scientific reports

Check for updates

OPEN

Multimodal fusion transformer network for multispectral pedestrian detection in low-light condition

Gong Li[1], Guoyin Ren[1]✉, Jingyu Wang[1], Mobing Zhi[3], Zhijie Yu[1], Bo Jiang[1], Haoliang Guan[2] & Qidan Guo[1]

Multispectral pedestrian detection has attracted significant attention owing to its advantages, such as providing rich information, adapting to various scenes, enhancing features, and diversifying applications. However, most existing fusion methods are based on convolutional neural network (CNN) feature fusion. Although CNNs perform well in image processing tasks, they have limitations in handling long-range dependencies and global information. This limitation is addressed by Transformers through their self-attention mechanism, which effectively captures global dependencies in sequential data and excels in processing such data. We propose a Multimodal Fusion Transformer (MFT) module to effectively capture and merge features. This module utilizes the Transformer's self-attention mechanism to capture long-term spatial dependencies of intra- and inter-spectral images, enabling effective intra- and inter-modal fusion to improve performance in downstream tasks, such as pedestrian detection. Additionally, the Dual-modal Feature Fusion (DMFF) module is introduced to more effectively capture between RGB and IR modalities on a broader scale. To assess the network's effectiveness and generalization, various backbones were developed for experimentation, yielding impressive results. Additionally, extensive ablation studies were performed, varying the positions and quantities of fusion modules to determine the optimal fusion performance.

Keywords Multispectral pedestrian detection, Feature fusion, Cross-modality

In recent years, pedestrian detection has become a prominent research area in computer vision, with extensive applications in video surveillance[1–3], autonomous driving[4–6], and UAV small object detection. Traditional methods primarily depend on visible light images, but their effectiveness is limited by factors such as lighting conditions[7], complex backgrounds, and occlusions[8], which can significantly compromise detection accuracy. To overcome these challenges, researchers are increasingly exploring the fusion of infrared and visible light images to improve pedestrian detection performance[9–11].

Infrared and visible light imaging techniques provide complementary information that enhances pedestrian detection. As illustrated in Fig. 1, infrared images capture thermal radiation emitted by objects, facilitating reliable detection of pedestrians in low-light or nighttime environments. Conversely, visible light images offer rich texture information, aiding in the differentiation of pedestrian features. By integrating the advantages of these two modalities, a more robust and accurate pedestrian detection system can be developed.

Existing image fusion methods can primarily be categorized into traditional techniques and deep learning-based approaches. Traditional algorithms typically carry out feature extraction in either the spatial or transform domain while relying on manually designed fusion rules. Classical frameworks encompass a variety of techniques, including multi-scale transforms, sparse representations, subspace methods, saliency-based approaches, and variational models. Although these methods can achieve satisfactory results in many cases, they still have some problems. Firstly, they tend to use the same transformations or representations to extract features from the source images and fail to take into account the essential differences between the source images. Second, manually designed fusion rules and activity level measurements perform poorly in complex fusion scenarios with gradually increasing design complexity.

[1]School of Digital And Intelligent Industry, Inner Mongolia University of Science and Technology, BaoTou 014010, China. [2]Baotou Vocational abd Technical College, Baotou, China. [3]Weaver Changcheng Technology Co, Baotou, China. ✉email: renguoyin@imust.edu.cn

Fig. 1. The first column is a RGB image and the second column is an IR image. The first row captures images taken at night, where the infrared image distinctly highlights the positions of pedestrians. In contrast, the second row showcases daytime images, in which the visible light photograph distinctly highlights the background details.

In contrast, deep learning-based image fusion[12] methods primarily address three key challenges: feature extraction, feature fusion, and image reconstruction. Based on their network architecture, these methods can be categorized into three groups: self-encoder-based methods, self-convolutional neural networks, and generative adversarial networks.

Auto-Encoder (AE) framework: Initially, the auto-encoder is pre-trained on large datasets for the purposes of feature extraction and image reconstruction., after which deep features are integrated using manually designed fusion strategies. Notable examples of such datasets include MS-COCO (Microsoft Common Objects in Context) and ImageNet[13]. However, these manual strategies may not be applicable to deep features, thus limiting performance.

Convolutional Neural Network (CNN) framework: This approach achieves end-to-end feature extraction and image reconstruction by designing the network architecture and loss function, thereby eliminating the need for tedious manual design[14]. A popular CNN-based image fusion framework constructs a loss function by assessing the similarity between the fused image and the source image, guiding the network for end-to-end training[15]. Many mainstream methods focus on building the loss function based on this similarity measurement[16]. Additionally, some CNN-based approaches utilize convolutional networks for feature extraction or activity level assessment as components of the overall method[17].

Generative Adversarial Network (GAN) framework: image fusion is regarded as an adversarial process between a generator and a discriminator. GANs constrain the generator using the discriminator to ensure that the generated fusion result aligns with the object distribution, thereby facilitating feature extraction and image reconstruction. Current GAN-based fusion methods establish object distributions from source images[18] or pseudo-labeled images[19].

Despite numerous studies on multispectral pedestrian detection, effectively fusing visible and thermal images to enhance feature consistency continues to pose challenges. Visible images can capture valuable features, such as skin tone and hair, which thermal images do not provide. To tackle this issue, it is essential to develop a method that fully leverages the features from visible light while incorporating information from thermal images, thus enhancing the accuracy of pedestrian detection.

Many current methods predominantly use convolutional layers to enhance modality-specific features; however, the restricted receptive field of these layers hampers their ability to capture long-range spatial dependencies. In contrast, Transformers excel at processing sequence data, allowing them to effectively capture long-range dependencies. This capability allows for improved integration of information from different sensors and enhances the representation of fused features by treating the feature representations of infrared and visible images as sequences. Transformer, as a general-purpose sequence modelling module, is able to flexibly handle inter-modal feature representations between different modalities to achieve better image information fusion.

The existing Transformer based fusion methods (such as CFT[20]) mainly have two limitations: (1) single-stage attention mechanisms are difficult to simultaneously model long-range dependencies within modalities and cross modal global interactions; (2) Feature fusion is mostly concentrated at a single scale, lacking collaborative enhancement of multi-level semantics. To overcome these limitations, we propose a phased fusion paradigm: firstly, the MFT module synchronously enhances intra modal feature consistency and inter modal complementarity through a hierarchical self attention mechanism; Secondly, the DMFF module establishes cross

modal global associations in the high-level semantic space and achieves multi-scale information collaboration through dual path feature enhancement. This divide and conquer design enables MFTNet to fully exploit the potential of bimodal features from both local global and single scale multiscale dimensions. The key contributions of this paper are summarized as follows:

(1) This article proposes a progressive fusion architecture of MFT and DMFF, which synchronously models long-range dependencies within modalities and cross-modal semantic associations through a "local-global" decoupling design.
(2) This article integrates the dynamic network characteristics of YOLOv11 to construct an efficient adaptive detection framework. While maintaining real-time performance, it greatly reduces the number of model parameters and significantly improves robustness in complex scenarios.
(3) The broad applicability and high efficiency of this feature fusion method enable seamless integration with various backbone networks and detection frameworks, such as ResNet and VGG, thereby improving both flexibility and performance.
(4) Through extensive experiments, good results are achieved on the challenging multispectral datasets LLVIP and FLIR by our method.

