frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 19 January 2024
DOI 10.3389/fphy.2023.1322232

OPEN ACCESS

EDITED BY
Arianna Morozzi,
Istituto Nazionale di Fisica Nucleare di Perugia, Italy

REVIEWED BY
Guohui Wang,
Xi'an Technological University, China
Chengfang Zhang,
Sichuan Police Academy, China

*CORRESPONDENCE
Jinjin Wang,
22207223042@stu.xust.edu.cn

RECEIVED 16 October 2023
ACCEPTED 21 December 2023
PUBLISHED 19 January 2024

CITATION
Ma L, Wang J, Dai X and Gao H (2024), Deep saliency detection-based pedestrian detection with multispectral multi-scale features fusion network.
Front. Phys. 11:1322232.
doi: 10.3389/fphy.2023.1322232

Deep saliency detection-based pedestrian detection with multispectral multi-scale features fusion network

Li Ma[1,2], Jinjin Wang[1,2]*, Xinguan Dai[1,2] and Hangbiao Gao[3]

[1]College of Communication and Information Engineering, Xi'an University of Science and Technology, Xi'an, China, [2]Xi'an Key Laboratory of Heterogeneous Network Convergence Communication, Xi'an, China, [3]Safety Supervision Department, Shaanxi Cuijiagou Energy Co., Ltd., Tongchuan, China

In recent years, there has been increased interest in multispectral pedestrian detection using visible and infrared image pairs. This is due to the complementary visual information provided by these modalities, which enhances the robustness and reliability of pedestrian detection systems. However, current research in multispectral pedestrian detection faces the challenge of effectively integrating different modalities to reduce miss rates in the system. This article presents an improved method for multispectral pedestrian detection. The method utilises a saliency detection technique to modify the infrared image and obtain an infrared-enhanced map with clear pedestrian features. Subsequently, a multiscale image features fusion network is designed to efficiently fuse visible and IR-enhanced maps. Finally, the fusion network is supervised by three loss functions for illumination perception, light intensity, and texture information in conjunction with the light perception sub-network. The experimental results demonstrate that the proposed method improves the logarithmic mean miss rate for the three main subgroups (all day, day and night) to 3.12%, 3.06%, and 4.13% respectively, at "reasonable" settings. This is an improvement over the traditional method, which achieved rates of 3.11%, 2.77%, and 2.56% respectively, thus demonstrating the effectiveness of the proposed method.

KEYWORDS

multispectral pedestrian detection, visible and infrared image, saliency detection, multiscale feature fusion, illumination perception

1 Introduction

Pedestrian target detection is a fundamental and basic task for various applications, including autonomous driving [1, 2] and intelligent video surveillance [3, 4]. It has been a popular research topic for decades, and significant progress has been made in recent years. However, developing a highly reliable and robust pedestrian detector for practical applications remains a challenge. Many pedestrian detection methods [5, 6] only use visible light images, which can lead to inadequate detection in complex environments (e.g., cloudy and rainy weather, insufficient light, and cluttered backgrounds, etc.). Pedestrian temperature during the day similar to the ambient temperature or lower than the surrounding objects, makes infrared image characterization insignificant. In or poor light conditions or at night, visible light images may not fully characterise the imaged scene. Therefore, relying solely on a single modality for detection has significant limitations. To overcome this, a widely used technique is to fuse infrared and visible light images [7, 8] to

achieve a more effective and comprehensive characterisation of the scene. This enhances the robustness of pedestrian target detection methods. Despite the significant progress made in pedestrian detection using infrared and visible images in past research, there are still several areas that require improvement.

The development of various image fusion techniques, including traditional approaches [9] and deep learning techniques [10], has enabled researchers to achieve better results when combining infrared and visible images. Conventional image fusion algorithms involve mapping the image to the transform domain, followed by applying a set of fusion rules to the two images to create a single image that combines infrared and visible information. Representative methods for complex scenarios include multiscale transform [11], sparse representation [12], and subspace learning [13]. Traditional algorithms have achieved certain results, but their high reliance on manual design makes it challenging to adapt to complex situations and their time-consuming nature imposes limitations. It is of great importance to explore alternative methods that can overcome these limitations.

Recent advances in deep neural network technology have demonstrated their ability to extract features and combine multiple modes. Therefore, deep learning frameworks have become increasingly popular for multispectral information fusion. Li et al. [8] proposed DenseFuse, which introduces dense connectivity into an encoder network to extract underlying image features for feature reuse. In the following step, the encoder fuses the deep features extracted by the encoder using the over-L1 paradigm or addition. Finally, the decoder generates the fused image. Additionally, deep features are challenging to interpret, so merging strategies developed manually cannot accurately determine the weights and, therefore, do not adequately cover the attributes of deep features. To prevent the limitations of an artificially constructed merger approach, Ma et al. [10] proposed an alternative method for image combining. They introduced the image combining problem into a generative network that is commonly used to resolve conflict problems between features. Meanwhile, Hou et al. [14] determined the saliency of the source image based on pixel intensities, which guided the fusion network in preserving rich salient information from the source images when generating images.

Although the network structures described above perform well in recovering image detail, they are designed at a single scale, which limits their ability to capture contextual information in the image. To better meet the need for broader context awareness, this limitation needs to be addressed. While there have been some successes, deep learning-based approaches still face a number of obstacles that need to be carefully considered, in particular the problem of light imbalance, which has not yet been fully investigated. Light imbalance, which refers to the difference in lighting conditions between daytime and nighttime scenes, has not yet been extensively studied [15, 16]. While visible images typically have sharper texture details, infrared images provide more salient targets and richer texture information at night compared to visible images. However, current approaches [17, 18] generally assume that texture information is only present in images visible to the human eye. This assumption is reasonable in daytime scenes, but at night, the fused image may lose texture details, which can affect its quality.

In order to achieve a better fusion of the two modalities and thus improve the robustness of the pedestrian detector, we propose a novel multispectral pedestrian detection method in this paper.

First, a saliency detection method is used to improve the accuracy of the detection algorithm by enhancing the pedestrian features to overcome the problem of inconspicuous pedestrian features in daytime images. Then, to solve the problem of information loss in merged images, a multiscale feature extraction method is used to capture various contextual information from images at different scales. The image fusion network can comprehend the content and structure of the image by introducing multiple parallel branches, each extracting and fusing features at different levels. This multiscale approach enables the network to encode both local and global features simultaneously, enhancing its ability to handle complex scenarios.

To address the problem of illumination imbalance, we implemented an illumination-aware sub-network that determines the probability of visible images during the day and at night. This sub-network provides valuable information regarding the lighting conditions of visible images. The introduction of this sub-network not only improves the quality and accuracy of image fusion methods but also enhances their efficiency. This paper primarily contributes to the following areas, as outlined below:

(1) This paper proposes an improved method for multi-spectral pedestrian detection. It utilises a deep saliency detection technique to obtain saliency maps of infrared images, which enhances pedestrian features. This method solves the problem of inconspicuous pedestrian features in infrared images during daytime.
(2) A self-coding Multi-scale Image Features Fusion (MIFF) network is designed. MIFF uses a feature fusion module to combine different modal features at the same scale. The image quality can be improved by fusing features of different dimensions;
(3) An illumination-aware subnetwork has been developed to calculate the probability of visible images during day and night. To facilitate fusion network training, illumination-aware loss, light intensity loss, and texture information loss have also been developed. By detecting the illumination, the meaningful information of the source image can be fused around the clock.
(4) The experimental results demonstrate that our approach outperforms other advanced methods in detecting pedestrians when applied to the publicly available KAIST dataset.

2 Related technologies and principles

The first section of this chapter provides an overview of the saliency target detection method, followed by a description of deep learning-based fusion methods, as well as a brief overview of multispectral pedestrian detection capabilities and algorithms. Furthermore, we intend to introduce illumination awareness as a key concept aimed at improving the effectiveness of multimodal image fusion.

