frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 26 March 2024
DOI 10.3389/fphy.2024.1356248

OPEN ACCESS

EDITED BY
Zhiqin Zhu,
Chongqing University of Posts and Telecommunications, China

REVIEWED BY
Yonghang Tai,
Yunnan Normal University, China
Taisong Jin,
Xiamen University, China
Jie Liu,
North China University of Technology, China

*CORRESPONDENCE
SuWei Zhai,
suwei_zhai@163.com

RECEIVED 15 December 2023
ACCEPTED 15 January 2024
PUBLISHED 26 March 2024

CITATION
Feng Y, Luo E, Lu H and Zhai S (2024), Cross-modality feature fusion for night pedestrian detection.
Front. Phys. 12:1356248.
doi: 10.3389/fphy.2024.1356248

Cross-modality feature fusion for night pedestrian detection

Yong Feng, Enbo Luo, Hai Lu and SuWei Zhai*

Electric Power Research Institute, Yunnan Power Grid Corporation, Kunming, China

Night pedestrian detection with visible image only suffers from the dilemma of high miss rate due to poor illumination conditions. Cross-modality fusion can ameliorate this dilemma by providing complementary information to each other through infrared and visible images. In this paper, we propose a cross-modal fusion framework based on YOLOv5, which is aimed at addressing the challenges of night pedestrian detection under low-light conditions. The framework employs a dual-stream architecture that processes visible images and infrared images separately. Through the Cross-Modal Feature Rectification Module (CMFRM), visible and infrared features are finely tuned on a granular level, leveraging their spatial correlations to focus on complementary information and substantially reduce uncertainty and noise from different modalities. Additionally, we have introduced a two-stage Feature Fusion Module (FFM), with the first stage introducing a cross-attention mechanism for cross-modal global reasoning, and the second stage using a mixed channel embedding to produce enhanced feature outputs. Moreover, our method involves multi-dimensional interaction, not only correcting feature maps in terms of channel and spatial dimensions but also applying cross-attention at the sequence processing level, which is critical for the effective generalization of cross-modal feature combinations. In summary, our research significantly enhances the accuracy and robustness of nighttime pedestrian detection, offering new perspectives and technical pathways for visual information processing in low-light environments.

KEYWORDS

pedestrian detection, YOLOv5, vision transformer, CNNs, feature fusion

1 Introduction

Pedestrians are a vital element in traffic scenarios, and the ability to detect pedestrians quickly and accurately has increasingly become a critical research topic in the field of computer vision. Pedestrian detection plays an essential role in various practical applications, such as autonomous driving perception systems [1–3] and intelligent security monitoring systems [4–6]. Additionally, pedestrian detection serves as the foundational task for downstream tasks like pedestrian tracking [7–9], action recognition and prediction [10–12], with its accuracy directly impacting the performance of these tasks. With the significant advancements in convolutional neural networks (CNNs), pedestrian detection models [13–16] have been continually updated and iterated, bringing forth models with outstanding performance. However, most pedestrian detection models are trained on single-modality, well-illuminated visible light datasets [17–19]. When faced with low-light conditions such as at night, their performance significantly declines due to excessive noise and decreased discriminability [4, 20]. Pedestrian detection using only nighttime visible light images is particularly challenging

because the data modality itself lacks a valid target area. Therefore, an increasing amount of research is focusing on cross-modality fusion learning, such as the fusion detection of visible and infrared images [21–26].

Infrared vision sensors operate on the principle of thermal imaging, distinguishing pedestrians from the background by differences in thermal radiation. Infrared imagery is robust against interference and is not easily affected by adverse environmental conditions [27, 28]. Even at night, infrared images can reveal the shape of pedestrians, effectively compensating for the vulnerability of visible light images to lighting conditions. However, infrared images also have drawbacks, such as lower resolution and a lack of texture information. On the other hand, visible light images provide rich detail and texture information [22]. Therefore, cross-modal fusion aims to extract complementary information between these two modalities, enhancing the flow of information between them and improving the perceptibility and robustness of detection algorithms. In the field of image fusion, a lot of work [29] has been carried out on the effective fusion of infrared images and visible light images.

In the field of pedestrian detection that fuses visible and infrared imaging, many approaches rely solely on Convolutional Neural Networks (CNN) to extract deep features [21, 23, 25, 26], with artificially designed complex fusion mechanisms to integrate features from different modalities. Extensive research has demonstrated the powerful representational capabilities of CNNs for expressing visual features in single-modality scenarios [30–32]. However, due to the limited receptive field, CNNs, while adept at capturing local information, exhibit weaker capabilities in capturing global texture information across modalities in fusion tasks. Transformer [33, 34] is equipped with self-attention mechanisms, possess a global receptive field and excel at learning long-range dependencies. Therefore, combining CNNs with transformers for cross-modality nighttime pedestrian detection can leverage the strengths of both, resulting in complementary advantages and enhanced detection performance.

Recently, vision transformers [33, 35–37] have been processing inputs as sequences and have demonstrated the capability to capture long-range correlations, offering a promising avenue towards a unified framework for multi-modal tasks. However, it remains to be clarified whether vision transformers can bring potential improvements to vis-inf pedestrian detection compared to existing multi-modal fusion modules [38–40] based on Convolutional Neural Networks (CNNs). Crucially, while some earlier studies have employed a simplistic global multi-modal interaction strategy, such an approach has not been universally applicable across various sensing data combinations [41–43]. We posit that in vis-inf pedestrian detection, which involves a variety of supplementary information and uncertainties, a comprehensive cross-modal interaction should be implemented to fully leverage the potential of cross-modal complementary features.