Related works

Multispectral pedestrian detection

The proposal of several infrared and visible datasets, such as FLIR and LLVIP, has garnered significant attention from researchers in multispectral pedestrian detection. Recent advances in multimodal registration[21] have highlighted that robust feature fusion fundamentally requires solving geometric misalignment between sensors—a prerequisite often overlooked in existing detection frameworks. Recent studies on multi-focus image fusion have demonstrated the effectiveness of adaptive weighting strategies[22] and dynamic transformer architectures[23] in addressing similar cross-domain alignment challenges. Peng et al.[24] proposed Hierarchical Attentive Fusion Network (HAFNet), an adaptive cross-modal fusion framework aimed at enhancing multispectral pedestrian detection performance. Zhang et al.[25] proposed TFDet, addressing RGB pedestrian detection under low-light conditions and enhancing overall multispectral pedestrian detection performance. These approaches align with findings from FusionGCN[26], which emphasizes the importance of hierarchical feature reconstruction through graph-based interactions. Unlike other methods, this approach thoroughly analyzes how noise-fused feature maps affect detection performance, demonstrating that enhancing feature contrast significantly mitigates these issues. Bao et al.[27] proposed Dual-YOLO, an infrared object detection network designed to address high misdetection rates and decreased accuracy resulting from insufficient texture information. Yang et al.[28] introduced a bi-directional adaptive attention gate (BAA-Gate) cross-modal fusion module, which optimizes feature representations from two modalities through attention mechanisms and incorporates an adaptive weighting strategy based on illumination to enhance robustness. Wang et al.[29] developed the Redundant Information Suppression Network (RISNet) to suppress cross-modal redundant information between RGB and infrared images, facilitating the effective fusion of complementary RGB-infrared data. Li et al.[30] proposed a recurrent multispectral feature refinement method that employs multiscale cross-modal homogeneity enhancement and confidence-aware feature fusion to deepen the understanding of complementary content in multimodal data and to explore extensive multimodal feature fusion. Cao[31] focused on generating highly distinguishable multimodal features by aggregating human-related cues from all available samples in multispectral images, achieving multispectral pedestrian detection through locally guided cross-modal feature aggregation and pixel-level detection fusion. Ding et al.[32] recently proposed LG-Diff, a diffusion-based framework that achieves high-quality visible-to-infrared translation in nearshore scenarios through local class-regional guidance and high-frequency prior modeling, demonstrating the potential of diffusion models in cross-modality feature alignment. Despite significant progress in multispectral pedestrian detection from previous studies, CNN convolution-based fusion strategies find it challenging to effectively capture global information in both intra-spectral and inter-spectral images. To address this limitation, this paper proposes a Transformer-based attention scheme.

Transformers

Transformer, known for its significant breakthrough in NLP and outstanding performance, has garnered considerable attention from researchers in computer vision. Increasingly, researchers are applying Transformers to various vision tasks, yielding promising results. Carion et al.[33] introduced DETR (Detection Transformer), marking the inaugural application of Transformer in object detection. Dosovitskiy et al.[34] introduced the ViT (Vision Transformer) model, which employs a self-attention mechanism for image classification. Esser et al.[35] developed VQGAN (Vector Quantized Generative Adversarial Network), combining Transformer and CNN for various applications. Transformer has since been increasingly adopted by researchers in multispectral pedestrian detection. Qingyun et al.[20] introduced the Cross-Modality Fusion Transformer (CFT), a cross-modal feature fusion method designed to fully leverage the combined information from multispectral image pairs, thereby enhancing the reliability and robustness of object detection in open environments. Unlike previous CNN-based approaches, this network learns long-range dependencies guided by Transformer and integrates global contextual information during feature extraction. Lee et al.[36] proposed a Cross-modality Attention Transformer (CAT), aiming to fully exploit the potential of modality-specific features to enhance pedestrian detection accuracy. Notably, Ding et al.[37] proposed a cross-modality bi-attention transformer (CBT) to decouple and guide RGB-thermal fusion in dynamic nearshore environments, demonstrating that transformer-based architectures can effectively align global contextual features across modalities while mitigating temporal degradation—a critical advancement for multispectral fusion in complex scenarios. Shen et al.[38] improved feature fusion in

multispectral object detection through a framework that employs dual Cross Attention Transformers, enhancing the integration of global feature interactions while simultaneously capturing complementary information across modalities. Zang et al.[38] introduced two sub-networks, Fusion Transformer Histogram day (FTHd) and Fusion Transformer night (FTn), tailored for multispectral pedestrian detection in day and night conditions, respectively. However, existing feature fusion methods using Transformer-based self-attention mechanisms have not fully exploited the potential of attention to efficiently capture and integrate complementary information across different modalities. To address this gap, this paper introduces a cross-modal attention feature fusion algorithm, leveraging Transformer architecture for enhanced multispectral pedestrian detection.

Proposed method

Architecture

As shown in Fig. 2, our proposed network model redesigns the YOLOv11 feature extraction network as a two-stream backbone architecture and embeds MFT and DMFF modules to facilitate cross-modal feature fusion. The network consists of four core components: (1) the dual stream CSPDarknet53 backbone network extracts multi-scale features from infrared and visible light images, respectively; (2) Embedding MFT modules at various feature levels to achieve pixel-level modal interaction; (3) Deploy the DMFF module on the Neck end for high-level semantic fusion; (4) The detection head completes pedestrian positioning and recognition. This progressive architecture preserves modality-specific information through a dual-stream backbone and enhances pedestrian detection robustness in complex scenes through the hierarchical fusion of MFT and DMFF.

The MFT module synchronously executes two key tasks in each stage (P3-P5) of the backbone network: firstly, establishing global context correlation of unimodal features through self-attention mechanism (Eq. 5) to strengthen long-range dependencies within the modality; Secondly, designing a cross-modal Q-K projection mechanism (Eqs. 2–4) to my local feature complementarity between modalities through query key-value mapping, achieving pixel level cross-modal interactive response. This design does not require explicit geometric alignment, but instead adaptively captures semantic correspondences between modalities through attention weights.

The DMFF module aggregates multi-level MFT outputs in the Neck section, and its dual path design acts on both spatial and channel dimensions: the Spatial Feature Shrinkage (SFS) path suppresses redundant background noise through channel attention, while the Cross-Modal Enhancement (CFE) path establishes global channel correlations between modalities. This hierarchical fusion strategy enables collaborative optimization of high-level semantic information (such as pedestrian contours) and low-level geometric details (such as texture edges), significantly improving the discriminative ability of feature expression.

The progressive fusion architecture of MFT and DMFF achieves a "locally global" decoupling design: MFT processes high-resolution features at various levels of the backbone network and captures pixel level cross-modal correspondence; DMFF integrates multi-scale information in the high-level semantic space, suppresses