2.1 Principles of saliency detection technology

Saliency detection, a method of highlighting targets in salient regions of an image, has seen a number of approaches emerge over the last few decades. Traditional methods of saliency detection include global contrast [19], local contrast [20] and hand-crafted saliency-based approaches such as colour and texture [21].

The majority of current saliency detection methods implement saliency detection tasks with convolutional neural networks. In these methods, the core task of saliency detection is to calculate the weight and extract the salient target information by analyzing the infrared image. This is then used to calculate weights and extract salient information about the target. Specifically, firstly, we pass through a weight calculation stage to subtly divide the original image into an underlying image and a detail image. Then, we applied a specialised saliency detection method to generate the saliency maps of the bottom and detail maps respectively. Furthermore, the weight distribution for the bottom image and detail image was determined through the clever integration of these saliency maps.

Salient target extraction is an important part of the saliency detection process. Its main goal is to extract information about salient regions from visible and infrared images, reflecting the key information of the image. Hou et al. [22] introduced short connections to the network architecture and proposed a saliency detection method that considers hopping connections for the first time. Luo et al [23] developed an innovative multi-resolution mesh structure that combines local and global information. Zhao et al [24] proposed a saliency detection network that uses feature and attention modules to capture richer contextual information based on pyramidal feature attention. Zhou et al. [25] proposed a confidence-aware saliency extraction (CSD) method that obtains rich saliency knowledge from noisy labels. The experimental results demonstrate that the obtained saliency maps have more precise boundaries. Although these innovative methods promote the development and research of saliency detection, a single saliency detection method may not be sufficient to express all features. Therefore, it is necessary to find more effective ways to express the features of saliency detection.

Saliency detection techniques are employed to fuse the original and saliency maps of infrared images. The information fusion method enhances the accuracy and reliability of salient target detection under infrared spectral conditions by capturing saliency features in images. Classical saliency detection methods are closely related to saliency detection tasks, which improves the recognition and analysis of salient targets in complex scenes.

2.2 Deep learning-based image fusion methods

Deep learning-based image fusion methods are a current research focus and Frontier in image fusion. These methods efficiently and accurately fuse images from various data sources. Neural networks are well-suited to this task due to their ability to model nonlinear functions. In deep learning image fusion, there are four main methodologies: autoencoder (AE), convolutional neural networks (CNNs), generative adversarial networks (GANs), and multiscale feature fusion.

2.2.1 Auto-encoder (AE)-based image fusion method

An Auto-Encoder (AE)-based image fusion method, comprised of an encoder and a decoder, belongs to the unsupervised learning neural network class. The encoder compresses the input image, and the decoder remaps the compressed features back to the original data space. The method involves two main processes: feature extraction and image reconstruction. DRF et al [26] independently fuse the source image after decomposing it into scene and attribute components. However, this method only addresses interpretability in terms of feature extraction and not fusion methods. Xu et al. [27] have developed an image fusion method based on Auto-Encoder (AE) technology to facilitate image interpretation. To enhance the interpretability of the fusion method, they designed multiple encoders to extract specific features. The experiments showed that the Auto-Encoder (AE)-based image fusion method demonstrates superior performance in dealing with the problem of fusing images from different sources. Furthermore, due to its efficient training and low computational costs, this method has become prevalent in image fusion applications.

2.2.2 Convolutional neural network (CNN)-based image fusion method

The CNN-based image fusion method, based on the Encode-Decoder framework, uses two independent encoders to combine the features of the input and output images. Zhang et al [28] propose an end-to-end image fusion framework that implements the process using the intensity and gradient paths while maintaining the ratio of gradient and intensity. Xu et al [29] examine the interactions between different image fusion tasks and co-train a unified model to solve multitask fusion by using flexible weighting. Additionally, the attention-based mechanism enables the network to accurately distinguish important information in different regions by introducing an attention model for more precise image fusion.

2.2.3 Generative adversarial network (GAN)-based image fusion method

Generative Adversarial Network (GAN)-based image fusion methods utilise a generative adversarial network including a generator and a discriminator. The generator maps the input image and the image to be fused to a high-quality output image, and the discriminator determines the consistency between the generator output and the real fused image. Considering the fusion of infrared images and IR images as a game between a generator and the discriminator innovatively, Ma et al [10] use the discriminator to drive the generator to synthesise a more texturally-rich fused image. To solve the single classification problem, Ma et al [30] propose a GAN-based multiclassification fusion method, which transforms image fusion into multiclassification by introducing a multiclassifier to more comprehensively and significantly enhance the texture information and global contrast of the fused image. The whole training process ensures that the final fused image reflects multimodal data information. It is recommended that the

generator output be gradually increased to achieve the maximum fusion effect. A multiclassifier has resulted in a significant improvement in image quality, as well as a considerable improvement in the visual effect of the fused image. As a result, we can introduce more texture details and contrast in image brightness during the image fusion process.

2.2.4 Method based on multi-scale image features fusion

The Method based on multiscale image features fusion is used to combine information from different sources or different feature representations to improve the performance of a task. It usually involves integrating features from multiple scales to obtain richer and more accurate image information capable of enhancing the model. Li et al [31], who have developed a fusion strategy based on spatial or channel attention, proposed a multi-scale image fusion autoencoder framework, NestFusion, in which the encoder extracts the multiscale features using sequential downsampling, and the decoder uses honeycomb connections to fuse multiscale features, thereby enhancing the details of the background and the salient regions of the image; in order to solve the problem of unlearnable fusion strategy of NestFuse, Li et al. propose the RFN-Nest fusion network [32], and further designed the Residual Fusion Network (RFN) on the basis of the NestFuse framework to replace the hand-crafted fusion strategy. In high design complexity, the above multi-scale feature network architectural models have high requirements on computational power, memory consumption, and graphics memory capacity, thus hindering their application on resource-constrained devices. In addition, the training of the RFN-Nest fusion strategy is separate from the training of the encoder and decoder, which means that features of different modes cannot be extracted effectively. Therefore, in this paper, a self-encoder-based Multiscale Image Features Fusion (MIFF) network is designed.

2.3 Illumination aware

In fact, some real-world computer vision applications have incorporated illumination into the modelling stage. Sakkos et al [33] developed a three-fold multi-task generated opponent for the network, with the goal of fusing functions from different illumination conditions on the branch division, which significantly improves the foreground division property. Li et al [34] proposed an illumination-aware Faster R-CNN that achieves adaptive convergence of visible and infrared image sub-networks by introducing a gating function on the output from an illumination-aware network.

However, multi-modal datasets present a key problem: visible images contain useful information, while infrared images capture supplementary information. The classical approach has addressed the problem of under-capture of IR images at night, but it is difficult for visible and infrared images to capture important information in darkness or low light conditions, and the fusion images still need more texture information to achieve better results.

Therefore, we propose an improved approach that introduces an illumination-aware subnetwork designed to enhance meaningful information in visible images under dark or poor illumination conditions. By doing so, we can make better use of meaningful information when fusion occurs.

Specifically, the introduction of an illumination-aware sub-network enables us to perform targeted enhancement of visible images based on ambient illumination conditions. By combining the illumination-aware sub-network with other image fusion methods, this approach enhances the quality and information richness of the fused image under different lighting conditions. The use of multimodal data further improves the performance of the fused images in different environments.

Overall, we propose a method that eliminates information fusion in multimodal datasets by using visible and infrared images simultaneously. Additionally, the introduction of an illumination-aware sub-network under dark or poor illumination conditions helps to achieve more accurate and richer image fusion.

2.4 Multispectral pedestrian detection

Pedestrian detection methods use spectral images for target detection. Monospectral approaches rely only on a single gamma or multicolour object, but are susceptible to interference from factors such as illumination, colour and texture. However, variations in these factors may cause the degradation of the performance of traditional methods under different conditions, limiting their application in complex scenes.

