection.

Traditional pedestrian detection methods primarily rely on manually designed features for implementation. However, these methods are susceptible to external factors such as lighting, viewing angles, and occlusions when dealing with visible light images [15]. Currently, most pedestrian detection research focuses on the use of visible light visual sensors, which acquire images by capturing and synthesizing data from red, green, and blue channels. However, a single sensor is prone to losing its effective capture capability of pedestrian targets in adverse weather conditions such as poor lighting, smoke, or dust, which can significantly impact the real-time detection performance. To overcome these challenges, pedestrian detection research has introduced infrared visual sensors to enhance the detection technology [16]. Infrared visual sensors, based on the thermal imaging principle, can effectively distinguish between pedestrian targets and backgrounds and exhibit good performance even in adverse environments. Therefore, when visible light cameras face limitations in monitoring and safety in specific scenarios, choosing to use infrared visual sensors to capture infrared images has become a solution [17]. Although using visible light images and infrared images as input sources in multispectral pedestrian detection technology can effectively counteract the impact of environmental conditions on applications like autonomous driving and intelligent video surveillance, our comparative analysis of pedestrian image information from different modalities in complex environments revealed that the actual multispectral data results in excessive computational requirements during the fusion process due to modality differences [18]. Furthermore, a significant increase in parameter calculations during model detection can result in longer detection times, which can have a substantial impact in practical applications.

Currently, there is a growing demand for multi-modal pedestrian recognition and detection models in various applications. These models need to maintain high accuracy while ensuring that detection

speed does not decrease or even improves. Therefore, when constructing network models, we must prioritize both accuracy and detection speed to strike a balance. Our contributions are summarized as follows:

1. In response to the modality imbalance issue in multispectral pedestrian detection, we propose a network model that leverages Resnet50, lightweight modules, and attention mechanisms to address and improve the modality imbalance problem caused by modal inconsistencies.
2. During the feature extraction process, a spatial reweighted attention module is introduced to enhance the model's focus on different regions, reinforcing the critical features of the two modalities and enabling more efficient learning by the model.
3. After incorporating the attention mechanism, the number of parameters and computational load of the model grow. To address this, we introduce depth-wise separable convolutions, utilizing channel-wise and point-wise convolutions to reduce the computational load while maintaining strong feature representation capabilities.

The remaining sections of this paper are organized as follows. In Section 2, we introduce the relevant work in the field of cross-modal pedestrian detection. Section 3 provides a detailed overview of the overall structure of our proposed model and the key concepts of its components. Section 4 covers the introduction of the dataset, evaluation criteria, and experimental results. Section 5 summarizes our research and discusses its potential significance.

2 Related Works

A significant portion of research in multispectral pedestrian detection builds upon and adapts object detection as a foundation, as nighttime lighting conditions with low illumination make pedestrians less visible and challenging to detect [19]. However, due to the inherent temperature characteristics of pedestrians and the sensitivity to temperature in thermal infrared cameras, researchers have increasingly focused on the use of infrared cameras in conjunction with visible light cameras to acquire images for pedestrian detection. Lai et al. [20] developed a cross-modal image acquisition system and designed a set of imaging hardware, including a color camera, a thermal imager, a beam splitter, and a three-axis camera fixture, to capture visible light and infrared images. They achieved precise optical registration to obtain pairs of registered visible light and infrared images, creating the largest publicly available multispectral dataset to date, known as the KAIST dataset. Subsequent researchers mostly enhanced the input by adding infrared images to the visible light images from this dataset [21,22], achieving more accurate all-weather pedestrian detection through the fusion of complementary information in cross-modal images [23,24]. Based on the focus of cross-modal pedestrian detection research, it can be divided into two categories: 1) studying feature extraction strategies for visible light and infrared image information. 2) studying strategies for processing cross-modal data of visible light and infrared images, and enhancing the speed of feature fusion algorithms.

(1) Cross-Modal Information Feature Extraction

Lee et al. [25] designed a cross-modal bidirectional adaptive attention gating fusion module based on attention mechanisms, which was used to extract information features from different modalities and gradually recalibrate these feature representations. To enhance the characteristics of the two modalities and guarantee the modality's specificity during feature propagation, researchers used a multi-stage, bidirectional fusion technique. Additionally, this model further enhances robustness to lighting changes through adaptive interaction. Yang et al. [26] introduced a mixed attention mechanism that combines local and global key information to enhance the quality of feature fusion between visible

and infrared images, thereby improving the fine representation ability of multi-modal features. In order to achieve a balance between accuracy and efficiency in multispectral pedestrian detection models, a lightweight anchor-free detection pipeline was employed, reducing computational complexity and thus improving model detection speed. In order to balance accuracy and efficiency in the multispectral pedestrian detection model, a lightweight anchor-free detection pipeline was employed, reducing computational complexity and thus improving model detection speed. Zuo et al. [27] employed YOLOv4 as the framework, using two identical feature extraction backbone networks to simultaneously extract relevant features from visible light and infrared photos. In terms of fusion methods, they proposed a multispectral channel feature fusion to achieve information fusion between the two modalities and perform predictions. This method achieved satisfactory detection results. However, due to its fusion method focusing solely on weighted fusion in the channel direction and neglecting spatial regions of interest, there is still room for improving detection performance. Cao et al. [28] introduced a method utilizing a multi-layer fusion network, incorporating spatial attention modules and channel attention modules in the middle section of the network to address the relationships between different modalities. The role of the channel attention module is to dynamically adjust feature weights through a self-supervised manner. Zhang et al. [29] proposed a feature enhancement fusion module based on attention mechanisms to block out ambient noise in various environments. This module can enhance the semantic and positional information of input images, effectively reducing errors in object detection. This provides strong data support for object detection algorithms.

While the above method can improve the model's detection accuracy, it is challenging to accurately determine the weight ratio between different modalities when fusing their features. Because the importance of modalities may vary depending on the task, this could result in the neglect of valuable information from certain modalities. Therefore, there is still room for improvement in the process of extracting features from two modalities. Considering the contributions and importance of each modality, it is necessary to explore alternative techniques for better cross-modal feature fusion. Through such improvements, useful information between different modalities can be better utilized before modal fusion, further optimizing the model's performance.