To address the challenges in vis-inf nighttime pedestrian detection, we propose an interactive cross-modal fusion framework based on yolov5, named FRFPD. This framework aims to enhance the performance of detection algorithms through efficient information fusion. FRFPD is constructed as a dual-stream architecture, specifically handling visible light (VIS) and infrared (Inf) data streams. On this foundation, we have designed feature interaction and fusion modules to optimize model performance: The Cross-Modal Feature Rectification Module (CMFRM) fine-tunes VIS and Inf features at a granular level, utilizing their spatial correlations to enhance the model's focus on complementary information and effectively reduce the uncertainty and noise from different modalities. This process precisely handles the complexity of multi-source data, paving the way for more effective feature extraction and interaction. Moreover, the Feature Fusion Module (FFM) [41] is structured in two stages, ensuring ample information exchange before feature fusion on a global scale. In the first stage, we introduce a cross-attention mechanism for cross-modal global reasoning, propelled by a wide receptive field facilitated by the self-attention mechanism. In the second stage, a mixed channel embedding is employed to generate enhanced feature outputs. In essence, the interaction strategy we introduce is multidimensional: within the CMFRM module, we correct feature maps on a spatial dimension; while in the FFM module, it apply a cross-modal attention mechanism for feature fusion across the global channel dimension. These approaches are vital for the effective generalization of cross-modal feature combinations, enhancing the model's capability to process information from diverse sensory modalities. Our contributions are summarized as follows:

(1) A dual-stream architecture is proposed in the FRFPD framework, leveraging YOLOv5, to handle visible light (VIS) and infrared (INF) data streams separately, tailored for addressing low-light challenges in nighttime pedestrian detection.
(2) The Cross-Modal Feature Rectification Module (CMFRM) is introduced to fine-tune visible and infrared features, exploiting their spatial correlations to enhance focus on complementary information, significantly reducing uncertainty and noise from different modalities. NF.
(3) An advanced Feature Fusion Module (FFM) developed in [41] is introduced, in two stages to promote ample information exchange and utilize a mixed channel embedding for generating enhanced feature outputs, improving detection capabilities.

2 Related works

The widespread application of Transformers in the field of Natural Language Processing (NLP) has proven their excellence and convenience in handling sequential data, which has also made them popular for visual tasks.

2.1 Vision transformer

The widespread application of Transformers in the field of Natural Language Processing (NLP) has proven their excellence and convenience in handling sequential data, which has also made them popular for visual tasks [35, 36, 44]; [45, 46]. ViT [35] addresses the high computational cost issue of Transformers in traditional visual tasks by flattening images into a series of pixel blocks (patches), transforming image processing tasks into a form similar to the word sequence processing in NLP. DeiT [47] further proposes a convolution-free Transformer structure, introducing a

teacher-student strategy through distillation tokens, with training conducted solely on ImageNet. Moreover, the positional encoding feature of Transformers is used to capture the order information of sequence data, which can be either fixed or learnable [48].

In the field of computer vision, Visual Transformer (VT) have demonstrated significant capabilities across various tasks such as image Fusion [49, 50]), pedestrian detection [51], particularly excelling in multispectral detection tasks [52–55] where they can focus on important features scattered across different spectral bands. Their self-attention mechanism's ability to model long-range dependencies and capture global context is especially valuable. Unlike convolutional neural networks [26, 56–58], VT operate on sequences of image patches (tokens) and are adept at learning to concentrate on the most informative parts of the input, making them inherently suited for multispectral detection where significant features may be sparsely distributed across spectral bands. However, the application of VT in multispectral detection, especially under challenging lighting conditions, remains a developing field. Our work is inspired by the intrinsic advantages of VT to tackle unique challenges in low-light multispectral scenarios. We have introduced a novel VT-based framework, specifically designed for this purpose, that incorporates modules sensitive to the nuances of multispectral data. Our proposed Cross-Modal Feature Rectification Module (CMFRM) expands the concept of VT by integrating cross-modal learning directly into the transformer architecture, serializing tokens along the spatial dimension, thereby enhancing the model's ability to perform fine-grained feature adjustment. This is critical for aligning features across different modalities, particularly when contending with varying levels of illumination and noise inherent in low-light conditions.

2.2 Multispectral pedestrian detection

The field of pedestrian detection has seen the emergence of numerous outstanding studies, including early traditional detection methods [59, 60] and the surge of CNN-based detection technologies [61–64] that came with the rapid development of Convolutional Neural Networks (CNN). However, the majority of research is still focused on single-modality visible light images. In nocturnal environments, relying solely on visible light images for pedestrian detection often fails to achieve satisfactory results, mainly because conventional visible light cameras perform poorly in night-time imaging, with target areas not being distinct and substantial noise interference. For this reason, it becomes extremely difficult for models like CNNs to extract effective features from nighttime visible light images. As research has deepened, infrared imagery, with its unique advantages in night-time settings, has started to be used to complement the shortcomings of visible light images. This has attracted increasing attention from researchers and has spurred the advancement and exploration of multispectral pedestrian detection technologies, especially those based on CNN approaches.

In the field of multispectral detection, fusion algorithms play a crucial role. The AR-CNN [65] model introduces an end-to-end region alignment algorithm, which addresses the subtle misalignments caused by positional offsets between multimodalities. This fusion approach reweights features to prioritize more reliable characteristics and suppress ineffective ones. Meanwhile, the CIAN [26] model leverages the interactive properties of multispectral input sources, proposing a cross-channel interactive attention network. This network extracts global features from each channel of the two modalities and recalibrates the channel responses of intermediate feature maps using an attention mechanism by computing the inter-channel correlation. In existing multispectral detection research, models like AR-CNN and CIAN offer solutions for minor misalignments between modalities and feature recalibration; however, these methods still show limitations in complex scenarios under low-light conditions, such as night-time pedestrian detection. These limitations manifest in two aspects: firstly, feature information loss due to insufficient lighting under low-light conditions cannot be compensated for by simply reweighting features; secondly, despite the CIAN model employing an interactive attention mechanism, more efficient strategies for information exchange and fusion are needed to handle the complex interactions between different modalities. CFT [66] proposed a fusion algorithm that combines transformer and CNN, which can learn remote dependencies and extract global context information. Self-attention can fuse features within and between modes. It is a relatively novel method recently, but this model uses traditional transformer, which has the problems of positional encoding and multi-head attention mismatch cross-modality fusion. ProbEn [67] research primarily focuses on the issue of multimodal object detection, with a particular emphasis on addressing the challenges of object detection in low-light conditions. It introduces the ProbEn probabilistic ensemble technique to effectively fuse object detection results from different sensors, thereby significantly enhancing the performance of multimodal object detection. UGC [68] is dedicated to addressing crucial challenges in multispectral pedestrian detection, encompassing issues such as image calibration and disparities between different modalities. The authors introduce a novel approach that aims to enhance pedestrian detection performance by incorporating Region of Interest (RoI) uncertainty and predictive uncertainty into the feature fusion and modality alignment processes.