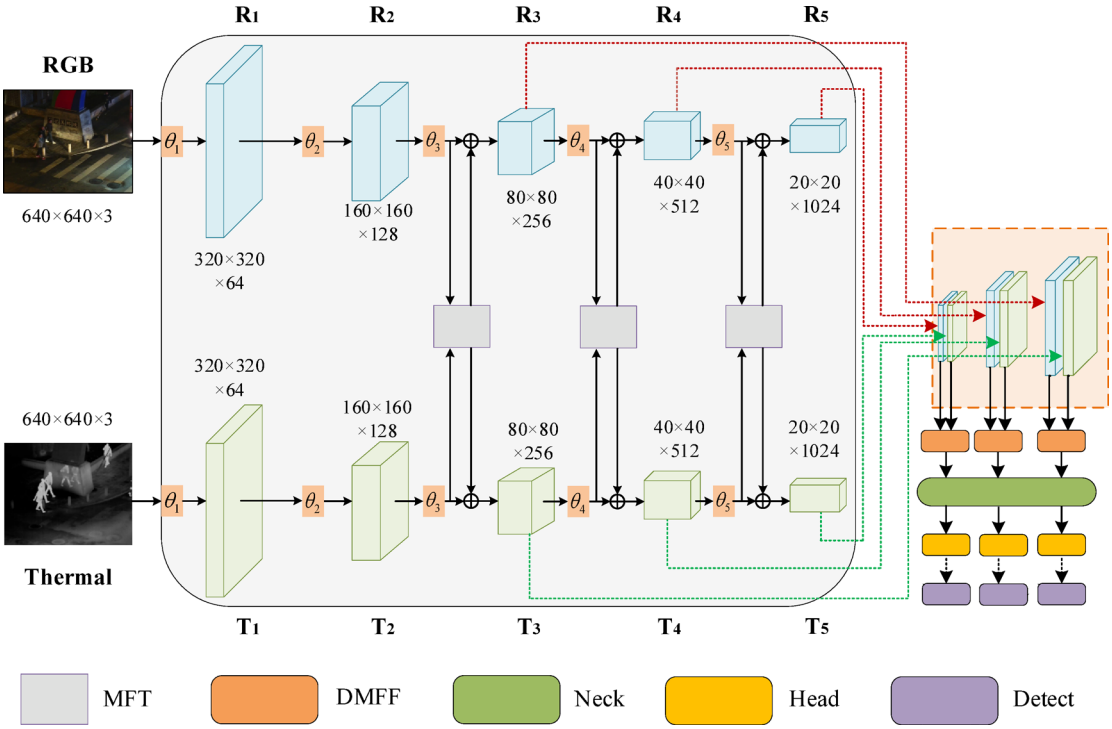


Fig. 2. Multimodal fusion backbone framework. R_i and T_i denote the RGB feature mapping and thermal modal feature mapping after convolution, respectively. θ_i denotes the convolution module. MFT represents our proposed multimodal feature fusion module, DMFF represents the introduced bimodal feature fusion module.

background interference, and enhances semantic consistency. The two form a cascade optimization through a multi-scale feature pyramid, which gradually improves the detection accuracy from low-level detail alignment to high-level semantic correlation, ultimately forming a complementary enhancement effect in complex and varied pedestrian detection tasks.

Multimodal fusion transformer (MFT)

An MFT module is proposed to aggregate multispectral features. As illustrated in Fig. 3, the MFT comprises SAB, SAM, and MLP modules. This configuration is described by Eq. 1:

$$F^i_{R+T} = (SAB(LN(F^i_R + F^i_T))) + (SAM(LN(F^i_R + F^i_T))) + (MLP(LN(F^i_R + F^i_T))) \quad (1)$$

$F^i_{R+T} \in R^{W \times H \times C}$ represents the fused features at the layer indexed by i, H represents the height of the feature map, W the width, and C the total number of channels. respectively. SAB and SAM indicates the feature fusion function with a specified parameter, MLP denotes Multilayer Perceptron, LN denotes LayerNorm. F^i_R represents the RGB feature maps, while F^i_T corresponds to the thermal fused features at the layer indexed by i. The spatial information from multi-scale input feature maps can be processed, the relationship between input features and input channels of different modules can be effectively established, and the interference of background noise can be reduced. In the following sections, the SAB and SAM modules will be introduced in detail.

Self attention Block(SAB)

The SAB module mitigates variability between the two modalities through a self-attention mechanism, enhancing representation and learning capabilities through local-global attention interactions. This enables multi-layer learning of sequence models to effectively extract semantic information regarding location, context, and dependencies within the sequences. In SAB, complementary features between different modalities are

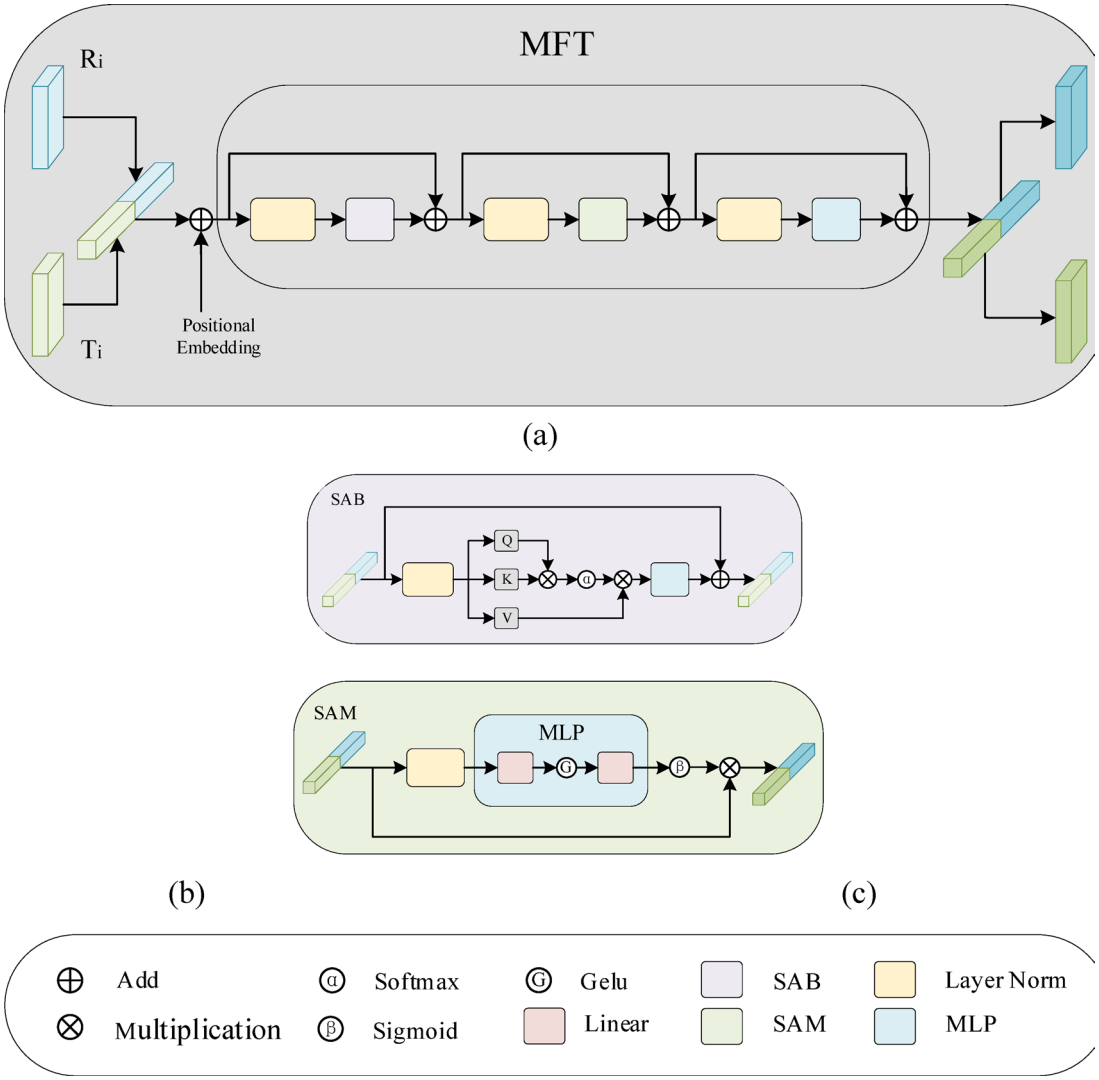


Fig. 3. (**a**) Multimodal fusion transformer module, (**b**) self attention block, (**c**) self aggregation module.

obtained to provide spatial weights for the subsequent attention mechanism. CNN convolution has only local receptive fields, whereas the Transformer can consider global spatial information. Inspired by[20], the Transformer is used for cross-modal feature extraction. The details are shown in Fig. 3b.