(2) Cross-Modal Data Processing

The process of cross-modal data processing aimed at effectively integrating and handling visible light and infrared image data, as depicted in Fig. 1. Jiang et al. [30] introduced two sub-networks: one for fusing transformer histograms of daytime images, and another for fusing transformer histograms of nighttime images. These fused features are subsequently processed through a feature fusion module to facilitate their combination and interaction. Additionally, by merging the processing streams of these two modalities into a single stream and utilizing the self-attention mechanism of the Transformer model, the fusion of information both within and between modalities is achieved. This not only simplifies computational requirements but also enhances the overall efficiency of the model. In order to address the issue of positional misalignment between visible light and infrared images, Zang et al. [31] perceived the cross-modal image's positional offset and adaptively align regional features for position calibration. To effectively fuse visible light and infrared features, a multi-modal weighting module is introduced to focus on more reliable features and suppress irrelevant ones. Simultaneously, researchers devised an RoI jitter strategy to enhance the scalability of cross-modal pedestrian detection algorithms across various perception devices and detection system configurations. To achieve the fusion of information between different multispectral network hosts, Ryu et al. [32] proposed a method called the multispectral interactive convolutional neural network. Researchers enforced interaction by adjusting the weights of feature mappings between multiple spectral networks. The network does not require

additional parameter learning, modifications to the network structure, or increased computational burden, thereby enhancing model efficiency while improving dual-stream feature fusion.

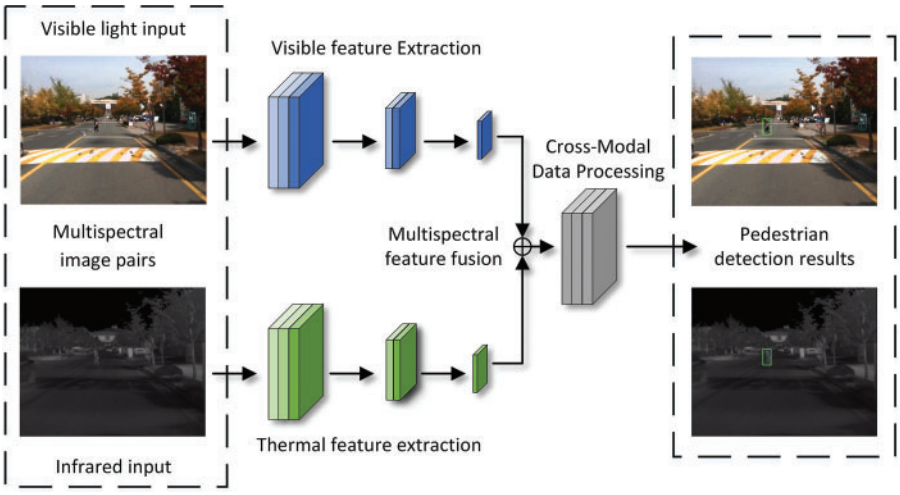

Figure 1: Cross-modal data detection pipelines

Despite continuous efforts by researchers to enhance model detection accuracy, this often comes at the cost of sacrificing some portion of the model's detection speed. In practical applications, real-time performance is crucial, and existing methods often fall short of meeting real-time requirements. Therefore, it is of utmost importance to improve the model's detection speed while ensuring detection accuracy. In response to the aforementioned challenges, we propose a modality balancing framework for visible-infrared cross-modal applications. This framework addresses how to effectively leverage cross-modal information when dealing with modality inconsistencies. Additionally, spatial attention mechanisms and lightweight modules are incorporated into the network architecture. When predicting relevant feature regions, the model automatically captures crucial region features for various deformable data, enhancing cross-modal key regions while reducing model parameters and computational load. This balances the overall detection performance of the model.

3 Method

3.1 Modality Balancing Framework for Cross-Modal

The network architecture is shown in Fig. 2. The model extends the Single Shot MultiBox Detector (SSD) framework and is primarily composed of a feature extraction module and a feature fusion module. Initially, the network captures the illumination values of both infrared and visible light images simultaneously using a miniature neural network. However, infrared images may struggle to reflect environmental lighting conditions during daytime. To reduce computational complexity, the visible light and infrared images are scaled to 56×56 and fed into the lighting perception module, which consists of two lightweight convolutional layers and three fully connected layers. Subsequently, taking advantage of the characteristics of lightweight convolution, feature compression and extraction are performed using ReLU activation functions and 2×2 max-pooling layers. Images are processed through convolutional layers equipped with batch normalization and ReLU activation functions, enhancing the network's non-linear representation capability. Subsequently, lightweight convolutional layers and pooling layers are used to reduce the network's parameter count and computational complexity. This step extracts the primary features from the feature maps. Then, a residual connection

adds the original input tensor to the feature tensor obtained through the convolutional block, resulting in a new feature representation.

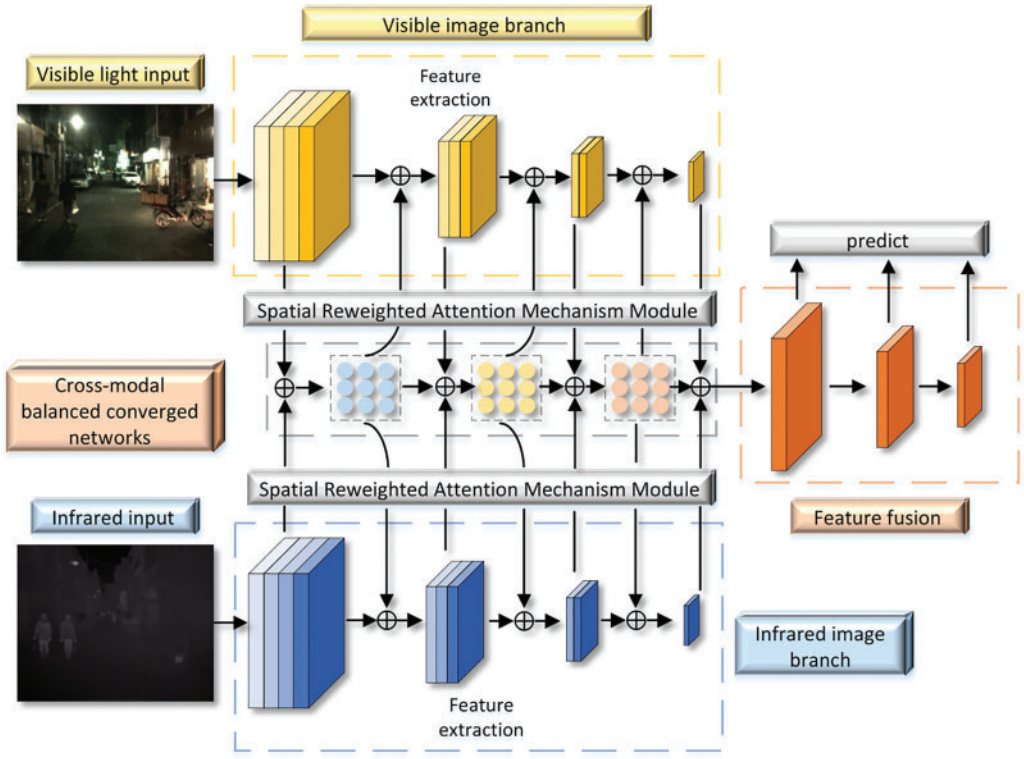

Figure 2: Overview framework of modal balance network

The main idea of this approach is to input the features of both modalities into two separate weight factor learning networks, thereby generating a weight factor for each modality. These weight factors are then multiplied with the features to enhance the feature representation of the other modality. Through this approach, the network is capable of efficiently transferring information from each modality to the other, resulting in a more enriched and informative output feature. The network architecture of the model is designed to adapt to machine memory constraints and computational affordability as much as possible. The structure of the network model is shown in Table 1.