To overcome these limitations, we propose the FRFPD framework, central to which are the Cross-Modal Feature Rectification Module (CMFRM) and the Feature Fusion Module (FFM). The CMFRM is motivated by the need to serialize tokens in the spatial dimension for fine-grained feature adjustment, aligning features within the visible and infrared modalities. Its design aims to finely tune features across modalities by exploiting their spatial correlations to amplify complementary information, thereby significantly reducing uncertainty and noise in low-light conditions. This approach is crucial for enhancing the accuracy and robustness of detection under varied lighting conditions. Concurrently, the FFM addresses the challenge of integrating diverse modalities effectively. It serializes tokens globally in the channel dimension, first performing global reasoning between modalities through a cross-attention mechanism, then refining the feature output with hybrid channel embedding. This strategy is driven by the need to provide not only an in-depth exchange of information but also a more nuanced enhancement of channel responses than the CIAN model. The motivation behind FFM is to improve the overall quality of feature fusion, enhancing the detection capabilities in complex scenarios. The FRFPD

10.3389/fphy.2024.1356248

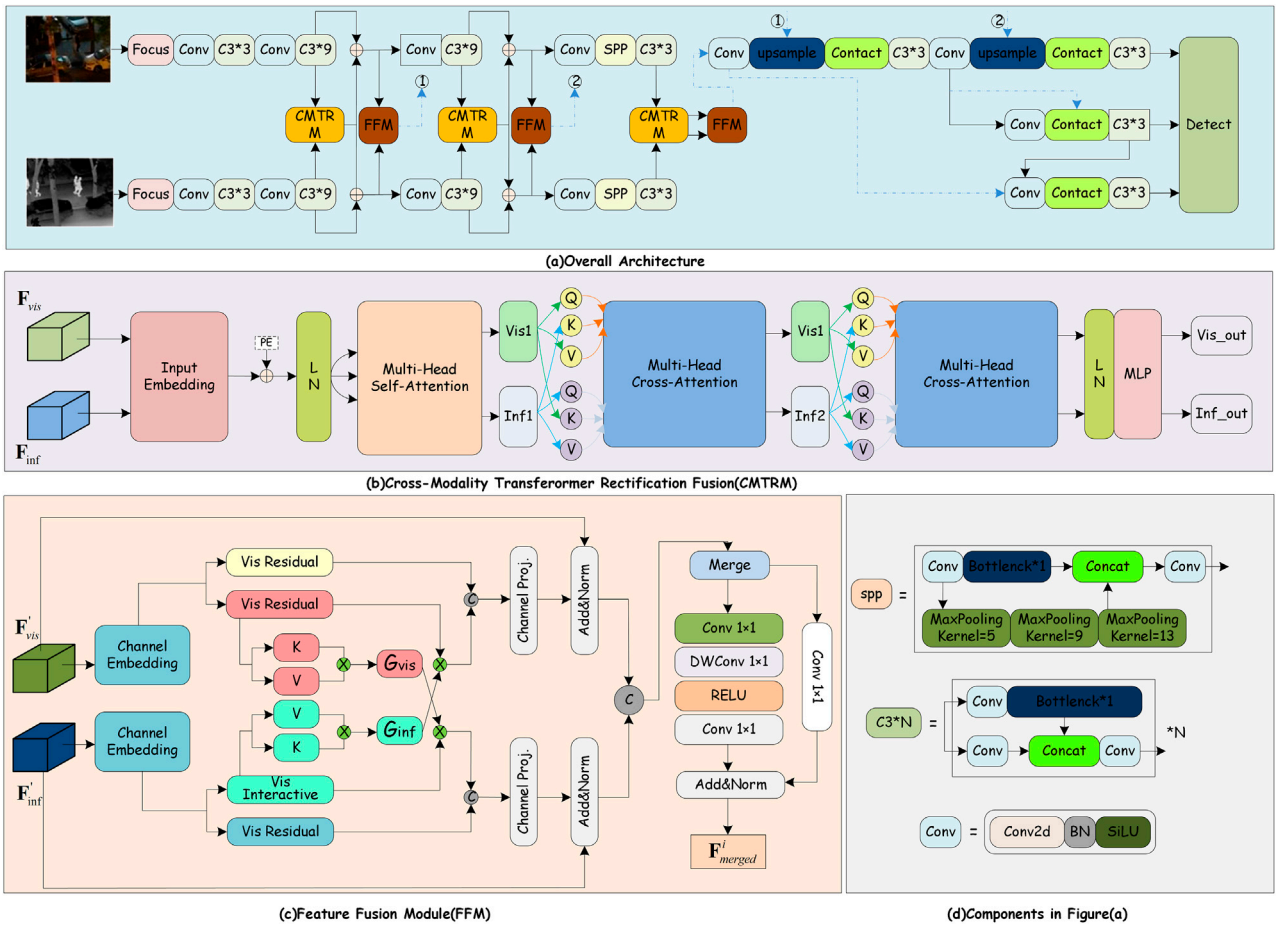

FIGURE 1
The Network structure of our proposals. **(A)** shows our overall network architecture, which adopts a novel combination of CNN and transformer. The deep features of visible and infrared images are extracted by two-stream CNN, and the proposed CMFRM module is used to leverag the features from one modality to rectify the features of the other modality. Feature Fusion Module (FFM) operates through a bifurcated process, as illustrated in **(C)** an initial stage of global information exchange followed by a stage of comprehensive global feature fusion. This structure is designed to facilitate extensive information interchange preceding the fusion of features at a global level. In addition, **(D)** shows the structure of the components in **(A)**.

framework sets a new performance benchmark for cross-modal feature fusion through its multi-dimensional interaction strategy, correcting feature maps on the channel and spatial dimensions, and implementing cross-attention at the sequence processing level.

3 Proposed method

3.1 Overview

Among the numerous target detection CNN models, YOLOv5 [69] is a highly reliable algorithm with fast recognition speed, which is easier to deploy and train. It is also one of the most popular detection frameworks currently and has a wide range of applications. Therefore, in this paper, we choose YOLOv5 to extract deep features and extend the transformer fusion algorithm to a dual-stream architecture. The backbone of YOLOv5 is modified from a single-stream structure to a dual-stream structure to separately extract deep features of the input visible light and infrared images. The rectification module, called Cross-Modal Feature Rectification Module (CMFRM), is implemented three times in the backbone. CMFRM is corrected one feature against another, and *vice versa*. In this way, the features of both modalities can be corrected. Additionally, as illustrated in Figure 1B, we introduced a Feature Fusion Module (FFM) [41] that merges features belonging to the same level into a single feature map. Then, a detection head is used to predict the final pedestrian positions. Our proposed network framework is illustrated in Figure 1.