During the past few years, multispectral imaging technology has expanded pedestrian detection areas. Multispectral pedestrian detection extracts more feature information, reduces interference and improves accuracy through the combination of multispectral bands, including acquisition of multispectral images, data pre-processing, feature extraction and pedestrian detection. Multispectral pedestrian detection has stronger anti-interference ability and accuracy compared to single-spectrum methods. However, it requires more data processing and storage resources.

Kim et al [35] proposed a Multispectral Pedestrian Detection Framework that incorporated enhancement methods and multiple labels for pedestrian detection. Tang et al [36] developed a bi-directional image alignment module and introduced semantic constraints based on segmentation to address fusion network requirements. Ding et al. [37] developed a multispectral pedestrian single-shot detection method that integrates visible and infrared image information to balance accuracy and speed. Li et al. [25] proposed confidence-aware multispectral pedestrian detection algorithms that fuse different branches of the image for prediction based on Dempster-Shafer theory. The experimental results demonstrate an improvement in detection accuracy. Nati Ofir et al [38] developed a new method for achieving multispectral image fusion and retaining a large amount of detail, based on hyperpixel segmentation.

These innovative methods strongly support the research and practical application of multispectral pedestrian detection, and promote the development and progress of this field. The continuous evolution of technology is expected to enable multispectral pedestrian detection to demonstrate its potential in various fields and offer more precise and reliable solutions to practical problems.

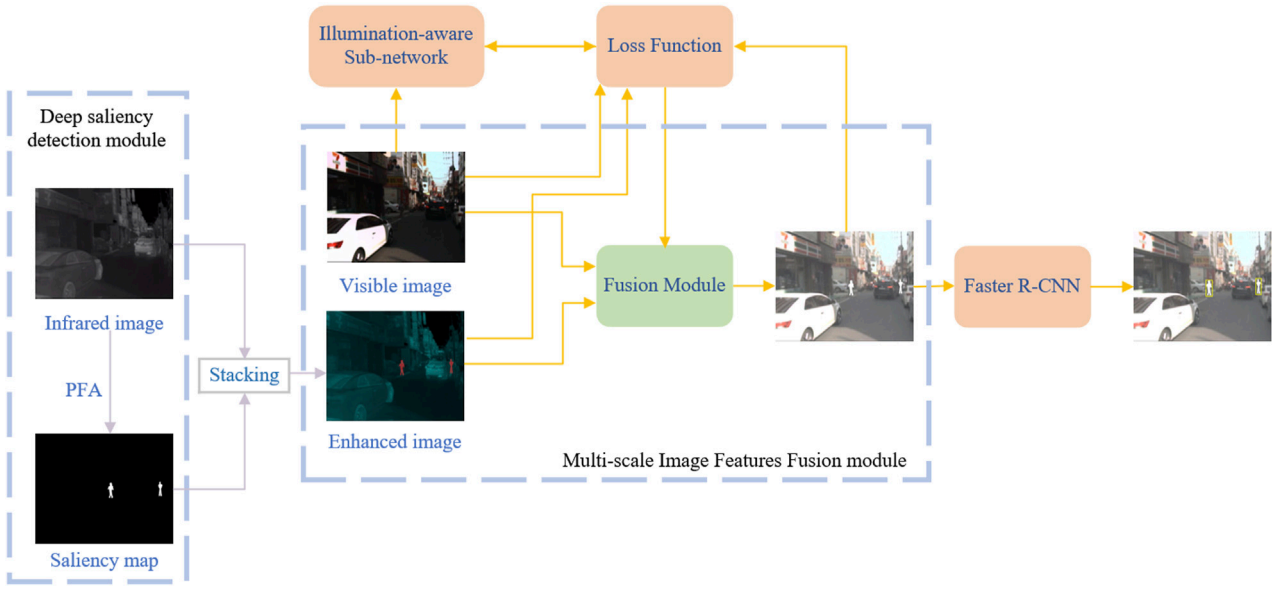

FIGURE 1
In this paper, we propose a general structure for the multispectral pedestrian detection method. Adapted from KAIST Multispectral Pedestrian Detection Benchmark, licensed under CC BY-NC-SA 4.0.

3 Methods of work

In this section, we present our proposal for a multispectral pedestrian detection method which includes three original contributions: saliency detection, multiscale image feature fusion, and illumination awareness. Detailed information about each contribution will be described in this section.

3.1 Overall structure

As shown in Figure 1, our proposed improved pedestrian detection method is based on a technical framework.

Firstly, a saliency map of the infrared image is generated by using saliency detection technique to enhance the pedestrian features in the image. However, the saliency map discards all texture information from the IR image. To solve this problem, we enhanced the IR image with a saliency map by replacing one channel of the three-channel IR image with the saliency map to obtain the enhancement map. By using this method, we can produce a synthesised image with significant pedestrian features. In addition, we retain other texture information in the infrared image.

Following the enhancement of the image and the visible image, the images are fed into a multiscale image feature fusion network. This is done to create the final fused image. In this fusion network, we introduce the feature enhancement fusion module, the main task of which is to efficiently merge these two features at the same scale. This innovative module improves image fusion so that the final fused image generated is more compatible with the desired visual criteria and requirements.

The illumination-aware sub-network is introduced to capture key details of the image under different lighting conditions, thus enhancing the quality and information richness of the merged image. By sensing lighting conditions, this sub-network helps capture and retain meaningful information in fused source image, which improves the network's performance under different lighting conditions, resulting in more accurate and effective fused images.

To train the fusion network effectively, we introduce several loss functions, including illumination-aware loss, light intensity loss, and texture information loss, which jointly guide the training process of the fusion network. Illumination-aware loss helps the model to better understand the lighting conditions in the image. Light intensity loss helps to control the overall brightness of the image. Loss of texture information preserves image details and texture characteristics.

By combining various loss functions, we can train fusion networks that are more adaptable to different lighting conditions and image contents, thereby enhancing the quality and robustness of image fusion.

Finally, Faster R-CNN is employed for target detection to achieve improved multispectral pedestrian detection.

3.2 Deep saliency detection

The paper presents a saliency map of the infrared image, which aids the pedestrian detector in performing better at night. Saliency detection techniques are classified into two categories: static saliency detection and depth saliency detection.

The static saliency detection method is mainly based on assumptions such as high contrast between salient targets and background, simple background, and single light source. Manual features with color and texture contrast are designed, or information such as spatial position priors are introduced to measure the saliency of images. However, manual features and *a priori* cannot capture high-level and global semantic information about a given object.

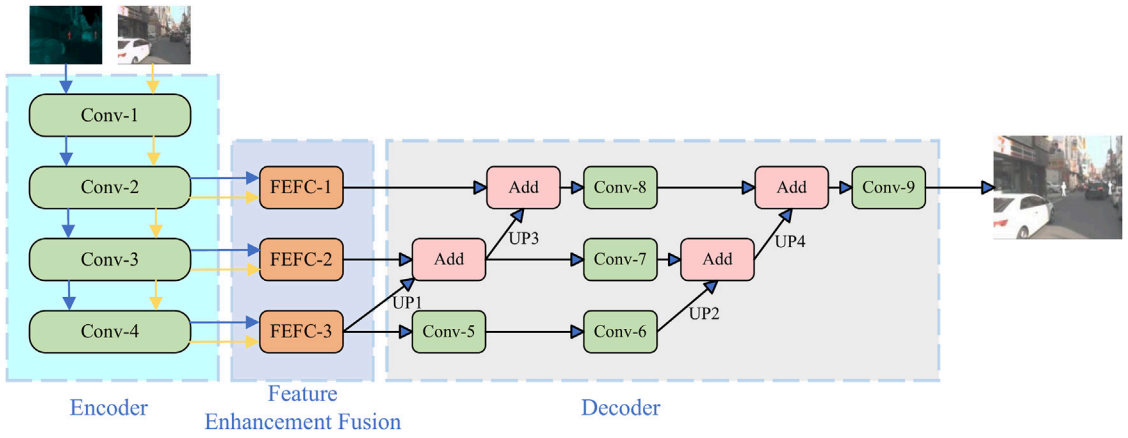

FIGURE 2
Auto-Encoder based multi-scale image feature fusion network. Adapted from KAIST Multispectral Pedestrian Detection Benchmark, licensed under CC BY-NC-SA 4.0.