Table 1: Network structure table

Network hierarchy	Network layer	Output	Parameter settings
Layer1	Convolution1	56 × 56	7 × 7, 64, stride 2, padding 3
Layer2	MaxPool1	28 × 28	3 × 3, stride 2
Layer3	Convolution2 × 3	14 × 14	1 × 1, [64, 64, 256], stride 1 padding same
Layer4	BatchNormalization1 Activation1	14 × 14	relu

(Continued)

Table 1 (continued)

Network hierarchy	Network layer	Output	Parameter settings
Layer5	Convolution3 × 3	14 × 14	1 × 1, [64, 64, 256], stride 1 padding same
Layer6	BatchNormalization Activation	14 × 14	relu
Layer7	Convolution4 × 3	14 × 14	1 × 1, [128, 128, 512], stride 2 padding same
Layer8	Activation1	14 × 14	relu
Layer9	Convolution5	14 × 14	1 × 1, 512, stride 1
Layer10	L2Normalization	14 × 14	gamma_init = 10
Layer11	Convolution6	7 × 7	7 × 7, stride 1
Layer12	Fully connect	1 × 1	

3.2 Spatial Reweighted Attention Mechanism Module

The purpose of spatial domain attention is to enhance key regional features in an image. It accomplishes this by effectively reorganizing spatial information from the original image using a spatial transformation module, mapping it to another space while preserving essential details. To mitigate the issue of multi-modal feature interference during the fusion of visible and infrared features, we designed an effective Spatial Reweighted Attention Module (SRAM). This is shown in Fig. 3. It allows for a focus on significant spatial feature regions during the feature fusion process, thus reducing interference between different modalities. We added the spatial reweighted attention module for feature extraction and fusion. By introducing this attention mechanism between different spatial locations in multispectral data, focusing on different regions and assigning them varying importance weights, we can more accurately capture subtle image features and spatial relationships. The extraction and fusion of such feature information contribute to improving the model's recognition and classification capabilities in complex environments. The input features are divided into visible features and infrared features. During the feature extraction process, the SRAM module is introduced. Each modality obtains feature information extracted by the SRAM module and adds them together as an intermediate feature map. The intermediate feature map is then subjected to the attention mechanism once more, and this information is shared between the two modalities. This allows both modalities to continuously supplement their own information with features from the other modality during the feature extraction process, ultimately resulting in enhanced feature information.

This module calculates weights for each position in the image and applies these weights to the original features, thereby reinforcing the representation of specific target regions of interest and reducing the influence of background areas unrelated to the current task. This approach effectively performs adaptive adjustments on the spatial level of the image, concentrating the focus to better emphasize task-relevant information. During the feature extraction stage, a spatial attention module is integrated to emphasize the more critical feature regions within the input image. This module will perform element-wise addition with features extracted from another image to compensate for feature information differences between the two different modalities, enhancing the final fused feature representation. Subsequently, we apply global average pooling and global max pooling operations to the features, obtaining two sets of statistical features representing different information for each

channel. Subsequently, these two sets of statistical features are merged using a convolutional kernel with a broad receptive field to achieve feature fusion. Finally, weight maps are generated through an activation function and superimposed onto the original input feature map, enhancing the target region.

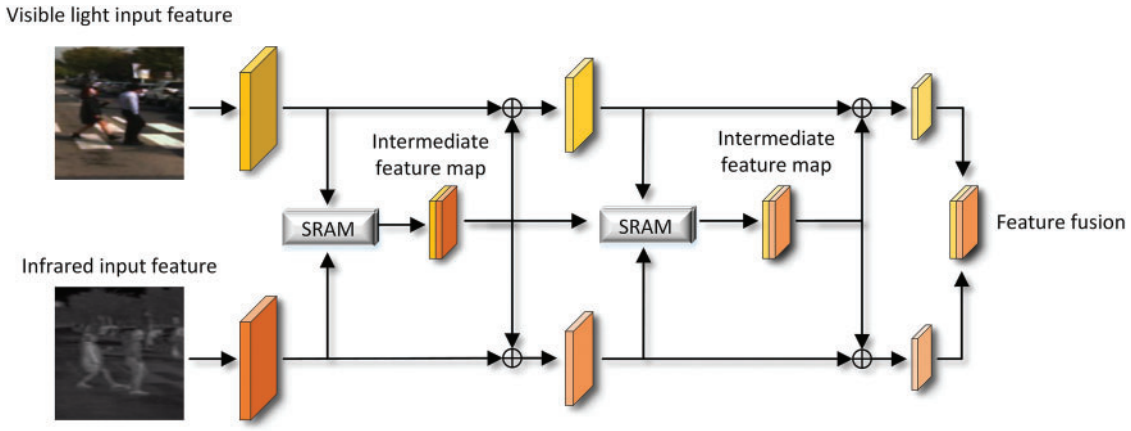

Figure 3: Spatial reweighted attention mechanism pipelines

As shown in Fig. 4, during the feature extraction process, features first undergo average pooling $F^s_{avg} \in \mathrm{R}^{1*H*W}$ and max pooling $F^s_{\max} \in \mathrm{R}^{1*H*W}$ operations along the channel axis, and then they are concatenated. A convolution operation is employed to generate a two-dimensional spatial attention map, which is then added to the original input features. After passing through the sigmoid function, the output feature G is obtained, and finally, feature emphasis or suppression is achieved through encoding. As shown in Eqs. (1)–(3), $F \in \mathrm{R}^{C*H*W}$ represents the input features, $M_s \in \mathrm{R}^{1*H*W}$ is the two-dimensional convolution of the spatial reweighted attention module, and σ is the sigmoid function.

$$\mathrm{M}_s(F) = \left(f^{7\times7}([AvgPool(F); MaxPool(F)]\right) \tag{1}$$

$$\mathrm{M}_s(F) = \left(f^{7\times7}\left(\left[F^s_{avg}; F^s_{\max}\right]\right)\right) \tag{2}$$

$$G = \sigma\left(F + M_s(F)\right) \tag{3}$$

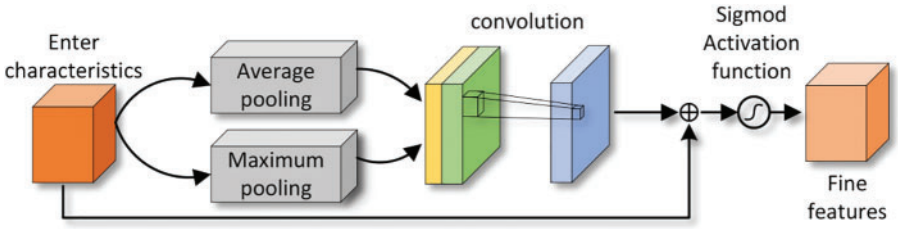

Figure 4: Spatial reweighted attention mechanism module

As shown in Eq. (4), v represents the visible light image, l represents the infrared image, and $v\left(C_x, C_y\right)$ and $l\left(C_x, C_y\right)$ represent their corresponding position features, with $\left(C_x, C_y\right)$ being the coordinates in the image. Fusion of visible light and infrared information is achieved through attention weights, $F\left(C_x, C_y\right)$ represents the fused features, and $\alpha\left(C_x, C_y\right)$ and $\beta\left(C_x, C_y\right)$ are the attention weights.