3.2 Cross-modality feature rectification module

In this paper, we explore the complementarity of information from different sensors [8], [9], noting that while this information is valuable, it is often affected by noise. To address this issue, we introduce a novel Cross-Modal Feature Rectification Module (CMFRM) in Figure 1B, which is capable of performing precise feature correction at each stage of feature extraction on parallel data streams. Utilizing Transformer technology for spatial feature correction, the CMFRM provides a

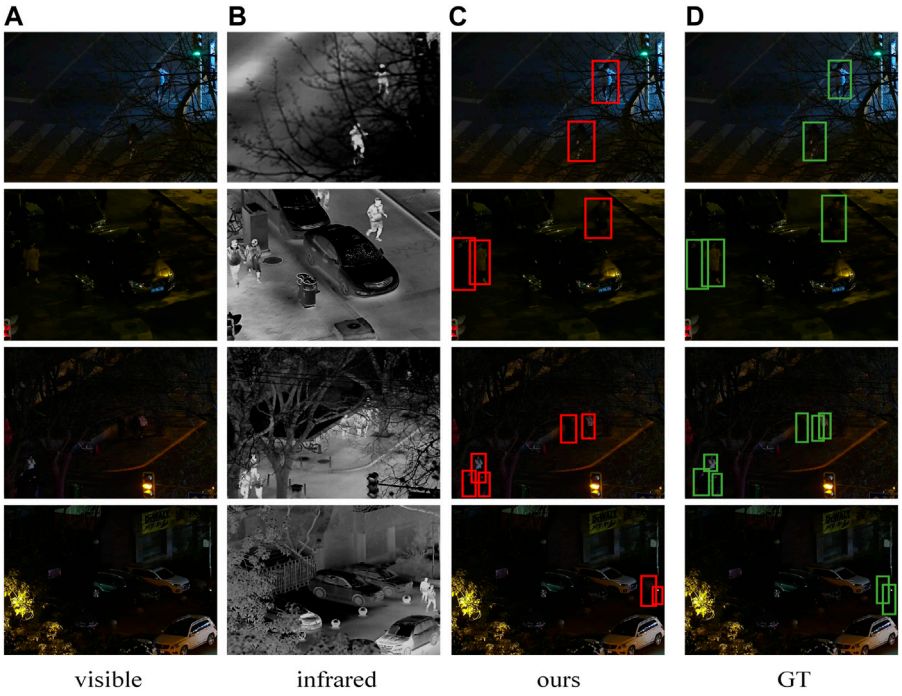

FIGURE 2
The visualization of the detection results, subfigure **(A)** shows the input visible lr images, subfigure**(B)** is the corresponding infrared images, subfigure **(C)** is the prediction result of our model, and subfigure **(D)** is the ground truth. These images are selected from the dataset listed at https://soonminhwang.github.io/rgbt-ped-detection/

granular correction mechanism. This not only effectively handles noise and uncertainty across different sensory modalities but also enhances the extraction and interaction of multimodal features, thereby improving the overall performance of the system.

In a two-stream structure, we extract features from visible and infrared images independently through Convolutional Neural Networks (CNN), obtaining visible feature and infrared feature, respectively. Both feature sets have the shape (B, C, H, W), where B is the batch size, C is the number of channels, and H and W are the dimensions of the spatial size. To adapt these features for the transformer, we flatten them into the shape (B, N, C), while proceeding along the spatial dimensions. where N is the number of tokens, given by $N = H \times W$. This step is a crucial phase in the transition of CNN features to transformer-based CMFRM module.

$$\text{flat}_{vis} = F_{vis} \cdot \text{view}(B, C, -1) \tag{1}$$

$$\text{flat}_{inf} = F_{inf} \cdot \text{view}(B, C, -1) \tag{2}$$

$$\text{flat}_{\text{cat}} = \text{concat}\left(\left(flat_{vis}, flat_{\text{inf}}\right), \text{dim} = 2\right) \tag{3}$$

$$Z = \text{flat}_{\text{cat}}.\text{permute}(0, 2, 1) \tag{4}$$

where F_{vis} and F_{inf} represent the visible and infrared features from the CNN, respectively. The `view` function reshapes the tensor of specified shape without changing its data, and `concat` concatenates the given tensors along the specified dimension. The `permute` function outputs a tensor after permuting the dimensions of the input tensor. Thus, in Eq 4, the shape of Z is $(B, 2N, C)$.

Positional embeddings enable the model to discern spatial relationships between different tokens during training. After positional embedding, the input sequence Z is then projected onto three weight matrices to compute a set of queries, keys, and values (Q, K, *and*V), expressed as $Q = ZW^Q$, $K = ZW^K$, $V = ZW^V$. In this context, the weight matrices are defined as $W^Q \in \mathbb{R}^{C\times D_Q}$, $W^K \in \mathbb{R}^{C\times D_K}$, and $W^V \in \mathbb{R}^{C\times D_V}$. Furthermore, the dimensions D_Q, D_K, and D_V are equivalent in our transformer model, such that $D_Q = D_K = D_V = C$. The Multi-head Self-Attention layer computes the attention weights by calculating the scaled dot products between Q and K. These weights are then applied to V to infer the refined output $\hat{Z}$.

$$\hat{Z} = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{D_K}}\right)V \tag{5}$$

However, multimodal data is distributed across different spatial domains, and relying solely on self-attention is insufficient for fully exploiting the mixed modality information, which may result in inadequate rectification. Based on the principle of self-attention, we speculate that exchanging the "values" and "keys" between different modalities might better enhance the vital information and facilitate the flow of complementary information. Building on these considerations, we have extended the traditional multi-head attention based on a cascading strategy by incorporating two instances of Cross-Attention (CA), as shown in Figure 1B. Additionally, the process of information exchange during the two instances of Cross-Attention can be represented by Eqs. 6–9.

$$CA^1_{vis}\left(Q_{vis}, K_{inf}, V_{vis}\right) = \text{softmax}\left(\frac{Q_{vis}K^T_{inf}}{\sqrt{d_k}}\right)V_{vis} \tag{6}$$

$$CA^1_{inf}\left(Q_{inf}, K_{vis}, V_{inf}\right) = \text{softmax}\left(\frac{Q_{inf}K^T_{vis}}{\sqrt{d_k}}\right)V_{inf} \tag{7}$$

and

$$CA^2_{vis}\left(Q_{vis}, K_{vis}, V_{inf}\right) = \text{softmax}\left(\frac{Q_{vis}K^T_{vis}}{\sqrt{d_k}}\right)V_{inf} \quad (8)$$

$$CA^2_{inf}\left(Q_{inf}, K_{inf}, V_{vis}\right) = \text{softmax}\left(\frac{Q_{inf}K^T_{inf}}{\sqrt{d_k}}\right)V_{vis} \quad (9)$$

where vis, inf represent visible token and infrared token from $\hat{Z}$ respectively. After processing through two cascaded multi-head cross-attention layers, the visible and infrared features are subjected to Layer Normalization (LN) and Multi-Layer Perceptron (MLP), ultimately producing two output features, $\tilde{\mathbf{F}}_{vis}$ and $\tilde{\mathbf{F}}_{inf}$.