Initially, the input sequence ϕ_X undergoes Layer Normalization before being mapped onto three weight matrices to generate a set of queries, keys, and values (Q, K, and V). using the following formulas.

$$Q = \varphi_x W^Q \tag{2}$$

$$K = \varphi_x W^K \tag{3}$$

$$V = \varphi_x W^V \tag{4}$$

W^Q, W^K and W^V are the weight matrices. Furthermore, the self-attention layer calculates the attentional weights using the scaled dot product of Q and K, and subsequently multiplies these weights with V to derive the output φ_Y.

$$\varphi_Y = soft\max\left(\frac{QK^T}{\sqrt{D_K}}\right)V \tag{5}$$

Where $\frac{1}{\sqrt{D_K}}$ is a scaling factor used to prevent the softmax function from converging to regions with the smallest gradient when the dot product becomes large. Subsequently, it passes through a multilayer perceptron (MLP) and finally produces the output sequence φ_Z. The overall formula is as follows:

$$\varphi_Z = W\varphi_Y + \varphi_X \tag{6}$$

The SAB module focuses on the global dependencies within a single modality through self-attention mechanism. As shown in Fig. 3b, its Q/K/V all come from the feature mapping of the same mode, and the feature consistency of the mode itself is enhanced through the global interaction within the layer. In contrast, the CFE module in DMFF (see "Cross-modal Feature Enhancement") is specially designed with a cross-modal interaction mechanism, where Q comes from one mode and K/V comes from another, to achieve cross-spectral attention guidance.

Self aggregation module (SAM)

The feature mapping φ_Z is obtained by the self-attentive weighting module, which encompasses feature mappings for both RGB and IR modalities. Each feature mapping functions as a feature detector. More weight must be assigned to both the visible and infrared feature detectors. We refer to the channel attention block in CBAM, which is shown in Fig. 3c. The SAM module performs feature weighting and adaptive fusion along the channel dimension using an adaptive gating mechanism, thereby enhancing feature expressiveness and diversity.

The input consists of the hybrid feature mapping $\varphi_z \in R^{W\times H\times C}$, which is normalized using Layer Norm across each feature dimension of every sample, ensuring that each feature has a mean of 0 and a variance of 1. Subsequently, the normalized data is fed into a multilayer perceptron (MLP) comprising two linear layers and a Gelu activation function. The sigmoid activation function then generates the channel attention weights $\varphi_W \in R^{C\times 1\times 1}$. Finally, the output feature mapping φ_o is obtained. The specific formula is as follows:

$$\varphi_W = Sig \bmod (MLP(LN(\varphi_Z)) \tag{7}$$

$$\varphi_O = W\varphi_W + \varphi_Z \tag{8}$$

Dual-modal feature fusion

The DMFF module consists of two primary components: the Spatial Feature Shrinkage module and the Cross-modal Feature Enhancement module, Which is illustrated in Fig. 4. Detailed descriptions of these modules are provided in the following sections.

Spatial feature shrinking

In the Spatial Feature Shrinkage (SFS) module, we employ two commonly used pooling methods in deep learning: average pooling and maximum pooling. Average pooling effectively captures the overall information of an image by calculating the average of pixel values within the pooling window, whereas maximum pooling emphasizes salient features by selecting the maximum value from that same window. Each method offers distinct advantages: average pooling enhances global information integration, while maximum pooling focuses on local key features. To capitalize on the advantages of both approaches, we introduce an adaptive weighted pooling mechanism inspired by hybrid pooling[39]. This mechanism enables flexible adjustment of the weights for average and maximum pooling, facilitating more effective extraction of both global and local image features. This operation can be expressed as:

$$\begin{aligned} avg1 &= AvgPool(F_R), \max 1 = MaxPool(F_R) \\ avg2 &= AvgPool(F_I), \max 2 = MaxPool(F_I) \end{aligned} \tag{9}$$

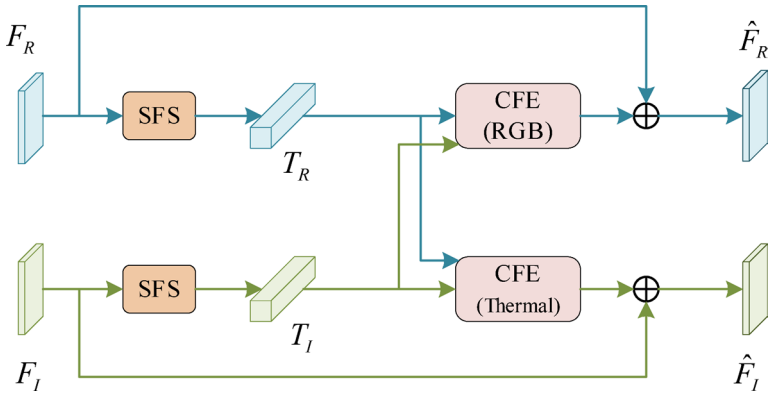


Fig. 4. This figure illustrates the DMFF module, which includes both the SFS and the CFE module.

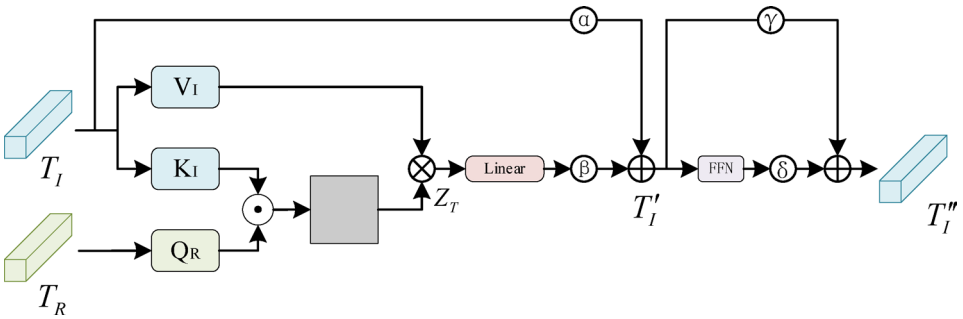


Fig. 5. Shows the details of the Cross-modal Feature Enhancement module. This module enhances feature representations by integrating information from different modalities, with the goal of improving the overall performance of multispectral pedestrian detection systems.

$$\begin{aligned} T_R &= \lambda_{rgb} \cdot avg1 + (1 - \lambda_{rgb}) \cdot \max 1 \\ T_I &= \lambda_{ir} \cdot avg2 + (1 - \lambda_{ir}) \cdot \max 2 \end{aligned} \tag{10}$$

Where, F_R and F_I represent the for-input feature mapping for visible and infrared images respectively. λ_{rgb} and λ_{ir} represent the weights between 0 and 1, which is a learnable parameter. T_R and T_I are the visible and infrared images obtained by hybrid pooling respectively.

Cross-modal feature enhancement

The input feature mappings F_R and $F_I \in R^{H \times W \times C}$ are passed through the SFS module to obtain a set of tokens $T_R, T_I \in R^{HW \times C}$. These are then used as inputs to the CFE module. Given that RGB-IR image pairs often do not align perfectly, two distinct CFE modules are utilized to extract complementary information, enhancing both RGB and IR features. The parameters of these two CFE modules remain separate. In Fig. 5, for the sake of clarity, only an example of the CFE module for the thermal branch is illustrated, as shown in Eq. 11:

T_R and T_I represent the RGB and IR feature representations extracted from the input for the CFE module. Here, T_I'' represents the IR image features enhanced through the CFE module, while $\Gamma_{CFE-I}(\cdot)$ represents the IR branching CFE module.