$$F\left(C_x, C_y\right) = \alpha\left(C_x, C_y\right) * v\left(C_x, C_y\right) + \beta\left(C_x, C_y\right) * l\left(C_x, C_y\right) \tag{4}$$

$$\alpha\left(C_x, C_y\right) = soft\max\left(\theta * W\alpha * \left[v\left(C_x, C_y\right); l\left(C_x, C_y\right)\right]\right) \tag{5}$$

$$\beta\left(C_x, C_y\right) = soft\max\left(\theta * W\beta * \left[v\left(C_x, C_y\right); l\left(C_x, C_y\right)\right]\right) \tag{6}$$

As shown in Eqs. (5) and (6), $\left[v\left(C_x, C_y\right); 1\left(C_x, C_y\right)\right]$ represents the concatenation of visible light and infrared features at position $\left(C_x, C_y\right)$. $W\alpha$ and $W\beta$ are learned weight matrices used to map features into the attention weight space. θ is a hyperparameter that controls the attention distribution. By enhancing the focus on space, the model can automatically pay attention to more useful features when fusing visible light and infrared information, thereby improving task performance.

3.3 Lightweight Convolution Module

To enhance the performance of neural networks, the current research trend is inclined towards constructing deeper and more complex network structures. However, this trend often comes with a sharp increase in network parameters, resulting in excessively large models. Therefore, in response to the issue of large parameter sizes, researchers are increasingly focusing on the development of lightweight networks to seek a balance between performance and model complexity, which has become a highly prominent research direction. As shown in Fig. 5, conventional convolution is an operation that involves multiplying the pixel values within the convolution kernel with the corresponding pixel values in the feature map and then summing all the multiplication results. Unlike previously manually designed feature extraction methods, in deep neural networks, convolutional kernels no longer have predefined fixed parameter filters. On the contrary, they can be learned through the backpropagation algorithm. This means that the network can automatically adjust the parameters of the convolutional kernels based on the characteristics of the data. For instance, the size of the convolutional kernels (height and width) can be adjusted as needed, and the number of channels in the kernels is determined by the depth of the input images or the previous layer's feature maps to better capture the crucial information from the input data. Deep neural networks effectively enhance feature extraction through this learning process.

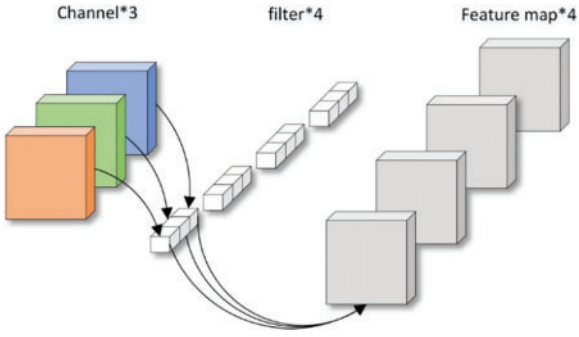

Figure 5: Image convolution map

In this paper, we introduce the Depthwise Separable Convolution module, which has a very simple computational process. In fact, it performs a separable convolution operation. In fact, it performs depth convolution and pointwise convolution operations separately. Unlike traditional convolutional networks where the number of output channels corresponds to the number of convolutional kernels, in depth separable convolution, each channel uses an independent convolutional kernel, resulting in an output channel count of 1 for each channel after the convolution operation. Therefore, the module generates output feature maps equal in channel count to the input feature maps. As shown in Fig. 6,

if the input feature map has N channels, each channel is individually processed by an independent convolutional kernel, resulting in N feature maps, each with a channel number of 1. These N feature maps are then concatenated in order to form an output feature map with N channels.

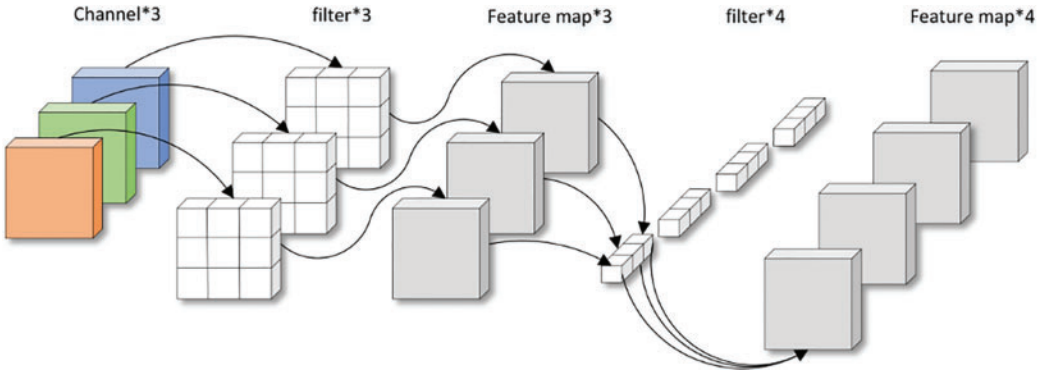

Figure 6: Lightweight convolution

As shown in Fig. 7, the left side represents the standard convolutional layer structure with Batch Normalization and ReLU activation function, while the right side represents the Depthwise Separable Convolutional layer structure with Batch Normalization and ReLU activation function. From the figure, it is evident that after performing depthwise convolution and pointwise convolution operations, batch normalization and ReLU activation are applied to each channel.