In recent years, a large number of scholars have begun to study deep neural network-based saliency detection methods. The significance detection was further improvement in accuracy by exploiting and utilising the properties of the individual feature maps of the contextual neural network. Proposed in PFA [24] (Pyramid Feature Attention Network), the deep saliency detection method effectively focuses on both low- and high-level spatial structure features for detecting saliency levels. The method has a powerful feature extraction capability and can accurately locate salient targets. Therefore, this paper uses PFA (Pyramid Feature Attention Network) as the saliency detection method. Details are described in the following.

3.3 Auto-encoder based multiscale image feature fusion

Figure 2 illustrates the network structure for the multiscale feature fusion network described in this paper, which employs an end-to-end training approach. The network comprises three sections: an encoder module, a feature enhancement fusion module, and a decoder module. Enhanced and visible Images as a result of the input are represented by $\mathbf{I}_{en}$ and $\mathbf{I}_{vi}$. $\mathbf{I}_{f}$ represent the final image after the fusion process is completed.

Due to the large variability of the two images, two mutually independent encoder sections are employed in this paper. These sections determine the multiscale features of the enhanced image and the visible image, respectively. For a single input image, four convolutional layers are used in the encoder. With Conv-1, convolutional layers map the input image to higher dimensions. Conv-2 is a 3 × 3 convolutional layer used to extract shallow features from the image. There are two maximal pooling layers, Conv-3 and Conv-4, which include a convolutional layer of 1 × 1 and a maximum pooling layer with a downsampling rate of 2. This is designed to extract features with high-level semantic information. It is recommended that the encoder output be divided into three scales: Level 1, Level 2, and Level 3. An enhanced and visible image feature is included with each scale.

A Feature Enhancement Fusion Module (FEFM) consists of three Feature Enhancement Fusion Layers (FEFCs). These layers fuse augmented map features and visible map features derived from the decoder at each scale in the system. The design and structure of the feature augmentation fusion layers are shown in Figure 3.

Firstly, the Feature Enhancement Fusion Layer implemented a 3 × 3 convolution operation on the decoder's augmented and visible map feature pairs at each scale to further extract the depth features, respectively, to obtain features $\mathbf{F}^{i}_{en}$ and $\mathbf{F}^{i}_{vi}$, where $\mathbf{i} \in \{1, 2, 3\}$ denotes the scale level at which they are located. The enhancement fusion of features from different modalities is then performed based on the idea of differential amplifiers. Features $\mathbf{F}^{i}_{en}$ and $\mathbf{F}^{i}_{vi}$ can be represented in terms of their common features and their respective private features with the following equations:

$$\mathbf{F}^{i}_{en} = \frac{\mathbf{F}^{i}_{en} + \mathbf{F}^{i}_{en}}{2} + \frac{\mathbf{F}^{i}_{vi} - \mathbf{F}^{i}_{vi}}{2} = \frac{\mathbf{F}^{i}_{en} + \mathbf{F}^{i}_{vi}}{2} + \frac{\mathbf{F}^{i}_{en} - \mathbf{F}^{i}_{vi}}{2} \tag{1}$$

$$\mathbf{F}^{i}_{vi} = \frac{\mathbf{F}^{i}_{vi} + \mathbf{F}^{i}_{vi}}{2} + \frac{\mathbf{F}^{i}_{en} - \mathbf{F}^{i}_{en}}{2} = \frac{\mathbf{F}^{i}_{en} + \mathbf{F}^{i}_{vi}}{2} + \frac{\mathbf{F}^{i}_{vi} - \mathbf{F}^{i}_{en}}{2} \tag{2}$$

In this paper, let $\mathbf{D}^{i}_{en-vi} = (\mathbf{F}^{i}_{en} - \mathbf{F}^{i}_{vi})/2$, $\mathbf{C}^{i}_{en+vi} = (\mathbf{F}^{i}_{en} + \mathbf{F}^{i}_{vi})/2$, $\mathbf{D}^{i}_{vi-en} = (\mathbf{F}^{i}_{vi} - \mathbf{F}^{i}_{en})/2$, where $\mathbf{C}^{i}_{en+vi}$ denotes the public features extracted from the enhanced image and the visible image, $\mathbf{D}^{i}_{en-vi}$ and $\mathbf{D}^{i}_{vi-en}$ denote the respective private features. Then the public and private feature states of the enhanced image and the visible image are enhanced with features using ECA (Efficient Channel Attention) module and then aggregated to get the fusion features $\mathbf{F}^{i}_{f}$, the mathematical expression for Eq. 3 can be seen in equation. Where ECA (−) denotes the Efficient Channel Attention module for attenuating the effect of redundant channels on the model performance.

$$F^{i}_{f} = ECA\left(D^{i}_{en-vi}\right) + ECA\left(D^{i}_{vi-en}\right) + ECA\left(C^{i}_{ir+en}\right) \tag{3}$$

Efficient Channel Attention schematic is shown in Figure 4. After applying Global Average Pooling (GAP) to the existing feature map $\mathbf{X}$, the resulting feature map will be approximately 1 × 1 in size. In order to determine the association between each channel and its four neighboring channels, it is applied to one-dimensional convolution with a kernel size of four. In addition, the importance weight of each channel is determined

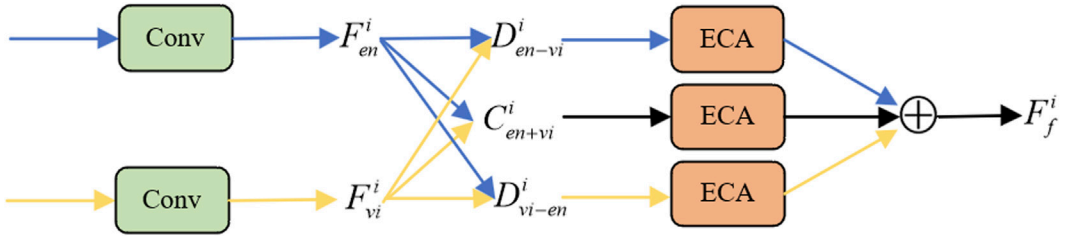

FIGURE 3
Feature enhanced fusion layer architecture.

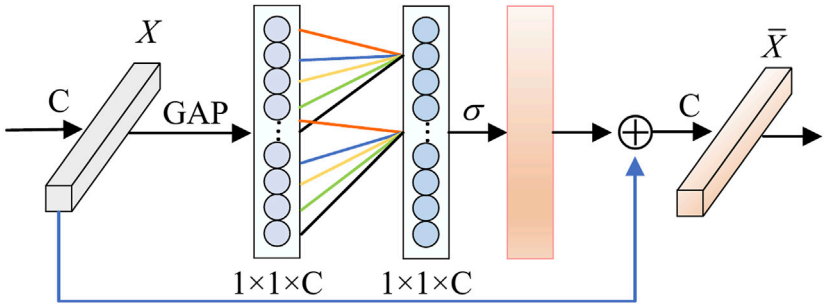

FIGURE 4
Diagram of the ECA module.

TABLE 1 Network parameters for encoders and decoders.