The CFE module functions as follows: First, tokens from the IR modality T_I are projected onto two independent matrices V_I, K_I to generate a set of values and keys. Next, tokens from the IR modality T_R is mapped onto a different independent matrix Q_R to derive a set of queries, as expressed in the following equation:

$$T_I' = \Gamma_{CFE-I}(\{T_R, T_I\}) \tag{11}$$

The cross-modal features output by the CFE module is enhanced through a feedforward network (FFN), which involves two mechanisms: (1) fusing complementary features extracted by cross-modal attention through the nonlinear transformation of multi-layer perceptrons; (2) By using residual connections to preserve the original modal features, an enhanced structure of "cross-modal interaction + intrinsic feature enhancement" is formed. Specifically, the activation function in FFN (such as GELU) provides non-linear representation capability for cross-modal features, while layer normalization ensures the stability of feature distribution, allowing thermal modal features to adaptively absorb texture clues from visible light modes, and vice versa.

$$V_I = T_I W^V, K_I = T_I W^K, Q_R = T_R W^Q \tag{12}$$

W^Q, W^K, W^V denotes the weight matrix.

Subsequently, a matrix is created using the dot product operation, which is then normalized via the softmax function to produce correlation scores that indicate the resemblance between the features of the RGB and IR modalities. This similarity is then utilized to enhance the RGB features by performing a multiplication of the matrix and the vector V_I, yielding the vector Z_I. Additionally, a multi-attention mechanism is incorporated to enhance the model's understanding of the relationship between RGB and thermal features.

$$Z_I = soft\max\left(\frac{Q_R K_I^T}{\sqrt{D_K}}\right) \cdot V_I \tag{13}$$

$$T_I' = \alpha \cdot Z_I W^O + \beta \cdot V_I \tag{14}$$

$$T_I' = \gamma \cdot T_I' + \delta \cdot FFN(T_I') \tag{15}$$

Third, the tensor $_I$ is transformed reverted to the original domain using a nonlinear mapping and then combined with the input sequence via a residual link (Eq. 14), where W^O refers to the output weight matrix prior to the FFN layer.

Finally, a two-layer fully-connected feedforward network (FFN) within the conventional Transformer framework is used to enhance the global information further, there by enhancing the model's robustness and accuracy, and outputting the enhanced features T_I'' (Eq. 15). Where $\alpha, \beta, \gamma, \delta$ represents all learning parameters.

Experiment

Experimental settings

Datasets

Next, we evaluate the effectiveness of our proposed method through experiments conducted on two publicly available multispectral datasets: LLVIP and FLIR.

LLVIP. The LLVIP dataset comprises 33,672 infrared and visible image pairs, totalling 16,836 pairs. It was trained and tested with 12,025 and 3,463 image pairs, respectively, predominantly captured in low-light conditions, with strict temporal and spatial pairing.

FLIR. The FLIR dataset includes 5,142 aligned RGB-IR image pairs capturing both day and night scenes. Of these, 4,129 pairs were utilized for training, while 1,013 pairs were set aside for testing. The dataset encompasses three object classes: "person," "car," and "bicycle." Due to the lack of alignment in the original dataset, we utilized the FLIR-aligned dataset for our experiments.

Evaluation indicators

For the evaluation of these two publicly available datasets, we employed the mean Average Precision (mAP), a widely used metric in pedestrian detection. This metric encompasses mAP50, mAP75, and mAP. Specifically, mAP50 averages AP values across all categories at IoU = 0.50, while mAP75 does so at IoU = 0.75. The mAP metric aggregates AP values across IoU thresholds between 0.50 and 0.95, with a step of 0.05. Higher values of these metrics indicate better performance of our method on the respective dataset.

Realization details

The code of MFTNet is implemented in PyTorch. The experiments are performed on 7 NVIDIA GeForce GTX 1080 Ti GPUs, with an input resolution of 640 × 640 pixels and a batch size of 28. All parameters in the network are updated using the SGD optimizer with an epoch of 200. We take the training weights of YOLOv11 on the COCO dataset as our pre-training weights.

Quantitative results

Evaluation of the LLVIP Dataset. Table 1 compares the performance of our network with other methods. Our approach achieves state-of-the-art results on this dataset, demonstrating significant performance improvements. Specifically, it outperforms other multimodal networks by a minimum of 0.6% and a maximum of 12.6% in terms of mAP50. When compared to the state-of-the-art RSDet[40] using ResNet50, our method shows superior performance with improvements of 0.6% and 1.4% in mAP50 and mAP, respectively.

Figure 6 demonstrates our method's detection performance on the LLVIP dataset through three key scenarios: (a) Distinctive Features Detection, (b) Occlusion Detection, and (c) Overlap Detection. The visualization shows consistent accuracy in pedestrian identification across varying scales and lighting conditions. Particularly, our approach maintains robust detection capability even in challenging cases with significant occlusion (b) and dense object overlap (c), while preserving fine feature discrimination (a).

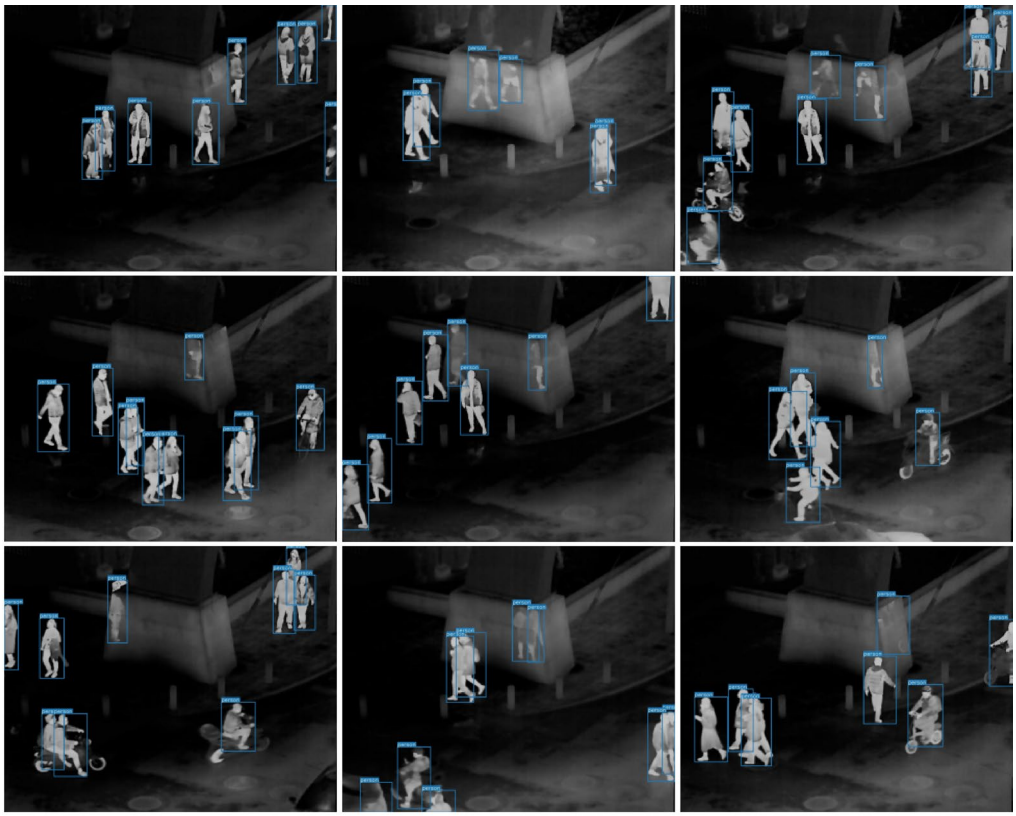
Figure 7 shows the experimental results of MFTNet with other state-of-the-art methods on the LLVIP dataset. As can be seen from the figure, our method achieves state-of-the-art results on the evaluation metric of mAP.