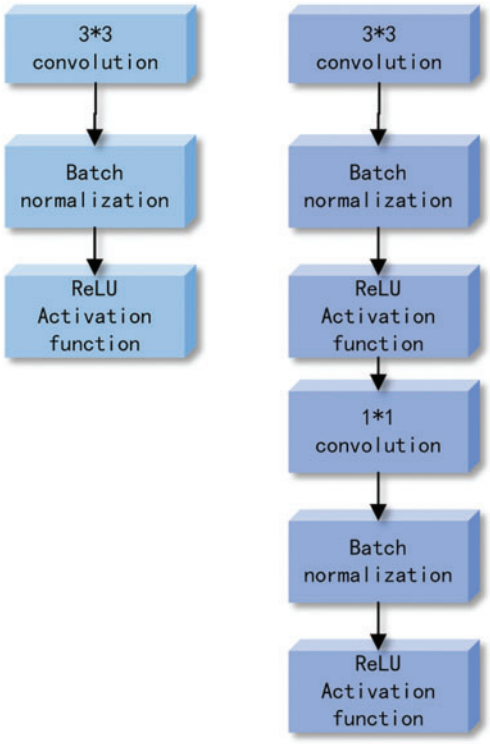

Figure 7: Comparison of Standard Convolution and Depthwise Separable Convolution Structures

During the computation process of the module, it first receives four input parameters: infrared tensor, visible light tensor, depth weight tensor, and normalization weight tensor. The input infrared tensor and visible light tensor are then multiplied by the depth weight tensor and normalization weight tensor, and they are concatenated along the channel dimension. The concatenated result is continuously subjected to operations of L2 normalization, depthwise convolution, pointwise convolution, and batch normalization, ultimately yielding the desired output tensor and weighted feature tensor. The calculation process is shown in Algorithm 1.

Algorithm 1: Lightweight modules

Input: Input rgb tensor $rt(x_i)$, depth weight tensor r_w, lwir tensor $lt(y_j)$, normalization weight tensor l_w
Output: output tensor $ot(z_k)$, processed rgb tensor rt, processed lwir tensor lt

Multiply visible and infrared tensors with weights element by element
Use the following methods to calculate the output tensor and the weighted image tensor
for i = 1 to Length (rgb tensor) **do**
    $rt(x_i) = rt(x_i) * (r_w)$
    for j = 1 to Length (lwir tensor) **do**
        $lt(y_j) = lt(y_j) * (l_w)$
        if i = j **then**
            concat = [rt, lt]
        else
            continue to next iteration of inner loop
        end if
    end for
    concat = L2Normalize (concat, gamma-init = 10)
    ot = Activation (BatchNormalization(PointwiseConv2D(DepthwiseConv2D()(concat))))
end for
return ot, rt, lt

After using depthwise separable convolution, the majority of computations and parameters in the network model are focused on the 1×1 convolutional layer. This design brings the advantage of parallel computation, allowing the network to perform calculations more efficiently. Convolution operations are implemented through matrix multiplication. For conventional square convolutional kernels, we need to rearrange the input features and convolutional kernel parameters, transforming them into matrix form before computation. In contrast, 1×1 convolutional operations do not require this kind of memory reordering and can directly utilize matrix multiplication to complete computations. Therefore, lightweight convolution significantly improves the model's inference speed. This is highly beneficial for addressing potential latency issues when deploying the model on mobile devices.

This paper presents a deep learning network model for cross-modal pedestrian detection based on Resnet50, focusing on the task of cross-modal pedestrian detection. In this model, we fully leverage the advantages of Resnet50 and integrate multi-modal information to achieve more accurate and comprehensive pedestrian detection. The model excels in maintaining high resolution and sensitivity. By integrating different image information within the network, we can capture rich scale information, enabling the model to effectively handle pedestrians of different sizes and distances. This design also helps preserve image details, especially in fine-grained image regions, thereby enhancing the model's detection accuracy. The model uses a spatial reweighted attention mechanism, effectively addressing

the complementarity issue among multimodal features. By passing the weighted feature data to different modal branches, we ensure that each modality can fully utilize its unique information without being influenced by other modalities. Additionally, through the operation of depth convolution, it reduces the model's parameter count, thereby increasing the model's computational speed without sacrificing performance. In practical applications, this balance allows the model to run more efficiently in resource-limited scenarios.

4 Experiments

4.1 Dataset

The multispectral pedestrian dataset provides more valuable information for the generalization capability of multispectral pedestrian detection.

The KAIST pedestrian dataset comprises a total of 95,328 images, as shown in Fig. 8, with each image containing two versions: RGB color and infrared images. In total, it includes 103,128 dense annotations. The dataset was captured during both daytime and nighttime, encompassing various common traffic scenes such as campuses, streets, and rural areas. The image size is 640 × 480. The dataset is divided into a total of 12 folders: set00-set11. The first 6 folders constitute the training set, containing 50,187 images, while the remaining 6 folders make up the test set, comprising 45,141 images.

Figure 8: KAIST dataset presentation image

4.2 Experimental Details

All methods were implemented using the Python programming language in the Tensorflow (version = 1.14) deep learning framework. The training batch size was set to 8, and the entire training process was divided into 20 epochs. Weight decay and momentum were set to 0.0005 and 0.9, respectively, and optimization was performed using the stochastic gradient descent algorithm.

GPUs excel in handling a large number of simple and highly repetitive operations and offer superior computational capabilities compared to CPUs. The experiments in this paper were conducted on an NVIDIA TITAN RTX GPU (Cuda 11.2 version).

In this paper, a comparison will be made with several state-of-the-art methods. These methods are categorized into three groups: (1) The AR-CNN method [33]. This method introduces a novel aligned region convolutional neural network to address end-to-end weakly aligned data issues. By constructing a Region Feature Alignment (RFA) module, it captures positional transformations, enabling adaptive alignment of region features from two modalities. Furthermore, this method introduces a multimodal fusion approach and an ROI (Region of Interest) jitter strategy to further enhance performance. (2) The CFR method [34]. This approach introduces an innovative neural network semi-feature fusion technique by incorporating a specific module into the network structure, cyclically fusing and refining each spectral feature. The purpose is to fully leverage the complementarity and consistency among multispectral features for improved balance and integration. (3) The MSDS-RCNN method [35]. This method proposes a network fusion structure. By jointly optimizing pedestrian detection and semantic segmentation tasks, unified learning of the overall network was achieved. This approach aims to tightly integrate pedestrian detection and segmentation tasks, thereby enhancing the model's performance.

4.3 Result

The superiority of this method is demonstrated from both the false negative rate and speed perspectives.

Miss Rate (MR): As shown in Eq. (7), the indicator for describing the miss rate of detection results in pedestrian detection. TP refers to samples that should be detected by the model and are actually detected, while FN refers to samples that should be detected by the model but are not detected. The sum of TP and FN equals the actual number of pedestrians. Recall indicates the detection rate of annotated pedestrians.

$$\mathrm{MR} = 1 - recall = \frac{FN}{FN + TP} \tag{7}$$

Our proposed method, at an IoU threshold of 0.5, achieved lower values on the reasonable subsets of daytime, nighttime, and the entire day, respectively, compared to the previous best competitor, AR-CNN. This indicates that our network exhibits significantly better localization accuracy compared to AR-CNN. As shown in Table 2, to comprehensively assess the detector's performance, we also evaluated it across all nine subsets, including pedestrian distance and occlusion levels. As depicted in Fig. 9, without additional processing, our network model outperforms other methods in handling small and congested pedestrians in most subsets.