	Layer	Input	k	s	Padding	n1	n2	Activate fnuction	Output
Encoder	Conv-1	640 × 512	1	1	—	3	16	ReLU	640 × 512
	Conv-2	640 × 512	3	1	Same	16	32	ReLU	640 × 512
	Conv-3	640 × 512	1	1	—	32	64	ReLU	320 × 256
	Conv-4	320 × 256	1	1	—	64	128	ReLU	160 × 128
Decoder	Conv-5	160 × 128	3	1	Same	128	128	-	160 × 128
	Conv-6	160 × 128	3	1	Same	128	128	-	160 × 128
	UP1	160 × 128	1	1	—	128	64	-	320 × 256
	UP2	160 × 128	1	1	—	128	64	-	320 × 256
	Conv-7	320 × 256	3	1	Same	64	64	-	320 × 256
	Conv-8	640 × 512	3	1	Same	32	32	-	640 × 512
	UP3	320 × 256	1	1	—	64	32	-	640 × 512
	UP4	320 × 256	1	1	—	64	32	-	640 × 512
	Conv-9	640 × 512	3	1	Same	32	3	Tanh	640 × 512

through the sigmoid activation function. A final output feature map $\bar{\mathbf{X}}$ is derived by multiplying the corresponding elements of the original input feature map. It is possible to determine channel attention weights for effective channel interaction through the Efficient Channel Attention module, since it takes into account cross-channel information interaction without dimensionality reduction.

The enhanced image and the visible image are obtained after the encoder module and the feature enhancement fusion module to obtain the enhanced features at three different scales $\mathbf{F}_{\mathbf{f}}^{\mathbf{1}}$, $\mathbf{F}_{\mathbf{f}}^{\mathbf{2}}$, and $\mathbf{F}_{\mathbf{f}}^{\mathbf{3}}$. The upward arrows in the decoder module in Figure 2 denote the up-sampling module, which consists of the 1 × 1 convolution and up-sampling operations. Add denotes the summation operation. The decoder

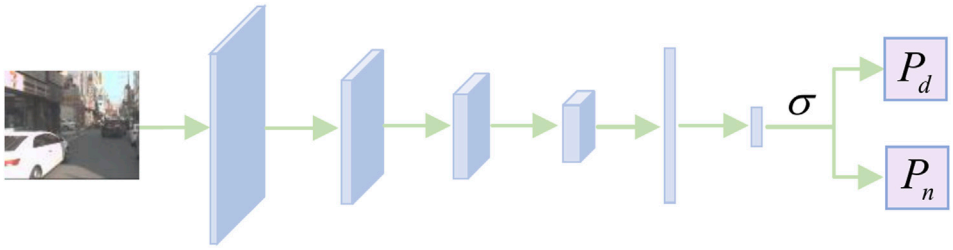

FIGURE 5
Schematic diagram of the lighting sensing sub-network. Adapted from KAIST Multispectral Pedestrian Detection Benchmark, licensed under CC BY-NC-SA 4.0.

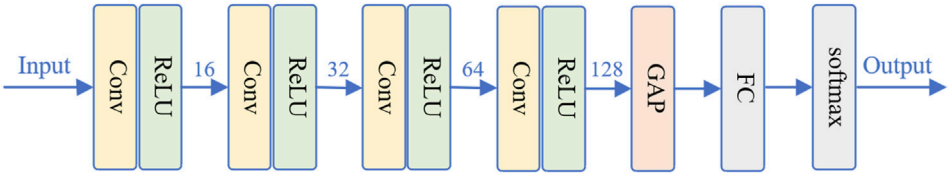

FIGURE 6
Illumination-aware sub-network structure.

undergoes a series of up-sampling, summation and convolution operations to obtain the final fused image $\mathbf{I_f}$. Table 1 lists the network construction the parameters of the encoder and decoder, including the concentration kernel size (k), the movement step (s), the number of input channels (n1), the number of output channels (n2), and the up-sampling module (UP).

3.4 Light factor based illumination-aware sub-network

Our research focuses on developing an Illumination-aware sub-network that approximates light conditions. Figure 5 below shows a schematic representation of this network and the specific network structure is shown in Figure 6. By feeding a visible image into this network, we performed a four-layer convolution operation to extract deep features of the image. Then, we transformed these features into a vector and generated a set of lighting probabilities by a global average pooling layer, a layer of fully connected neural network combined with a Sigmoid based activation function $\{\mathbf{P_d}, \mathbf{P_n}\}$. Where $\mathbf{P_d}$ indicates the approximate probability that the image is in daylight, and $\mathbf{P_n}$ indicates the approximate probability that the image is in nighttime. By combining multiscale image features, the multiscale fusion network creates a multiscale image that incorporates meaningful information from both enhanced and visible images. In this report, guidance is provided regarding illumination-aware loss, light intensity loss, and texture information loss.

3.5 Fusion loss function

As a means of ensuring that the multiscale image feature fusion network can adaptively fuse meaningful information using illumination information, we propose an illumination-aware loss $\mathbf{L_{ill}}$, which has the following definition:

$$\mathbf{L_{ill}} = \mathbf{P_d} \times \frac{\mathbf{1}}{\mathrm{HW}} \|\mathbf{I_f} - \mathbf{I_{vi}}\|_\mathbf{1} + \mathbf{P_n} \times \frac{\mathbf{1}}{\mathrm{HW}} \|\mathbf{I_f} - \mathbf{I_{en}}\|_\mathbf{1} \tag{4}$$

It is significant to note that H and W represent the image's height and width, and $\|\cdot\|_\mathbf{1}$ denotes the $\mathbf{L_1}$-paradigm. This loss adjusts merged image intensity constraints according to illumination probability. Consequently, the luminance information in the source image will be dynamically updated. There is, however, the possibility that the fused image may not maintain optimal luminance distribution. In order to address this issue, we will introduce the concept of light intensity loss $\mathbf{L_{aux}}$, defined as follows:

$$\mathbf{L_{aux}} = \frac{\mathbf{1}}{\mathrm{HW}} \|\mathbf{I_f} - \mathbf{max}(\mathbf{I_{en}}, \mathbf{I_{vi}})\|_\mathbf{1} \tag{5}$$

In this definition, $\max(\cdot)$ means taking the maximum value of each pixel of each image.

Considering the importance of texture details in the detection process, we then introduce texture information loss L_{tex}, which is defined as follows:

$$\mathbf{L_{tex}} = \frac{\mathbf{1}}{\mathrm{HW}} \|\|\nabla \mathbf{I_{en}}| - |\nabla \mathbf{I_{vi}}|)\|_\mathbf{1} \tag{6}$$

Assuming that ∇ denotes the gradient operator, we compute the gradient in this paper using the Sobel operator. $|\cdot|$ represents the operation of taking absolute values.

Finally, our total loss L_{total} is defined as:

$$\mathbf{L_{total}} = \mathbf{L_{ill}} + \lambda_\mathbf{1} \mathbf{L_{aux}} + \lambda_\mathbf{2} \mathbf{L_{tex}} \tag{7}$$

Where λ_1 and λ_2 are two balancing coefficients to regulate the percentage of each loss.

We have developed a multiscale image feature fusion network that dynamically maintains optimal intensity distribution no matter what the illumination situation is. This is under the guidance of illumination-aware loss, light intensity loss, and texture information loss. The ideal texture details can be obtained under the guidance of texture information loss. These loss functions ensure that we can get meaningful fused images throughout the day.

4 Experimentation

This section first presents a publicly available dataset and implementation details. Following this, Using three classical methods as a comparison, we verified the excellent performance of our proposed method. Finally, ablation experiments are conducted to determine module efficiency.

4.1 Datasets

The KAIST dataset is widely recognised in academia as a benchmark evaluation standard for multispectral pedestrian detection tasks. We chose to use this dataset mainly because it contains a rich diversity of pedestrian images covering a variety of complex scenarios including occlusions, multi-scale variations, and different background conditions. This diversity allows us to validate and evaluate the robustness and accuracy of the experimental results in a more comprehensive way. The KAIST dataset covers two viewpoints, i.e., the horizontal viewpoint and the top viewpoint. The horizontal viewpoint dataset is collected from urban environments and covers a variety of different scenes and complex backgrounds. The top view angle dataset, on the other hand, is taken from places such as school campuses and car parks, and contains dense groups of pedestrians as well as complex environmental backgrounds. In the training set of the KAIST dataset, an image is captured every two frames, covering a total of 25,086 images. The test set, on the other hand, captures an image every thirty frames, totalling 2,252 images. Of these, the daytime test set includes 1,455 images and the nighttime test set contains 797 images. To ensure fairness in comparing the results with our competitors, In order to evaluate the results, we used all reasonable subsets of scale and occlusion from the KAIST test dataset in [39].