As shown in Table 2, our method achieves state-of-the-art performance in most categories, with a 75.4% mAP50 that surpasses BU-LTT[51] (73.2%) and ThermalDet[52] (74.6%). This superiority stems from three key innovations:

(1) Global Cross-modal Interaction: Compared to CNN-based methods (e.g., BU-ATT[51] with 73.1% mAP50), our Transformer-based MFT module ("Multimodal fusion transformer (MFT)") establishes pixel-wise long-range dependencies between RGB and IR modalities through self-attention mechanisms (Eq. 5). This

Methods	Data	Backbone	mAP50↑	mAP↑
Halfwayfusion[41]	RGB + IR	VGG16	91.4	55.1
GAFF[42]	RGB + IR	ResNet18	94.0	55.8
ProbEn[43]	RGB + IR	ResNet50	93.4	51.5
CSAA[44]	RGB + IR	ResNet50	94.3	59.2
RSDet[40]	RGB + IR	ResNet50	95.8	61.3
FusionGAN[45]	RGB + IR	GAN	83.8	48.1
GANMcC[46]	RGB + IR	GAN	87.8	49.8
NestFuse[47]	RGB + IR	Encoder–decoder	86.9	49.7
DenseFuse[48]	RGB + IR	Encoder–decoder	88.2	50.4
SDNet[27]	RGB + IR	–	86.6	50.8
U2Fusion[49]	RGB + IR	VGG	87.1	47.6
DIVFusion[50]	RGB + IR	Encoder–decoder	89.8	52.0
Ours	RGB + IR	CSPDarknet53	**96.4**	**62.7**

Table 1. Comparison with advanced techniques on the LLVIP dataset.



(a)Distinctive Features Detection (b)Occlusion Detection (c)Overlap Detection

Fig. 6. Visualizing the results on the LLVIP dataset, pedestrians can be accurately detected even if they are occluded by an object.

enables adaptive fusion of complementary features, particularly improving pedestrian detection by 12.6% over YOLOv5s.

(2) Hierarchical Feature Enhancement: The DMFF module ("Dual-modal feature fusion") synergizes multi-scale features through dual-path fusion (SFS + CFE), significantly boosting car detection accuracy to 88.3% (+ 8.3% vs. YOLOv5s). As evidenced by ablation studies (Table 5), the combined use of MFT and DMFF contributes 2.6% mAP improvement.

(3) Computational Efficiency: Despite its superior performance, our model maintains lightweight with only 12.4 M parameters (Fig. 8), achieving 39.4% mAP that outperforms ResNet50-based methods (e.g., RSDet[40] at 61.3% mAP with 95.8 M parameters).

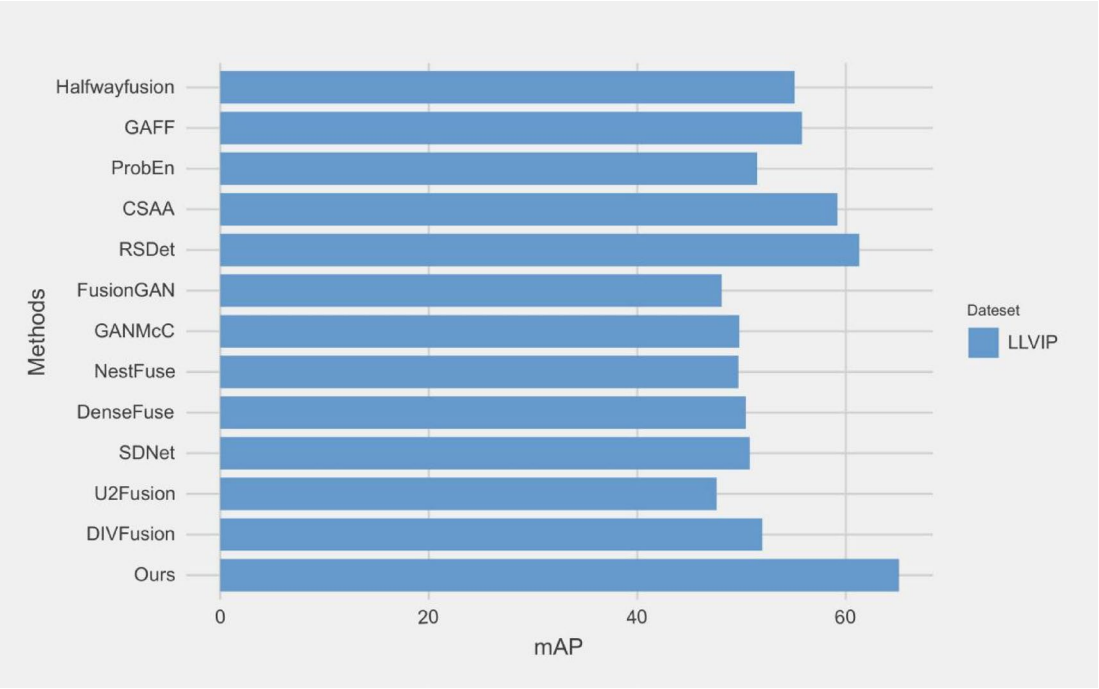


Fig. 7. Visualisation of MFTNet with other state-of-the-art methods on the LLVIP dataset on the evaluation metric for mAP.

Methods	Person↑	Bicycle↑	Car↑	mAP50↑
Faster R-CNN[53]	39.6	54.7	67.6	67.6
SSD[54]	40.9	43.6	61.6	48.7
RetinaNet[55]	52.3	61.3	71.5	61.7
FCOS[56]	69.7	67.4	79.7	72.3
MMTOD-UNIT[53]	49.4	64.4	70.7	61.5
MMTOD-CG[53]	50.3	63.3	70.6	61.4
RefineDet[57]	77.2	57.2	84.5	72.9
TermalDet[52]	78.2	60.0	85.5	74.6
YOLOv3-tiny[58]	67.1	50.3	81.2	66.2
IARet[58]	77.2	48.7	85.8	70.7
CMPD[59]	69.6	59.8	78.1	69.3
PearlGAN[60]	54.0	23.0	75.5	50.8
YOLOv5s[58]	68.3	67.1	80.0	71.8
YOLOF[61]	67.8	**68.1**	79.4	71.8
CFR[62]	74.4	57.7	84.9	72.4
BU-ATT[51]	76.1	56.1	87.0	73.1
BU-LTT[51]	75.6	57.4	86.5	73.2
Ours	**80.9**	57.1	**88.3**	**75.4**

Table 2. Comparison with advanced techniques on the FLIR dataset.

The visualization in Fig. 9 further validates our method's robustness in cross-modal scenarios, where attention maps (Fig. 10) demonstrate effective suppression of background interference while preserving critical pedestrian contours.

Comparison of parameter counts and accuracy on the FLIR dataset: In the FLIR dataset, we conducted a detailed comparison of various detectors in terms of parameter count and detection accuracy. As shown in Fig. 8, our detector achieved higher detection accuracy (75.4% mAP@50) while maintaining a relatively small parameter count (12.41 M). Compared to other multispectral detectors, our approach strikes a better balance between model complexity and performance.