Table 2: Experimental results of different methods on the KAIST dataset

Methods	All	Day	Night	Near	Medium	Far	None	Partial	Heavy
ACF [21]	47.32	42.57	56.17	28.74	53.67	88.20	62.94	81.40	88.02
Halfway fusion [36]	25.75	24.88	26.59	8.13	30.34	75.70	43.13	65.21	74.36
FusionRPN+BF [37]	18.29	19.57	16.27	0.04	30.87	88.86	47.45	0.38	0.41
IAF R-CNN [38]	15.73	14.55	18.26	0.96	25.54	77.84	40.17	48.40	69.76
IATDNN+IASS [39]	14.95	14.67	15.72	0.04	28.55	83.42	45.43	46.25	64.57

(Continued)

Table 2 (continued)

Methods	All	Day	Night	Near	Medium	Far	None	Partial	Heavy
CIAN [40]	14.12	14.77	11.13	3.71	19.04	55.82	30.31	41.57	62.48
MSDS-RCNN [35]	11.63	10.60	13.73	1.29	16.19	63.73	29.86	38.71	63.37
GAFF [41]	10.62	10.82	10.14	–	–	–	–	–	–
CFR [34]	10.05	9.72	10.80	–	–	–	–	–	–
AR-CNN [33]	9.34	9.94	8.38	0.00	16.08	69.00	31.40	38.63	55.73
Ours	8.93	9.16	8.19	0.00	16.62	57.10	28.22	37.25	61.74

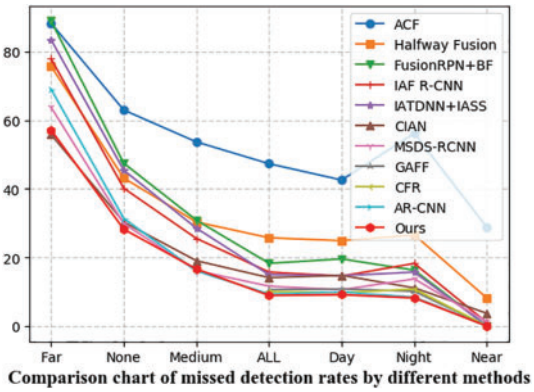

Figure 9: Comparison chart of miss detection rates among different methods

Speed: As shown in Table 3, we compared the runtime of our network model with state-of-the-art methods. Our network directly takes 640 × 512 multispectral images as input without the need for image upscaling, thereby reducing a significant amount of computational load.

Table 3: Computation speeds of different methods on the KAIST dataset

Methods	Platform	Speed (s)
ACF [21]	MATLAB	2.73
Halfway fusion [36]	TITAN X	0.43
FusionRPN + BF [37]	MATLABAB	0.80
IAF R-CNN [38]	TITAN X	0.21
IATDNN+IASS [39]	TITAN X	0.25
CIAN [40]	1080 Ti	0.07
MSDS-RCNN [35]	TITAN X	0.22
GAFF [41]	1080 Ti	0.06
CFR [34]	1080 Ti	0.05
AR-CNN [33]	1080 Ti	0.12
Ours	TITAN RTX	0.07

The speed comparison chart, as shown in Fig. 10, illustrates that our method incorporates an attention mechanism module and introduces a lightweight module while maintaining detection accuracy to achieve a balance. As a result, the model's computational speed has been slightly reduced compared to CFR and GAFF.

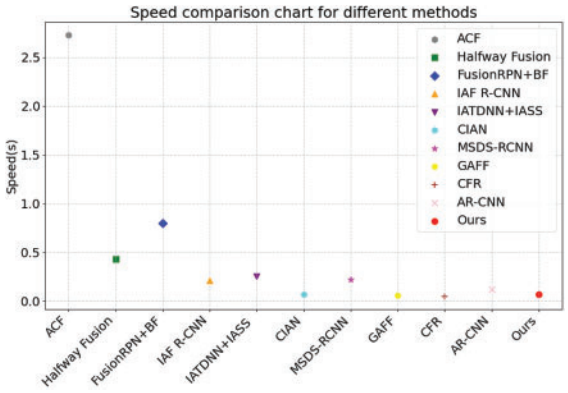

Figure 10: Speed comparison chart among different methods

4.4 Ablation Study

To demonstrate the effectiveness of the proposed spatial reweighted attention mechanism and lightweight module, this paper conducted ablation experiments on the KAIST dataset. We will present experimental results and provide detailed analysis in this section. In Table 4, it can be seen from the experimental results that solely adding the attention module leads to an increase in the model's computational time. After incorporating the lightweight module, model computation speed can be further improved without significantly sacrificing computation accuracy. The final experimental results demonstrate that the simultaneous inclusion of both the attention module and lightweight module achieves the best pedestrian detection performance.

To validate the algorithm's effectiveness, we will compare detection examples of the spatial reweighted attention module and the lightweight module on the KAIST multispectral pedestrian detection dataset for a more intuitive evaluation. As shown in Fig. 11, the experimental detection results of the model on the KAIST dataset are presented from left to right. The first two columns display the detection results with the sole inclusion of the attention module in the network structure, marked with green rectangular boxes. The middle two columns show the detection results with only the inclusion of the lightweight module in the network structure, marked with red rectangular boxes. The last two columns display the detection results with both the attention module and the lightweight module added to the network structure, marked with yellow rectangular boxes. The first two rows consist of visible light images in daytime environments and their corresponding infrared images. The last two rows contain visible light images in nighttime environments and their corresponding infrared images. From the detection results, it can be observed that when only the attention module or the lightweight module is added to the network structure, there are erroneous detections in both daytime and nighttime scenes, resulting in incorrect detection boxes around pedestrian instances. However, when these two modules are combined, the model shows a significant improvement in pedestrian detection performance. In visible light images, some pedestrian instances have complex backgrounds. including pedestrians under shadows, unclear boundaries between different pedestrians, or instances where pedestrians are partially occluded. However, there is no noticeable change in detection results

between daytime and nighttime scenes. This is because the algorithm proposed in this paper merges detailed and semantic information from both modalities into a fused feature with a relatively small modality gap, enhancing the model's generalization ability. The model can better adapt to various scenarios, such as challenging lighting conditions or background occlusion. Therefore, the algorithm in this paper can avoid incorrect detections, resulting in more accurate detection results.