3.3 Two-stage feature fusion module

After obtaining the feature mappings from each layer, a two-stage feature fusion module (Feature Fusion Module, FFM) [41] is introduced to enhance the interaction and integration of global information. As illustrated in Figure 1C, in the first stage, the two branches are kept separate, and a cross-attention mechanism is designed to facilitate the global exchange of information between the two branches. In the stage 2, the concatenated features are transformed back to the original scale through a mixed channel embedding.

Global Information exchange stage. We first flatten the input feature of size $\tilde{\mathbf{F}}_{vis}$ and $\tilde{\mathbf{F}}_{inf} \in \mathbb{R}^{H\times W\times C}$ into $R^{N\times C}$ along with channel dimension, where $N = H \times W$, and C is the number of tokens, Then, through linear embedding, we generate two vectors of the same size $R^{N\times C}$, named the residual vector X_{res} and the interactive vector X_{inter}. Building upon this, we propose an efficient cross-attention mechanism that applies to these two interactive vectors from different modal pathways, achieving comprehensive information exchange across modalities. This mechanism offers complementary interactions from a sequence-to-sequence perspective, surpassing the rectification-based interactions from the feature map perspective in CMFRM.

Our cross-attention mechanism, designed for improved cross-modal feature fusion, is an adaptation of the conventional self-attention mechanism [33]. The traditional method encodes inputs into Queries (Q), Keys (K), and Values (V), computing a global attention map via QK^T. This results in a computationally expensive $N \times N$ matrix. Alternatively [70], proposes using a global context vector $G = K^TV$, reducing the size to $C_{\text{head}} \times C_{\text{head}}$. Our approach builds on this by embedding interactive vectors into K and V for each head, with both matrices sized $N \times C_{\text{head}}$. The final output is a product of these interactive vectors and the context vector from an alternate modality, constituting the cross-attention process.

$$\begin{aligned} G_{vis} &= \hat{K}^T_{vis}\hat{V}_{vis} \\ G_{inf} &= \hat{K}^T_{inf}\hat{V}_{inf} \end{aligned} \quad (10)$$

$$\begin{aligned} U_{vis} &= X^{\text{inter}}_{vis}\text{Softmax}\left(G_{inf}\right) \\ U_{inf} &= X^{\text{inter}}_{inf}\text{Softmax}\left(G_{vis}\right) \end{aligned} \quad (11)$$

Note that G denotes the global context vector, while U indicates the attended result vector. To realize attention across different representational subspaces, we maintain the multi-head mechanism, where the number of heads corresponds to the number of elements in the transformer backbone. Subsequently, the attended result vector U and the residual vector are concatenated. Finally, we apply a second linear embedding and resize the feature back to $R^{H\times W\times C}$.

Global Feature Fusion Module. In the fusion component of the Feature Fusion Module (FFM), channel-wise integration is performed using 1×1 convolution for combining features from dual pathways. Considering the necessity of spatial context for Vis-Inf pedestrain detection, we adopt a strategy influenced by Mix-FFN [71] and ConvMLP [72], incorporating a depth-wise 3×3 convolution (DW Conv) to form a skip connection architecture. This approach facilitates the consolidation of the concatenated feature dimensions $R^{H\times W\times 2C}$ into the decoder output dimension $R^{H\times W\times C}$.

4 Experiments

In this section, we first introduce two multispectral datasets, KAIST [73] and LLVIP [22]. The KAIST dataset compiles data from day and night autonomous driving scenarios, while the LLVIP dataset is composed of night-time surveillance scenarios. Given our focus on nighttime pedestrian detection, we exclusively selected the nighttime subset of the KAIST dataset. Subsequently, we delve into some specifics of the model training phase. The evaluation metrics for pedestrian detection diverge slightly from those of traditional object detection, hence we will clarify the evaluation metrics utilized in this study. We benchmark our results against state-of-the-art methods and conduct ablation studies to assess the effectiveness of our proposed module. Lastly, the visualization of our proposals is provided to facilitate an intuitive understanding of their impact. At last, we provide a visualization of the predicted results as shown in Figure 2.

4.1 Dataset

KAIST. The KAIST dataset [73], introduced at CVPR2015, consists of 95k aligned pairs of visible and infrared images and has been extensively utilized. All annotations are manually labeled, including 1,182 pedestrian instances. Due to biased annotations in the original training set, this study employs the sanitized version [23]. The sanitized KAIST provides 7,601 training images with at least one valid pedestrian instance, filtered and sampled from the original training videos. There are 2,846 pairs for night training and 4,755 pairs for day training. The test set comprises 2,252 image pairs, with 797 for night and 1,455 for day. Test annotations from the improved version [31], which corrects the initial annotations, are used. The resolution of training and test images is 640×512.

LLVIP. LLVIP [22] is a nighttime pedestrian dataset for surveillance scenarios, presented at ICCV2021. It includes 15,488 strictly aligned visible-infrared image pairs, featuring numerous pedestrians and cyclists from diverse street locations between 6 and 10 p.m. [22]. The original resolution of the images is 1280×1024, but to reduce computational demands, we scale down the images by half to 640×512 in this paper.

TABLE 1 Results on KAIST night dataset and the results in bold indicate the optimal.