4.2 Implementation details

This study used the Python version 3.8 and PyTorch 1.9.0 deep learning framework located on Ubuntu 20.04 operating system. We randomly initialised the network parameters, settings $\boldsymbol{\lambda_1 = 2}$, $\lambda_2 = \mathbf{0.4}$. Our inputs were KAIST dataset images of original size 640 × 512 without any resizing operations. We used the Adam optimiser to train our proposed basic target detection network with an initial learning rate of 0.005 and a default batch size of 32, and the number of training periods was set to 100. We reduced the learning rate by a factor of 10 when the training loss was no longer decreasing and the validation accuracy was no longer improving. Subsequently, we continued to reduce the learning rate by two times until we eventually stopped training. All models were trained on GeForce RTX 3090 GPUs from NVIDIA. To compare different models' performance, evaluation metrics follow the standard KAIST evaluation, where we used the log-averaged MR (FPPI) as the evaluation metric, measured in the MR^{-2} interval in the range $[10^{-2}, 10^{0}]$.

4.3 Comparison of research with classical methods

We compared with ACF + T + THOG [39], Halfway Fusion [40], FusionRPN + BDT [41], IAF R-CNN [42], IATDNN + IASS [43], MSDS-RCNN [44], CIAN [45], MBNet [46] and LG-FAPF [47]. Comparison. We evaluated all the detection results on the KAIST test set.

As shown in Table 2, as in previous studies, we conducted experiments on three main subsets (all-day, daytime, and nighttime) as well as six other subsets under the "reasonable" setting, and observed that our method demonstrated excellent performance. The experimental the consequences show that our model is more efficient able to perform multimodal data fusion with improved speed and accurate detection.

In Figure 7, we present a comparison of our method with three different state-of-the-art multispectral pedestrian detectors MSDS-RCNN [44], MBNet [46] and LG-FAPF [47]. Our method is able to successfully detect pedestrians in very low light or darkness and effectively suppress human-like false alarms. This highlights that our method has excellent robustness and accuracy in pedestrian detection tasks. This means that we are better able to address the challenges of pedestrian detection in complex scenes and provide more reliable detection results.

The "reasonable" setting presents greater difficulties due to increased occlusions and unclear images, particularly at night. Figure 8 shows the FPPI-MR curves for the "reasonable" setting on the KAIST dataset. The MR-FPPI plots of our proposed multispectral pedestrian detection method are displayed alongside those of other existing methods. In the range of $[10^{-2}, 10^{0}]$, the grey curve consistently outperforms the blue curve. The results indicate that our method is more accurate than the latest LG-FAPF method. Table 2 further demonstrates the superiority of our approach.

4.4 Parameter sensitivity experiments

In order to verify the stability of the model to hyperparameters, we conducted parameter sensitivity experiments on all-day dataset under "reasonable" settings. We set the value of λ_1 to {0.02, 0.2, 2, 20, 200} and the value of λ_2 to {0.04, 0.4, 4, 40, 400}. As shown in Table 3, the experimental results show that the proposed model can get better detection results in larger parameter spaces. Specifically, the values of λ_1 and λ_2 cannot be too large or too small, and the best detection results are achieved when the value of λ_1 is equal to 2 and the value of λ_2 is equal to 0.4.

4.5 Ablation studies

To authenticate the validity of the significance detection, multiscale image feature fusion and illumination perception, we

TABLE 2 An analysis of the KAIST dataset compares our method with nine representative state-of-the-art multispectral pedestrian detectors, using log-averaged performance as the evaluation metric.

Methods	KAIST (MR^{-2})								
	Reasonable			Six subsets					
	All.day	Day	Night	Near	Medium	Far	None	Partial	Heavy
ACF + T + THOG [39]	46.48	40.57	54.17	28.63	51.71	85.57	54.69	70.49	81.50
Halfway Fusion [40]	27.07	25.88	28.59	8.01	30.69	74.57	47.44	61.17	72.04
FusionRPN + BDT [41]	19.91	16.67	21.38	0.12	30.87	84.42	40.57	46.66	70.46
IAF R-CNN [42]	15.87	14.55	18.26	1.02	23.12	71.16	41.65	46.48	63.03
CIAN [45]	14.92	14.77	16.13	3.71	19.04	55.82	30.31	41.57	62.48
IATDNN + IASS [43]	14.35	14.67	15.72	0.03	27.01	80.10	40.85	47.36	62.13
MSDS-RCNN [44]	9.61	9.09	10.92	1.26	16.13	67.36	31.22	37.67	60.62
MBNet [46]	8.52	8.28	9.86	**0.00**	16.07	55.99	27.74	35.43	59.14
LG-FAPF [47]	6.23	5.83	6.69	0.58	8.44	40.47	18.92	22.01	50.24
Ours	**3.12**	**3.06**	**4.13**	0.46	**6.32**	**32.56**	**14.87**	**17.34**	**46.15**

We tested the three main subsets (all-day, daytime, and nighttime) as well as six other subsets under "reasonable" settings. Bold indicates best in performance.