Table 4: Experimental results for various network design choices

Component		choice		
Spatial reweighted attention mechanism		√		√
Depth separable convolution			√	√
MR (%)	Day	10.60	10.60	**9.16**
	Night	7.60	**8.13**	8.19
	All	9.51	9.71	**8.93**
	Medium	16.76	**16.30**	**16.62**
	Far	57.92	**56.48**	57.10
	None	28.53	**27.95**	28.22
	Partial	37.65	**36.98**	37.25
	Heavy	62.28	62.08	**61.74**
Speed (s)		0.11	**0.06**	0.07
Parameters ($\times 10^6$)		12.89	**9.68**	10.82

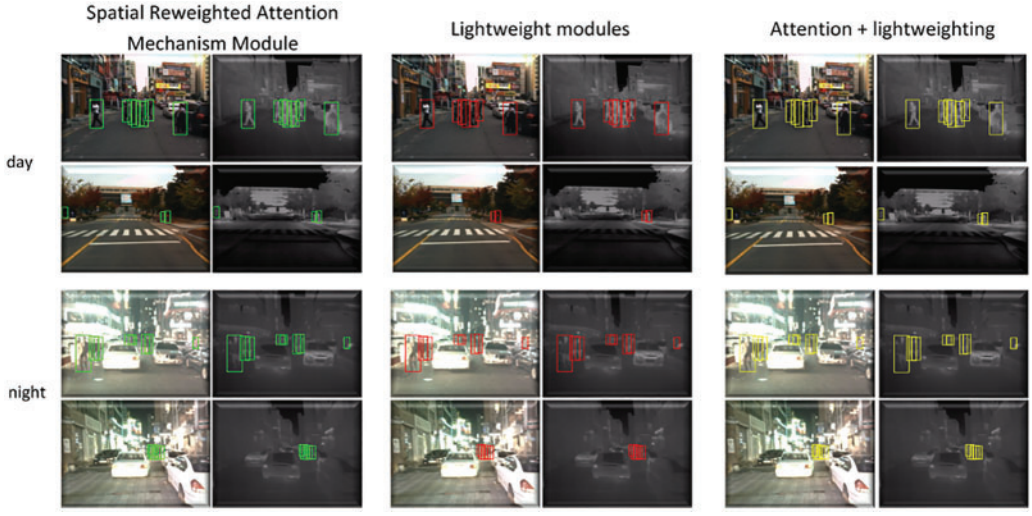

Figure 11: Comparison chart of different module selections

5 Conclusion

Accuracy and speed are both crucial in visual task applications. This paper proposes a deep learning network model for multimodal pedestrian detection, effectively combining spatial reweighted

attention mechanisms with lightweight modules. This enhances the model's attention to various spatial locations while improving the detection speed in multispectral pedestrian detection tasks, balancing the model's accuracy and efficiency. Experimental results indicate that our detection method achieved superior performance on the KAIST multispectral pedestrian dataset compared to other pedestrian detection methods. The experiments demonstrate that the improvement strategy is correct and effective. The overall performance of the improved model is more favorable for terminal deployment applications, achieving a relative balance between detection accuracy and speed to a certain extent. Our research findings can also provide technical support and a basis for studies in related fields such as video data analysis and multi-scene analysis.

Although this paper has made some progress in multispectral pedestrian detection tasks, there are still some challenges and issues that need to be addressed. These include issues related to position offset caused by weak alignment between visible light and infrared images, as well as effectively integrating feature information from different modalities. In the future, we will explore more reasonable multimodal alignment schemes and feature fusion mechanisms to more effectively utilize bimodal features, contributing valuable insights to the development of multispectral applications.

Acknowledgement: S. Hwang; J. Park; N. Kim; Y. Choi; I. S. Kweon. We would like to thank the above researchers for building the KAIST dataset.

Funding Statement: This paper was supported by the Henan Provincial Science and Technology Research Project under Grants 232102211006, 232102210044, 232102211017, 232102210055 and 222102210214, the Science and Technology Innovation Project of Zhengzhou University of Light Industry under Grant 23XNKJTD0205, the Undergraduate Universities Smart Teaching Special Research Project of Henan Province under Grant Jiao Gao [2021] No. 489-29, the Doctor Natural Science Foundation of Zhengzhou University of Light Industry under Grants 2021BSJJ025 and 2022BSJJZK13.

Author Contributions: Conceptualization, L.D.; Data curation, L.D. and R.F.; Formal analysis, L.D. and Z.L.; Investigation, Z.L.; Methodology, L.D. and R.F.; Software, M.X. and B.L.; Supervision, Z.L. and Y.C.; Writing–review & editing, R.F.

Availability of Data and Materials: The data presented in this study are openly available in (KAIST) at (10.1109/CVPR.2015.7298706), reference number [16].

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] C. Yan, H. Zhang, X. Li, Y. Yang, and D. Yuan, "Cross-modality complementary information fusion for multispectral pedestrian detection," *Neural. Comput. Appl.*, vol. 35, pp. 10361–10386, 2023. doi: 10.1007/s00521-023-08239-z.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[3] D. Ghose, S. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau and T. Rahman, "Pedestrian detection in thermal images using saliency maps," presented at the 2019 IEEE/CVF CVPRW, Long Beach, CA, USA, Jun. 16–17, 2019.