Methods	Data modality	LAMR (%)	AP50
Yolov5 [69]	Visible	63.65	43.95%
Yolov5 [69]	Infrared	14.73	77.51%
MLF-CNN [74]	Visible + Infrared	25.65	67.60%
IATDNN [75]	Visible + Infrared	26.88	67.02%
CWF-CNN [76]	Visible + Infrared	30.82	64.59%
L-SSD [77]	Visible + Infrared	35.38	48.77%
MSDS-RCNN [23]	Visible + Infrared	13.73	-
CS-RCNN [78]	Visible + Infrared	11.86	-
CIAN [26]	Visible + Infrared	11.13	-
MBNet [79]	Visible + Infrared	10.98	-
UGC [68]	Visible + Infrared	10.92	-
ProbEn [67]	Visible + Infrared	10.83	-
Our Method	Visible + Infrared	**10.79**	**82.48%**

4.2 Evaluation

Evaluation metrics. The first assessment metric is the Log-Average Miss Rate (LAMR), which is a specialized metric for evaluating the performance of pedestrian detection systems. The relationship between the Miss Rate (MR) and the False Positives Per Image (FPPI) is plotted on a log-log scale, and nine FPPI reference points are selected within the range $[10^{-2}, 10^{0}]$, evenly spaced in the logarithmic space. LAMR is defined as shown in Eq 14.

$$MR = \frac{FN}{TP + FN} \tag{12}$$

$$FPPI = \frac{FP}{imgs\,\mathrm{num}} \tag{13}$$

$$LAMR = \exp\left(\frac{1}{9}\sum_{f}\log\left(MR\underset{FPPI\le f}{\operatorname{argmax}}FPPI\right)\right) \tag{14}$$

where f is within the set $\{10^{-2}, 10^{-1.75}, \ldots, 10^{0}\}$, TP represents the number of True Positives, FP is the number of False Positives, and FN denotes the number of False Negatives. Additionally, we utilize AP50 as our second metric, complementing LAMR. In the evaluation process, all detected bounding boxes are matched to ground truth annotations for each image via a greedy algorithm. If the Intersection over Union (IoU) between the detection box and the ground truth exceeds a specified threshold, the detection is considered a True Positive (TP), indicating a successful prediction. Due to the highly non-rigid nature of pedestrians, we adopt the common IoU threshold of 0.5. Thus, AP50 denotes the Average Precision when the IoU threshold is 0.5.

4.3 Comparison of results on KAIST night dataset

We compared our model with the results of state-of-the-art models on the KAIST Night test set, as presented in Table 1. Our model builds upon a two-stream architecture extended from yolov5; hence, we assessed the single-modality detection capabilities of yolov5 with only visible and only infrared images on the same dataset. The task of night-time pedestrian detection using solely visible light images poses a substantial challenge, reflected in a high LAMR of 63.65%. Through the development of effective cross-modality fusion algorithms, such as MSDS-RCNN [23] and CFT [66], the LAMR for night-time pedestrian detection can be significantly decreased, improving detector performance. Furthermore, our proposed method records a LAMR of 10.79% and an AP50 of 82.48%, evidencing the effectiveness and competitive edge of our approach.

4.4 Ablation study

From the previous sections, we have familiarized ourselves with the architecture and proposed modules such as CMFRM, as well as the enhancements in our method. However, the exact quantitative improvements contributed by these modules remain uncertain. Therefore, in this section, we present a succinct and insightful ablation study to address the aforementioned inquiries. Table 2 illustrates that CMFRM has led to a decrease of 1.14% in LAMR and an enhancement of 1.47% in AP50 on the KAIST Night dataset, and a reduction of 0.63% in LAMR on the LLVIP dataset. FFM contributes to a decrease of 0.57% in LAMR and an improvement of 1.18% in AP50 on the KAIST Night dataset, and a reduction of 0.80% in LAMR on the LLVIP dataset. Finally, when compared to the baseline model CFT [66], our comprehensive model CMTF decreases LAMR by 1.38% and enhances AP50 by 3.2% on the KAIST Night dataset, and lowers LAMR by 1.62% on the LLVIP dataset.

TABLE 2 Results of ablation study and the results in bold indicate the optimal.

Method			KAIST night		LLVIP	
Base	CMFRM	FFM	LAMR (%)	AP50 (%)	LAMR (%)	AP50 (%)
✓			12.71	79.28	5.40	97.50
✓	✓		11.57	80.75	4.77	97.72
✓		✓	12.14	80.46	4.60	97.09
✓	✓	✓	**10.79**	**82.48**	**3.78**	**97.98**

4.5 Conclusion

In this paper, we introduce an interactive cross-modal fusion framework based on YOLOv5, designed to improve the performance of nighttime pedestrian detection algorithms through efficient information fusion. Our framework utilizes a dual-stream architecture to separately handle visible and infrared images, effectively addressing the challenges posed by low-light conditions. Our proposed FRFPD significantly enhance model performance by fine-tuning features across modalities, reducing uncertainty and noise, and focusing on complementary information. These modules also facilitate multi-dimensional feature interaction and rectification, including cross-attention mechanisms at the sequence processing level, which are crucial for the effective generalization of cross-modal feature combinations. Overall, our research not only boosts the performance of nighttime pedestrian detection but also offers new technical solutions and perspectives for visual information processing under low-light conditions.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://multispectral.kaist.ac.kr/pedestrian/data-kaist.

Author contributions

YF: Methodology, Writing–original draft. EL: Writing–review and editing. HL: Writing–review and editing. SZ: Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

Authors YF, EL, HL, and SZ were employed by Yunnan Power Grid Corporation.

Publisher's note

References

1. Chen L, Lin S, Lu X, Cao D, Wu H, Guo C, et al. Deep neural network based vehicle and pedestrian detection for autonomous driving: a survey. *IEEE Trans Intell Transportation Syst* (2021) 22:3234–46. doi:10.1109/tits.2020.2993926

2. Chen Z, Huang X. Pedestrian detection for autonomous vehicle using multi-spectral cameras. *IEEE Trans Intell Vehicles* (2019) 4:211–9. doi:10.1109/tiv.2019.2904389

3. Hbaieb A, Rezgui J, Chaari L. Pedestrian detection for autonomous driving within cooperative communication system. In: *2019 IEEE wireless communications and networking conference (WCNC)*. IEEE (2019). p. 1–6.

4. Wang X, Chen J, Wang Z, Liu W, Satoh S, Liang C, et al. When pedestrian detection meets nighttime surveillance: a new benchmark. *International Joint Conference on Artificial Intelligence* (2020) 20000:509–515. doi:10.24963/ijcai.2020/71

5. Kulbacki M, Segen J, Wojciechowski S, Wereszczyński K, Nowacki JP, Drabik A, et al. Intelligent video monitoring system with the functionality of online recognition of people?s behavior and interactions between people. In: *Intelligent information and database systems: 10th asian conference, ACIIDS 2018, dong hoi city, vietnam, march 19-21, 2018, proceedings, Part II 10*. Springer (2018). p. 492–501.

6. Rai M, Husain AA, Maity T, Yadav RK, Neves A. Advance intelligent video surveillance system (aivss): a future aspect. *Intell Video Surveill* (2019) 37. doi:10.5772/intechopen.76444

7. Huang L, Zhao X, Huang K. Bridging the gap between detection and tracking: a unified approach. *Proc IEEE/CVF Int Conf Comput Vis* (2019) 3999–4009. doi:10.1109/ICCV.2019.00410

8. Sun Z, Chen J, Chao L, Ruan W, Mukherjee M. A survey of multiple pedestrian tracking based on tracking-by-detection framework. *IEEE Trans Circuits Syst Video Technol* (2020) 31:1819–33. doi:10.1109/tcsvt.2020.3009717

9. Stadler D, Beyerer J. Improving multiple pedestrian tracking by track management and occlusion handling. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021). p. 10958–67.