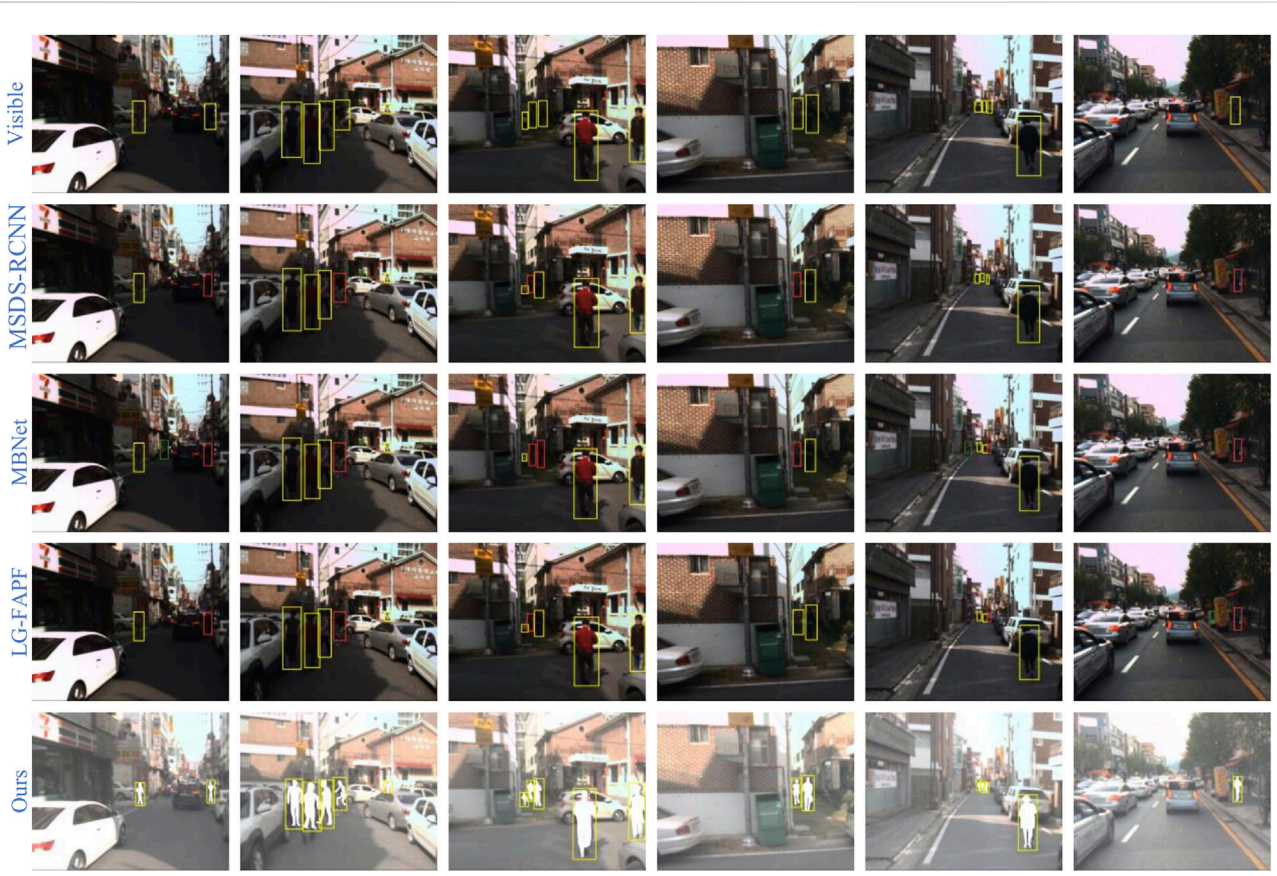

FIGURE 7
Comparison with three state-of-the-art multispectral pedestrian detectors (MSDS-RCNN [44], MBNet [46], and LG-FAPF [47]) by means of a uniform confidence threshold. The first series of columns shows the results of the reference pedestrian detection, which is performed on a visible image. The other columns show the pedestrian detection results for different methods on visible images. Yellow boundary boxes indicate positive tags, red boundary boxes indicate ignored tags, and green boundary boxes indicate false alarm tags. Adapted from KAIST Multispectral Pedestrian Detection Benchmark, licensed under CC BY-NC-SA 4.0.

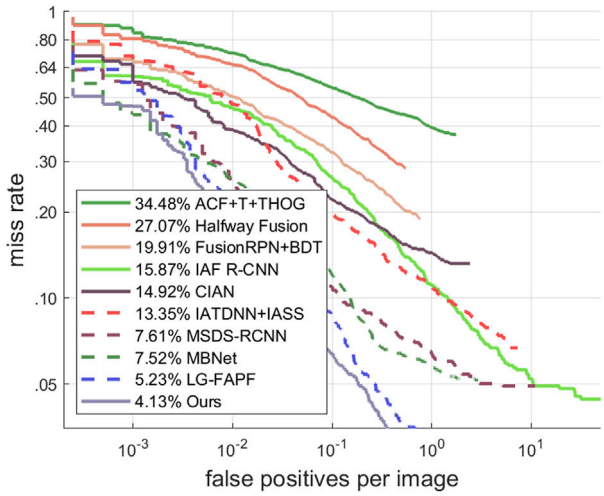

FIGURE 8
The analysis of detection results was compared using the KAIST dataset.

TABLE 3 Parameter sensitivity experiments on the KAIST all-day dataset.

KAIST All.day (MR^{-2})					
$\lambda_1\lambda_2$	0.02	0.2	2	20	200
0.04	6.69	5.42	5.01	6.34	9.97
0.4	6.13	4.79	**3.12**	3.98	10.36
4	7.47	5.37	4.23	7.01	12.49
40	9.02	7.43	8.60	11.42	10.27
400	14.64	14.73	13.22	13.61	18.36

Bold indicates best in performance.

TABLE 4 Ablation experiments on the KAIST dataset.

Methods	All day	Day	Night
(I)	24.79	19.71	36.55
(II)	20.91	26.54	14.89
(III)	15.39	18.21	7.66
(IV)	7.08	8.46	5.31
Ours	3.12	3.06	4.13

carried out ablation experiments. The experimental results are shown in Table 4.

Where, (I) denotes direct detection of visible images. (II) denotes detection directly on infrared images. (III) denotes detection using enhancement maps. (IV) denotes the addition of a multi-scale feature fusion network to (III), at which point P_d and P_n are removed from the loss function. Ours denotes the addition of an illumination-aware sub-network to (IV), which is our complete network.

From (I) and (II), it is evident that pedestrian detection using visible images performs better during the daytime, while infrared images perform better at night. This highlights the limitations of relying on a single modality for detection. The results of method (III) indicate that the use of augmented images can significantly enhance pedestrian detection effectiveness for the three subsets under the "reasonable" settings. This is because the saliency detection can highlight pedestrian features in the images. Method (IV) outperforms method (III) by using a multi-scale image feature fusion network. It is also possible to fuse the enhanced image and the visible image at high quality. Therefore, our fusion network has been demonstrated to be effective. Our results are better than those of method (IV) because the illumination-aware subnetwork can adaptively fuse the two images according to the ambient lighting conditions, which improves the quality of the fused images as well as the final detection results.

5 Conclusion

In this paper, a novel multi-spectral pedestrian detection method is proposed. Firstly, in this method, we use saliency detection technique to enhance the pedestrian features in infrared images and successfully solves the problem of inconspicuous pedestrian features in infrared images during daytime; then, a multi-scale image feature fusion network based on self-encoder is designed to effectively fuse different modal features at the same scale, which improves the feature extraction capability and the quality of image fusion; finally using an illumination-aware sub-network, the problem of light imbalance is solved, and meaningful information of the fused images is fused around the clock. Numerous experimental results demonstrate that our algorithm has excellent performance on the KAIST dataset. Specifically, our algorithm improves the logarithmic mean leakage rate (MR) by 3.11%, 2.77%, and 2.56% for three major subsets of different time periods (all-day, daytime, and nighttime) under the "reasonable" setting, respectively. Our results are significantly improved compared to classical multispectral pedestrian detection methods currently considered state-of-the-art. We aim to contribute to multispectral pedestrian detection advancement through our research.

In future work, we would further investigate finer image fusion methods to better fuse the bimodal features and further improve the detection performance. Meanwhile, we will also design lighter modules to speed up pedestrian detection so that our method can be applied to real-time scenarios. We hope that these developments will improve the accuracy and efficiency of pedestrian detection and will provide a reliable and efficient solution for real-world implementations.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The requirement of ethical approval was waived by Xi'an University of Science and Technology for the studies involving humans because Xi'an University of Science and Technology. The studies were conducted in accordance with the local legislation and institutional requirements.

The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

LM: Methodology, Writing–original draft, Writing–review and editing. JW: Investigation, Validation, Writing–original draft. XD: Project administration, Writing–review and editing. HG: Formal Analysis, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Key Industry Innovation Chain Project of Shaanxi Key Research and Development Plan (2021ZDLGY07–08).

Acknowledgments

I would like to express my gratitude to Professor Mary Ma for proofreading the first draft of my thesis and providing invaluable guidance and advice. I also thank XD for his teachings, which have been of great benefit to me. Lastly, I express my gratitude to Xi'an University of Science and Technology for providing me with an education that I will cherish for a lifetime. Additionally, I extend my thanks to the grant project "2021ZDLGY07-08" for its financial support.

Conflict of interest

Author HG was employed by the company Shaanxi Cuijiagou Energy Co., Ltd.

The remaining 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

References

1. Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The kitti vision benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 2012; Providence, RI, USA (2012). p. 3354–61.

2. Geronimo D, Lopez AM, Sappa AD, Graf T. Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans Pattern Anal Mach Intell* (2009) 32(7): 1239–58. doi:10.1109/tpami.2009.122

3. Wang X, Wang M, Li W. Scene-specific pedestrian detection for static video surveillance. *IEEE Trans Pattern Anal Mach Intell* (2013) 36(2):361–74. doi:10.1109/TPAMI.2013.124

4. Li X, Ye M, Liu Y, Zhang F, Liu D, Tang S. Accurate object detection using memory-based models in surveillance scenes. *Pattern Recognit* (2017) 67:73–84. doi:10.1016/j.patcog.2017.01.030

5. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; June 2005; San Diego, CA, USA (2005). p. 886–93.

6. Gool LV, Mathias M, Timofte R, Benenson R. Pedestrian detection at 100 frames per second. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; June 2012; Providence, RI, USA (2012). p. 2903–10.