[4] C. Wang, A. Bochkovskiy, and H. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," arXiv:2207.02696, 2022.
[5] Q. Deng, W. Tian, Y. Huang, L. Xiong, and X. Bi, "Pedestrian detection by fusion of RGB and infrared images in low-light environment," presented at the 2021 IEEE 24th Int. Conf. Fusion., Sun City, South Africa, Nov. 1–4, 2021, pp. 1–8.
[6] M. Zahid, M. Khan, F. Azam, M. Sharif, S. Kadry and J. R. Mohanty, "Pedestrian identification using motion controlled deep neural network in real-time visual surveillance," *Soft Comput.*, vol. 27, pp. 453–469, 2021.
[7] P. Wang, L. Zhou, M. Xiao, and P. Zhang, "Multi-spectral fusion network for full-time robust pedestrian detection," presented at the 2021 Int. Conf. EIECT, Kunming, China, Oct. 29–31, 2021.
[8] G. Li, W. Lai, and X. Qu, "Pedestrian detection based on light perception fusion of visible and thermal images," *Optics & Laser Technol.*, vol. 156, pp. 108466, 2022. doi: 10.1016/j.optlastec.2022.108466.
[9] Y. Zhang, B. Zhai, G. Wang, and J. Lin, "Pedestrian detection method based on two-stage fusion of visible light image and thermal infrared image," *Electronics.*, vol. 12, no. 14, pp. 3171, 2023. doi: 10.3390/electronics12143171.
[10] A. Benjumea, I. Teeti, F. Cuzzolin, and A. Bradley, "YOLO-Z: Improving small object detection in YOLOv5 for autonomous vehicles," arXiv:2112.11798, 2021.
[11] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for pedestrian detection," presented at the 2018 Conf. ECCV, Munich, Germany, Sep. 8–14, 2018, pp. 732–747.
[12] M. Marnissi, H. Fradi, A. Sahbani, and N. Amara, "Unsupervised thermal-to-visible domain adaptation method for pedestrian detection," *Pattern Recogn. Lett.*, vol. 153, pp. 222–231, 2022. doi: 10.1016/j.patrec.2021.11.024.
[13] S. Kanwal *et al.*, "Person re-identification using adversarial haze attack and defense: A deep learning framework," *Comput. Electr. Eng.*, vol. 96, pp. 107542, 2021. doi: 10.1016/j.compeleceng.2021.107542.
[14] M. Raza, M. Sharif, M. Yasmin, M. Khan, T. Saba and S. Fernandes, "Appearance based pedestrians' gender recognition by employing stacked auto encoders in deep learning," *Future Gener. Comput. Syst.*, vol. 88, pp. 28–39, 2018. doi: 10.1016/j.future.2018.05.002.
[15] M. Kieu, A. Bagdanov, and M. Bertini, "Bottom-up and layerwise domain adaptation for pedestrian detection in thermal images," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 1, pp. 1–19, 2021. doi: 10.1145/3418213.
[16] W. Lee, L. Jovanov, and W. Philips, "Cross-modality attention and multimodal fusion transformer for pedestrian detection," presented at the 2022 Conf. ECCV, Aviv, Israel, Oct. 23–27, 2022, pp. 608–623.
[17] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan and L. Shao, "Mask-guided attention network for occluded pedestrian detection," presented at the 2019 IEEE/CVF ICCV, Seoul, Korea (South), Oct. 27–Nov. 02, 2019, pp. 4967–4975.
[18] C. Feng, Z. Cao, Y. Xiao, Z. Fang, and J. Zhou, "Multi-spectral template matching based object detection in a few-shot learning manner," *Inform. Sci.*, vol. 624, pp. 20–36, 2023. doi: 10.1016/j.ins.2022.12.067.
[19] M. Rashid, M. A. Khan, M. Sharif, M. Raza, M. M. Sarfraz and F. Afza, "Object detection and classification: A joint selection and fusion strategy of deep convolutional neural network and SIFT point features," *Multimed. Tools Appl.*, vol. 78, pp. 15751–15777, 2019. doi: 10.1007/s11042-018-7031-0.
[20] K. Lai, J. Zhao, X. Huang, and L. Wang, "Research on pedestrian detection using optimized mask R-CNN algorithm in low-light road environment," *J. Phys. Conf. Ser.*, vol. 1777, no. 1, pp. 012057, 2021. doi: 10.1088/1742-6596/1777/1/012057.
[21] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," presented at the 2015 IEEE Conf. CVPR, Boston, MA, USA, Jun. 07–12, 2015, pp. 1037–1045.
[22] D. Heo, E. Lee, and B. Ko, "Pedestrian detection at night using deep neural networks and saliency maps," *Electronic Imaging*, vol. 2018, no. 17, pp. 060403-1–060403-9, 2018.

[23] Q. Wang, Y. Chi, T. Shen, J. Song, Z. Zhang and Y. Zhu, "Improving rgb-infrared pedestrian detection by reducing cross-modality redundancy," presented at the 2022 IEEE Int. Conf. ICIP, Bordeaux, France, Oct. 16–19, 2022, pp. 526–530.
[24] D. Pei, M. Jing, H. Liu, and F. Sun, "A fast RetinaNet fusion framework for multi-spectral pedestrian detection," *Infrared Phys. Technol.*, vol. 105, pp. 103178, 2020. doi: 10.1016/j.infrared.2019.103178.
[25] Y. Lee, T. Bui, and J. Shin, "Pedestrian detection based on deep fusion network using feature correlation," presented at the 2018 APSIPA ASC, Honolulu, HI, USA, Nov. 12–15, 2018, pp. 694–699.
[26] X. Yang, Y. Qian, H. Zhu, C. Wang, and M. Yang, "BAANet: Learning bi-directional adaptive attention gates for multispectral pedestrian detection," presented at the 2022 ICRA, Philadelphia, PA, USA, May 23–27, 2022, pp. 2920–2926.
[27] X. Zuo, Z. Wang, Y. Liu, J. Shen, and H. Wang, "LGADet: Light-weight anchor-free multispectral pedestrian detection with mixed local and global attention," *Neural Process. Lett.*, vol. 55, no. 3, pp. 2935–2952, 2023. doi: 10.1007/s11063-022-10991-7.
[28] Z. Cao, H. Yang, J. Zhao, S. Guo, and L. Li, "Attention fusion for one-stage multispectral pedestrian detection," *Sens.*, vol. 21, no. 12, pp. 4184, 2021. doi: 10.3390/s21124184.
[29] Y. Zhang, Z. Yin, L. Nie, and S. Huang, "Attention based multi-layer fusion of multispectral images for pedestrian detection," *IEEE Access*, vol. 8, pp. 165071–165084, 2020. doi: 10.1109/ACCESS.2020.3022623.
[30] Q. Jiang, J. Dai, T. Rui, F. Shao, J. Wang and G. Lu, "Attention-based cross-modality feature complementation for multispectral pedestrian detection," *IEEE Access.*, vol. 10, pp. 53797–53809, 2022. doi: 10.1109/ACCESS.2022.3175303.
[31] Y. Zang, C. Fu, D. Yang, H. Li, C. Ding and Q. Liu, "Transformer fusion and histogram layer multispectral pedestrian detection network," *Signal, Image and Video Process.*, vol. 17, pp. 1–9, 2023. doi: 10.1007/s11760-023-02579-y.
[32] J. Ryu, J. Kim, H. Kim, and S. Kim, "Multispectral interaction convolutional neural network for pedestrian detection," *Comput. Vis. Image Underst.*, vol. 223, no. 12, pp. 103554, 2022. doi: 10.1016/j.cviu.2022.103554.
[33] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," presented at the 2019 IEEE/CVF ICCV, Seoul, Korea (South), Oct. 27–Nov. 02, 2019, pp. 5127–5137.
[34] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Multispectral fusion for object detection with cyclic fuse-and-refine blocks," presented at the 2020 IEEE ICIP, Abu Dhabi, United Arab Emirates, Oct. 25–28, 2020, pp. 276–280.
[35] D. Guan, Y. Cao, J. Liang, Y. Cao, and M. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Inform. Fusion*, vol. 50, pp. 148–157, 2019. doi: 10.1016/j.inffus.2018.11.017.
[36] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," arXiv:1611.02644, 2016.
[37] D. König, 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 CVPRW, Honolulu, HI, USA, Jul. 21–26, 2017, pp. 243–250.
[38] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *Pattern Recognit.*, vol. 85, pp. 161–171, 2019. doi: 10.1016/j.patcog.2018.08.005.
[39] L. Zhang *et al.*, "Cross-modality interactive attention network for multispectral pedestrian detection," *Inform. Fusion*, vol. 50, pp. 20–29, 2019. doi: 10.1016/j.inffus.2018.09.015.
[40] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," arXiv:1808.04818, 2018.
[41] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Guided attentive feature fusion for multispectral pedestrian detection," presented at the 2021 IEEE WACV, Waikoloa, HI, USA, Jan. 03–08, 2021, pp. 72–80.