10. Zhang P, Lan C, Zeng W, Xing J, Xue J, Zheng N. Semantics-guided neural networks for efficient skeleton-based human action recognition. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020).

11. Liu K, Liu W, Ma H, Tan M, Gan C. A real-time action representation with temporal encoding and deep compression. *IEEE Trans Circuits Syst Video Technol* (2020) 31:647–60. doi:10.1109/tcsvt.2020.2984569

12. Kong Y, Fu Y. Human action recognition and prediction: a survey. *Int J Comput Vis* (2022) 130:1366–401. doi:10.1007/s11263-022-01594-9

13. Huang X, Ge Z, Jie Z, Yoshie O. Nms by representative region: towards crowded pedestrian detection by proposal pairing. *Proc IEEE/CVF Conf Comput Vis Pattern Recognition* (2020) 10750–9. doi:10.1109/CVPR42600.2020.01076

14. Ouyang W, Zeng X, Wang X. Modeling mutual visibility relationship in pedestrian detection. *Proc IEEE Conf Comput Vis pattern recognition* (2013) 3222–9. doi:10.1109/CVPR.2013.414

15. Tian Y, Luo P, Wang X, Tang X. Pedestrian detection aided by deep learning semantic tasks. *Proc IEEE Conf Comput Vis pattern recognition* (2015) 5079–87. doi:10.1109/CVPR.2015.7299143

16. Xu D, Ouyang W, Ricci E, Wang X, Sebe N. Learning cross-modal deep representations for robust pedestrian detection. *Proc IEEE Conf Comput Vis pattern recognition* (2017) 5363–71. doi:10.1109/CVPR.2017.451

17. Braun M, Krebs S, Flohr F, Gavrila DM. Eurocity persons: a novel benchmark for person detection in traffic scenes. *IEEE Trans pattern Anal machine intelligence* (2019) 41:1844–61. doi:10.1109/tpami.2019.2897684

18. Dollar P, Wojek C, Schiele B, Perona P. Pedestrian detection: an evaluation of the state of the art. *IEEE Trans pattern Anal machine intelligence* (2011) 34:743–61. doi:10.1109/tpami.2011.155

19. Zhang S, Benenson R, Schiele B. Citypersons: a diverse dataset for pedestrian detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017). p. 3213–21.

20. Li G, Zhang S, Yang J. Nighttime pedestrian detection based on feature attention and transformation. In: *2020 25th international conference on pattern recognition (ICPR)*. IEEE (2021). p. 9180–7.

21. Chen YT, Shi J, Mertz C, Kong S, Ramanan D. *Multimodal object detection via bayesian fusion* (2021). *arXiv preprint arXiv:2104.02904*.

22. Jia X, Zhu C, Li M, Tang W, Zhou W. Llvip: a visible-infrared paired dataset for low-light vision. *Proc IEEE/CVF Int Conf Comput Vis* (2021) 3496–504. doi:10.1109/ICCVW54120.2021.00389

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

24. Liu J, Zhang S, Wang S, Metaxas D. Multispectral deep neural networks for pedestrian detection. (2016) *arXiv preprint arXiv:1611.02644*.

25. Zhang H, Fromont E, Lefèvre S, Avignon B. Guided attentive feature fusion for multispectral pedestrian detection. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (2021). p. 72–80.

26. 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

27. Zhao B, Wang C, Fu Q. Multi-scale pedestrian detection in infrared images with salient background-awareness. *J Electron Inf Technol* (2020) 42:2524–32. doi:10.11999/JEIT190761

28. Li H, Yang M, Yu Z. Joint image fusion and super-resolution for enhanced visualization via semi-coupled discriminative dictionary learning and advantage embedding. *Neurocomputing* (2021) 422:62–84. doi:10.1016/j.neucom.2020.09.024

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

30. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proc IEEE Conf Comput Vis pattern recognition* (2016) 770–8. doi:10.1109/CVPR.2016.90

31. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. Ssd: single shot multibox detector. In: *Computer vision–ECCV 2016: 14th European conference*, Amsterdam, The Netherlands (Springer) (2016), 21–37.

32. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. *Proc IEEE Conf Comput Vis pattern recognition* (2016) 779–88. doi:10.1109/CVPR.2016.91

33. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst* (2017) 30.

34. Liu Y, Wang L, Li H, Chen X. Multi-focus image fusion with deep residual learning and focus property detection. *Inf Fusion* (2022) 86-87:1–16. doi:10.1016/j.inffus.2022.06.001

35. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. *An image is worth 16x16 words: transformers for image recognition at scale* (2020). *arXiv preprint arXiv:2010.11929*.

36. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers and distillation through attention. *Int Conf machine Learn* (2021) 10347–57.

37. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. *Proc IEEE/CVF Int Conf Comput Vis* (2021) 10012–22. doi:10.1109/ICCV48922.2021.00986

38. Hu X, Yang K, Fei L, Wang K. Acnet: attention based network to exploit complementary features for rgbd semantic segmentation. In: *2019 IEEE international conference on image processing (ICIP)* (IEEE (2019), 1440–4.

39. Xiang K, Yang K, Wang K. Polarization-driven semantic segmentation via efficient attention-bridged fusion. *Opt Express* (2021) 29:4802–20. doi:10.1364/oe.416130

40. Deng F, Feng H, Liang M, Wang H, Yang Y, Gao Y, et al. Feanet: feature-enhanced attention network for rgb-thermal real-time semantic segmentation. In: *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE (2021). p. 4467–73.

41. Zhang J, Liu H, Yang K, Hu X, Liu R, Stiefelhagen R. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers. IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 12, pp. 14679–14694, Dec. 2023. doi:10.1109/TITS.2023.3300537

42. Li H, Yu Z, Mao C. Fractional differential and variational method for image fusion and super-resolution. *Neurocomputing* (2016) 171:138–48. doi:10.1016/j.neucom.2015.06.035

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

44. Touvron H, Cord M, Sablayrolles A, Synnaeve G, Jégou H. Going deeper with image transformers. *Proc IEEE/CVF Int Conf Comput Vis* (2021) 32–42. doi:10.1109/ICCV48922.2021.00010

45. Zhou D, Liu Z, Wang J, Wang L, Hu T, Ding E, et al. Human-object interaction detection via disentangled transformer. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022). p. 19568–77.