To verify the cross-modal detection capability of our method, we conducted visual analysis on the FLIR dataset. As shown in Fig. 9, the results are organized into three key scenarios: (a) Occlusion Detection, (b)

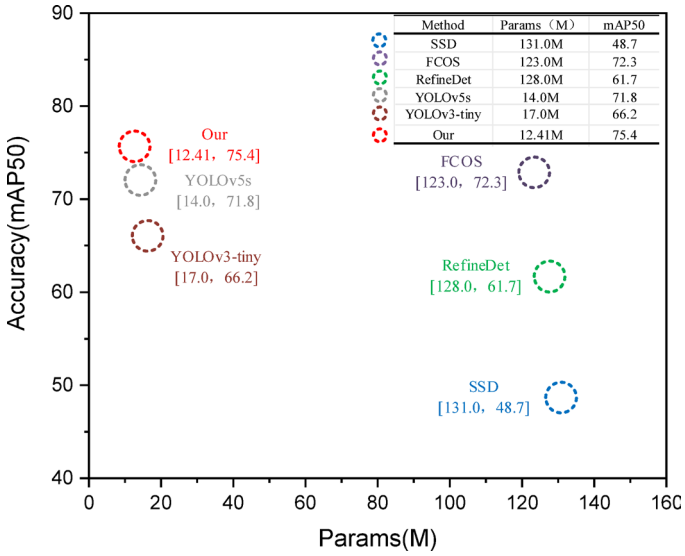
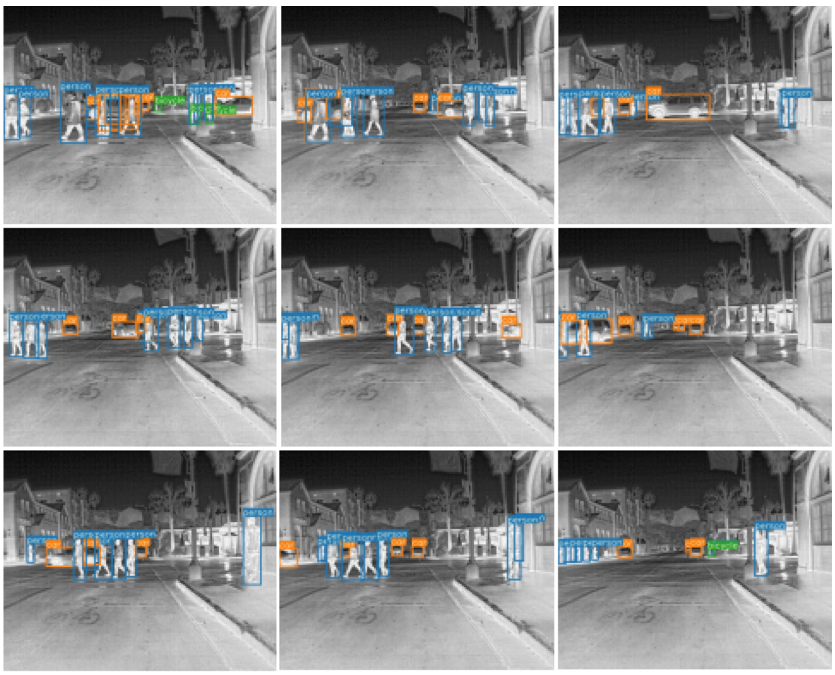


Fig. 8. Parameter vs. accuracy on the FLIR dataset.



(a)Occlusion Detection (b)Overlap Detection (c)Remote detection

Fig. 9. Visualize results on FLIR dataset.

Overlap Detection, and (c) Remote Detection, presented in a multi-scene grid layout. Different colored bounding boxes (blue for pedestrians, orange for vehicles, green for cyclists) clearly indicate detected objects across various conditions. The visualization demonstrates our method's robust performance in accurately identifying targets despite occlusion, overlapping objects, or long distances, providing intuitive validation of the cross-modal detection mechanism.

Figure 11 shows the experimental results of MFTNet with other state-of-the-art methods on the FLIR dataset. As can be seen from the figure, our method achieves state-of-the-art results on the evaluation metric of mAP.

Qualitative analysis

Figure 10 depicts a sample of the visualization results illustrating daytime and nighttime attention maps on the LLVIP and FLIR datasets. In the second and fifth columns of the figure, the baseline approach demonstrates

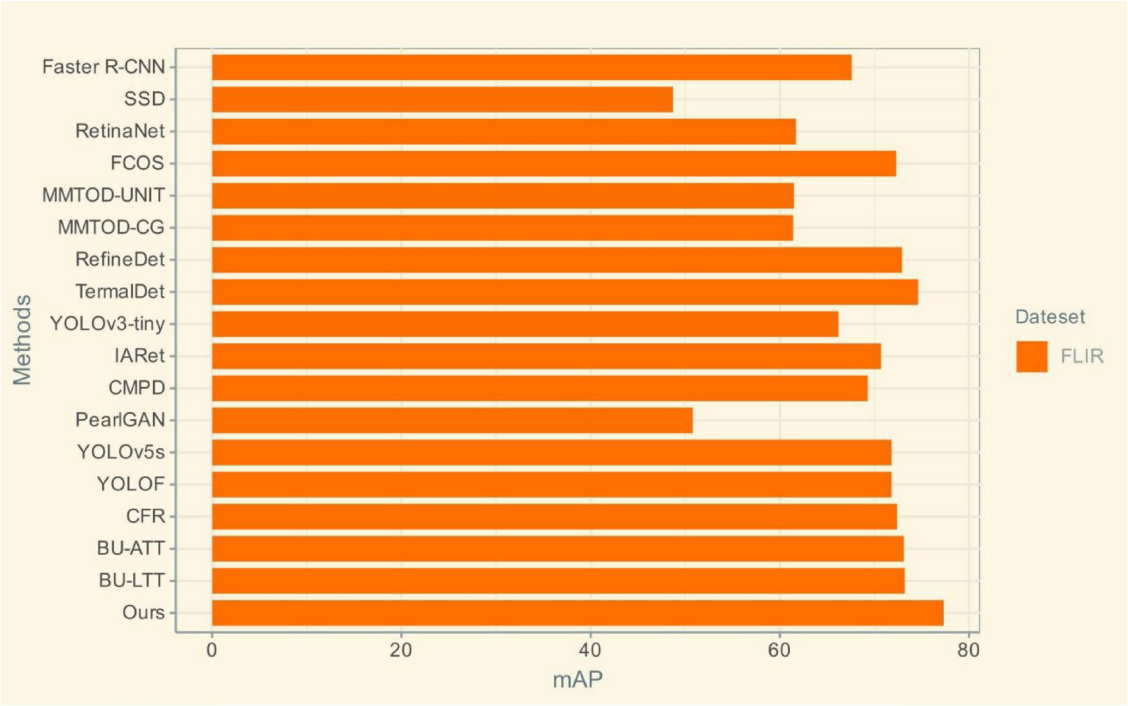


Fig. 10. The first and third rows are visible light images and the second and fourth rows are infrared images. The second and fifth columns visualize the baseline method, and the third and sixth columns visualize our method.