7. Ma J, Ma Y, Li C. Infrared and visible image fusion methods and applications: a survey. *Inf Fusion* (2019) 45:153–78. doi:10.1016/j.inffus.2018.02.004

8. Li H, Wu XJ. DenseFuse: a fusion approach to infrared and visible images. *IEEE Trans Image Process* (2018) 28(5):2614–23. doi:10.1109/tip.2018.2887342

9. Zhou Z, Wang B, Li S, Dong M. Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters. *Inf Fusion* (2016) 30:15–26. doi:10.1016/j.inffus.2015.11.003

10. Ma J, Yu W, Liang P, Li C, Jiang J, FusionGAN . FusionGAN: a generative adversarial network for infrared and visible image fusion. *Inf Fusion* (2019) 48:11–26. doi:10.1016/j.inffus.2018.09.004

11. Liu Y, Jin J, Wang Q, Shen Y, Dong X. Region level based multi-focus image fusion using quaternion wavelet and normalized cut. *Signal Process.* (2014) 97:9–30. doi:10.1016/j.sigpro.2013.10.010

12. Zhang Q, Maldague X. An adaptive fusion approach for infrared and visible images based on NSCT and compressed sensing. *Infrared Phys Tech* (2016) 74:11–20. doi:10.1016/j.infrared.2015.11.003

13. Liu Y, Chen X, Ward RK, Jane Wang Z. Image fusion with convolutional sparse representation. *IEEE Signal Process. Lett* (2016) 23(12):1882–6. doi:10.1109/lsp.2016.2618776

14. Hou R, Zhou D, Nie R, Liu D, Xiong L, Guo Y, et al. VIF-Net: an unsupervised framework for infrared and visible image fusion. *IEEE Trans Comput Imaging* (2020) 6: 640–51. doi:10.1109/tci.2020.2965304

15. Casal JJ, Candia AN, Sellaro R. Light perception and signalling by phytochrome A. *J Exp Bot* (2013) 65:2835–45. doi:10.1093/jxb/ert379

16. Gundel PE, Pierik R, Mommer L, Ballaré CL. Competing neighbors: light perception and root function. *Oecologia* (2014) 176:1–10. doi:10.1007/s00442-014-2983-x

17. Carvalho RF, Takaki M, Azevedo RA. Plant pigments: the many faces of light perception. *Acta Physiol Plant* (2011) 33:241–8. doi:10.1007/s11738-010-0533-7

18. Sanchez SE, Rugnone ML, Kay SA. Light perception: a matter of time. *Mol Plant* (2020) 13:363–85. doi:10.1016/j.molp.2020.02.006

19. Cheng M, Mitra NJ, Huang X, Torr PHS, Hu S. Global contrast based salient region detection. *IEEE Trans Pattern Anal Machine Intelligence* (2015) 37(3):569–82. doi:10.1109/tpami.2014.2345401

20. Klein DA, Frintrop S. Center-surround divergence of feature statistics for salient object detection. In: Proceedings of the 2011 International Conference on Computer Vision; November 2011; Barcelona, Spain (2011). p. 2214–9.

21. Cheng M-M, Mitra NJ, Huang X, Torr PH, Hu S-M. Global contrast based salient region detection. *IEEE Trans Pattern Anal Machine Intelligence* (2015) 37(3):569–82. doi:10.1109/tpami.2014.2345401

22. Hou Q, Cheng M-M, Hu X, Borji A, Tu Z, Torr P. Deeply supervised salient object detection with short connections (2017). Available at: https://arxiv.org/abs/1611.04849.

23. Luo Z, Mishra AK, Achkar A, Eichel JA, Li S, Jodoin P-M. Non-local deep features for salient object detection. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 2017; Honolulu, HI, USA (2017). p. 7.

24. Zhao T, Wu X. Pyramid feature attention network for saliency detection (2019). Available at: https://arxiv.org/abs/1903.00179.

25. Li Q, Zhang C, Hu Q, Fu H, Zhu P. Confidence-awareFusion using dempster-shafer theory for multispectral pedestrian detection. *IEEE Trans Multimedia* (2022): 3160589. doi:10.1109/TMM.2022.3160589

26. Xu H, Wang X, Ma J. Drf: disentangled representation for visible and infrared image fusion. *IEEE Trans Instrumentation Meas* (2021) 70:1–13. doi:10.1109/tim.2021.3056645

27. Xu H, Wang X, Ma J. Drf: disentangled representation for visible and infrared image fusion. *IEEE Trans Instrum Meas* (2021) 70:1–13. doi:10.1109/tim.2021.3056645

28. Zhang H, Xu H, Xiao Y, Guo X, Ma J, Rethinking the image fusion: a fast unified image fusion network based on proportional maintenance of gradient and intensity, In: Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto and AAAI (2020), p. 12797–804.

29. Xu H, Ma J, Jiang J, Guo X, Ling H. U2Fusion: a unified unsupervised image fusion network. *IEEE Trans Pattern Anal Mach Intell* (2022) 44(1):502–18. doi:10.1109/tpami.2020.3012548

30. Ma J, Zhang H, Shao Z, Liang P, Xu H. GANMcC: a generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Trans Instrum Meas* (2021) 70:1–14.

31. Hui LI, Wu X, Durrani T. NestFuse: an infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Trans Instrumentation Meas* (2020) 69(12):9645–56. doi:10.1109/tim.2020.3038013

32. Hui LI, Wu X, Kittler J. RFN-Nest: an end-to-end residual fusion network for infrared and visible images. *Inf Fusion* (2021) 73:72–86. doi:10.1016/j.inffus.2021.02.023

33. Sakkos D, Ho ES, Shum HP. Illumination-aware multi-task GANs for foreground segmentation. *IEEE Access* (2019) 7:10976–86. doi:10.1109/access.2019.2891943

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

35. Kim J, Kim H, Kim T, Kim N, Choi Y. MLPD: multi-label pedestrian detector in multispectral domain. *IEEE Robotics and Autumation* (2021) 6:7846–53. doi:10.1109/lra.2021.3099870

36. Tang L, Deng Y, Ma Y, Huang J, Ma J. SuperFusion: a versatile image registration and fusion network with semantic awareness. *IEEE/CAA J Automatica Sin.* (2022) 9:2121–37. doi:10.1109/jas.2022.106082

37. Ding L, Wang Y, Laganière R, Huang D, Luo X, Zhang H. A robust and fast multispectral pedestrian detection deep network. *Knowledge-Based Syst* (2021) 229:106990. doi:10.1016/j.knosys.2021.106990

38. Zhou H, Qiao B, Yang L, Lai J, Xie X. Texture-guided saliency distilling for unsupervised salient object detection (2023). Available at: https://arxiv.org/abs/2207.05921.

39. Hwang S, Park J, Kim N, Choi Y, So Kweon I. Multispectral pedestrian detection: benchmark dataset and baseline. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; June 2015; Boston, MA, USA. IEEE (2015). p. 1037–45.

40. Liu J, Zhang S, Wang S, Metaxas DN. Multispectral deep neural networks for pedestrian detection (2016). Available at: https://arxiv.org/abs/1611.02644.

41. König D, Adam M, Jarvers C, Layher G, Neumann H, Teutsch M. Fully convolutional region proposal networks for multispectral person detection. In: Proceedings of the Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on; July 2017; Honolulu, HI, USA. IEEE (2017). p. 243–50.

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

43. 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

44. Li C, Song D, Tong R, Tang M. Multispectral pedestrian detection via simultaneous detection and segmentation. In: Proceedings of the British Machine Vision Conference 2018, BMVC 2018; September, 2018; Newcastle, UK. Northumbria University (2018).

45. 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

46. Zhou K, Chen L, Cao X. Improving multispectral pedestrian detection by addressing modality imbalance problems (2020). Available at: https://arxiv.org/abs/2008.03043.

47. Cao Y, Luo X, Yang J, Cao Y, Yang MY. Locality guided cross-modal feature aggregation and pixel-level fusion for multispectral pedestrian detection. *Inf Fusion* (2022) 88:1–11. doi:10.1016/j.inffus.2022.06.008