46. Xia Z, Pan X, Song S, Li LE, Huang G. Vision transformer with deformable attention. *Proc IEEE/CVF Conf Comput Vis pattern recognition* (2022) 4794–803. doi:10.1109/CVPR52688.2022.00475

47. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. *Training data-efficient image transformers and distillation through attention* (2012). arxiv. 10.48550.

48. Shaw P, Uszkoreit J, Vaswani A. *Self-attention with relative position representations* (2018). *arXiv preprint arXiv:1803.02155*.

49. Li H, Liu J, Zhang Y, Liu Y. A deep learning framework for infrared and visible image fusion without strict registration. *Int J Comput Vis* (2023). doi:10.1007/s11263-023-01948-x

50. Li H, Zhao J, Li J, Yu Z, Lu G. Feature dynamic alignment and refinement for infrared?visible image fusion: translation robust fusion. *Inf Fusion* (2023) 95:26–41. doi:10.1016/j.inffus.2023.02.011

51. Yang Y, Xu K, Wang K. Cascaded information enhancement and cross-modal attention feature fusion for multispectral pedestrian detection. *Front Phys* (2023) 11:1–11. doi:10.3389/fphy.2023.1121311

52. Choi Y, Kim N, Hwang S, Kweon IS. Thermal image enhancement using convolutional neural network. In: *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE (2016). p. 223–30.

53. Choi Y, Kim N, Hwang S, Park K, Yoon JS, An K, et al. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Trans Intell Transportation Syst* (2018) 19:934–48. doi:10.1109/tits.2018.2791533

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

55. Kim N, Choi Y, Hwang S, Kweon IS. Multispectral transfer network: unsupervised depth estimation for all-day vision. *Proc AAAI Conf Artif Intelligence* (2018) 32. doi:10.1609/aaai.v32i1.12297

56. Guan D, Cao Y, Yang J, Cao Y, Tisse CL. Exploiting fusion architectures for multispectral pedestrian detection and segmentation. *Appl Opt* (2018) 57:D108–D116. doi:10.1364/ao.57.00d108

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

58. Wagner J, Fischer V, Herman M, Behnke S Multispectral pedestrian detection using deep fusion convolutional neural networks. *ESANN* (2016) 587:509–14.

59. Dollár P, Appel R, Belongie S, Perona P. Fast feature pyramids for object detection. *IEEE Trans pattern Anal machine intelligence* (2014) 36:1532–45. doi:10.1109/tpami.2014.2300479

60. Zhang S, Benenson R, Schiele B Filtered channel features for pedestrian detection. *CVPR* (2015) 1751–1760. doi:10.1109/CVPR.2015.7298784

61. Brazil G, Yin X, Liu X. Illuminating pedestrians via simultaneous detection and segmentation. *Proc IEEE Int Conf Comput Vis* (2017) 4950–9. doi:10.1109/ICCV.2017.530

62. Mao J, Xiao T, Jiang Y, Cao Z. What can help pedestrian detection? *Proc IEEE Conf Comput Vis pattern recognition* (2017) 3127–36. doi:10.1109/CVPR.2017.639

63. Wang X, Xiao T, Jiang Y, Shao S, Sun J, Shen C. Repulsion loss: detecting pedestrians in a crowd. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018). p. 7774–83.

64. Zhang S, Wen L, Bian X, Lei Z, Li SZ. Occlusion-aware r-cnn: detecting pedestrians in a crowd. *Proc Eur Conf Comput Vis (Eccv)* (2018) 637–53. doi:10.1007/978-3-030-01219-9_39

65. Zhang L, Zhu X, Chen X, Yang X, Lei Z, Liu Z. Weakly aligned cross-modal learning for multispectral pedestrian detection. *Proc IEEE/CVF Int Conf Comput Vis* (2019) 5127–37. doi:10.1109/ICCV.2019.00523

66. Qingyun F, Dapeng H, Zhaokui W. *Cross-modality fusion transformer for multispectral object detection* (2021). *arXiv preprint arXiv:2111.00273*.

67. Chen YT, Shi J, Ye Z, Mertz C, Ramanan D, Kong S. Multimodal object detection via probabilistic ensembling. *Eur Conf Comput Vis* (2022) 139–58. doi:10.1007/978-3-031-20077-9_9

68. Kim JU, Park S, Ro YM. Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. *IEEE Trans Circuits Syst Video Technol* (2021) 32:1510–23. doi:10.1109/tcsvt.2021.3076466

69. Jocher G, Stoken A, Borovec J, Changyu L, Hogan A, Diaconu L, et al. *ultralytics/yolov5: v3. 0*. Zenodo (2020).

70. Shen Z, Zhang M, Zhao H, Yi S, Li H. Efficient attention: attention with linear complexities. *Proc IEEE/CVF Winter Conf Appl Comput Vis* (2021) 3531–9. doi:10.1109/WACV48630.2021.00357

71. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. Segformer: simple and efficient design for semantic segmentation with transformers. *Adv Neural Inf Process Syst* (2021) 34:12077–90. doi:10.48550/arXiv.2105.15203

72. Li J, Hassani A, Walton S, Shi H. Convmlp: hierarchical convolutional mlps for vision. *Proc IEEE/CVF Conf Comput Vis Pattern Recognition* (2023) 6306–15. doi:10.1109/CVPRW59228.2023.00671

73. 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* (2015). p. 1037–45.

74. Chen Y, Xie H, Shin H. Multi-layer fusion techniques using a cnn for multispectral pedestrian detection. *IET Comput Vis* (2018) 12:1179–87. doi:10.1049/iet-cvi.2018.5315

75. 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

76. Park K, Kim S, Sohn K. Unified multi-spectral pedestrian detection based on probabilistic fusion networks. *Pattern Recognition* (2018) 80:143–55. doi:10.1016/j.patcog.2018.03.007

77. Zhuang Y, Pu Z, Hu J, Wang Y. Illumination and temperature-aware multispectral networks for edge-computing-enabled pedestrian detection. *IEEE Trans Netw Sci Eng* (2021) 9:1282–95. doi:10.1109/tnse.2021.3139335

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

79. Zhou K, Chen L, Cao X. Improving multispectral pedestrian detection by addressing modality imbalance problems. In: *Computer vision–ECCV 2020: 16th European conference*. Glasgow, UK (2020). Proceedings, Part XVIII 16 (2020).