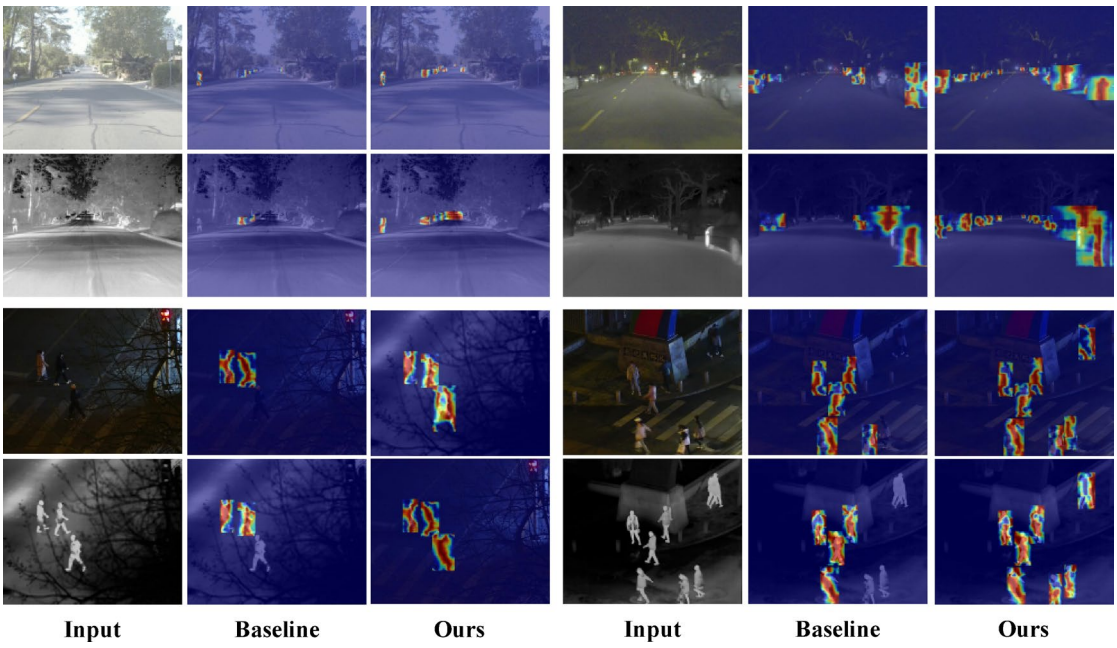


Fig. 11. Visualisation of MFTNet with other state-of-the-art methods on the FLIR dataset on the evaluation metric for mAP.

less comprehensive coverage of various regions in the input image. Conversely, in the third and sixth columns, our approach effectively utilizes global spatial positioning data and correlations between different objects to comprehensively capture all objects. In different datasets, the experimental results of the mAP50 rate are shown in Fig. 12.

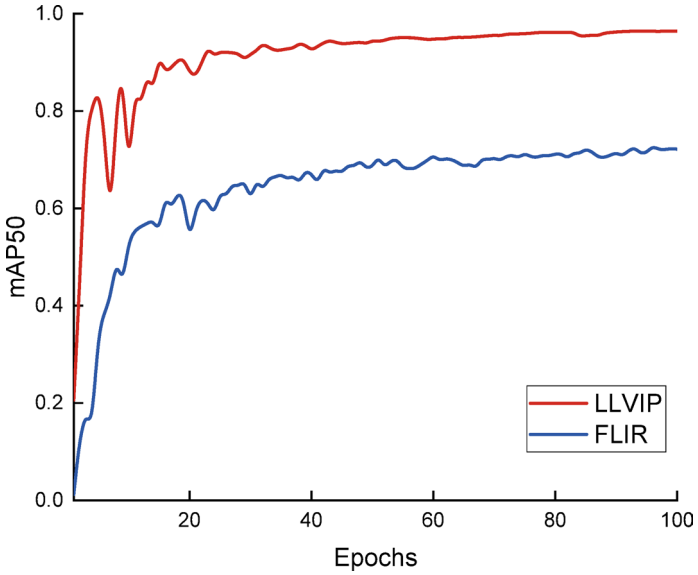


Fig. 12. Training process of mAP50 different dataset.

Detector	Backbone	Methods	Params(M)↓	mAP50↑	mAP75↑	mAP↑
YOLOv11	ResNet50	Baseline	48.73	69.5	29.5	34.8
		Ours	103.32	72.1	31.2	35.6
	VGG16	Baseline	40.53	68.5	27.9	32.8
		Ours	83.17	70.1	29.4	34.4
	CSPDarkNet53	Baseline	3.95	72.1	31.3	36.8
		Ours	12.41	75.4	34.1	39.4

Table 3. Comparison of baselines with our method in different network backbones on the FLIR dataset.

Ablation study

Comparisons with different backbones

To assess the effectiveness of the MFT and DMFF modules, experiments were first performed on the YOLOv11 detector using three different backbones: ResNet50, VGG16, and CSPDarkNet53. The results from the FLIR dataset, shown in Table 3, demonstrate that our approach using ResNet50, VGG16, and CSPDarkNet53 outperformed the baseline method, achieving improvements of 0.8%, 1.6%, and 2.9% in representative mAPs, respectively. Thus, it is concluded that our method is applicable to a variety of backbone networks.

Figure 13 illustrates the results of our ablation experiments conducted on the YOLOv11 detector using three different backbone networks: ResNet50, VGG16, and CSPDarkNet53. The bar chart compares the performance of our proposed method against the baseline for each backbone. As shown, our approach consistently outperforms the baseline across all three backbones, with the most significant improvement observed when using CSPDarkNet53. These results highlight the versatility of our method across different backbone architectures and its effectiveness in enhancing detection performance.

Effects of different module positions

As shown in Fig. 14, this section presents the mAP50, mAP75, and mAP values for various positions and numbers of MFT modules on the FLIR dataset. Table 4 presents the positional details of these modules. The results in Table 4 indicate that the highest mAP index values, 75.4%, 34.1%, and 39.4%, correspond to the positions of the MFT fusion modules at layers 3, 4, and 5, respectively. The Continued increase in the number of fusion modules results in a decrease in the mAP metric. Therefore, we conclude that the optimal fusion position occurs after the convolution of the third, fourth, and fifth layers.

Ablation of different modules

To assess the effectiveness of the MFT and DMFF modules, we excluded these modules from our method. Table 5 shows that integrating the MFT module enhances the mAP performance of the LLVIP dataset by 1.1% and the FLIR dataset by 1.9% compared to the baseline. Similarly, integrating the DMFF module enhances the mAP performance by 0.5% for the LLVIP dataset and 1.0% for the FLIR dataset compared to the baseline. Introducing both MFT and DMFF modules results in an improved mAP performance of 2.3% for the LLVIP dataset and 2.9% for the FLIR dataset

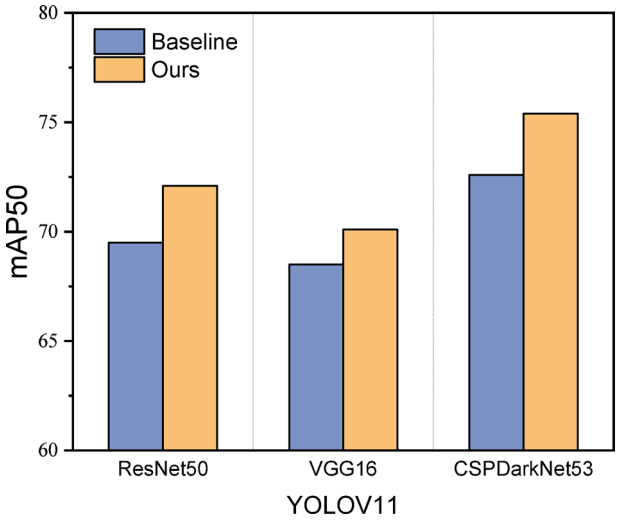


Fig. 13. The bar chart indicates that our modules are embedded in different network backbones.

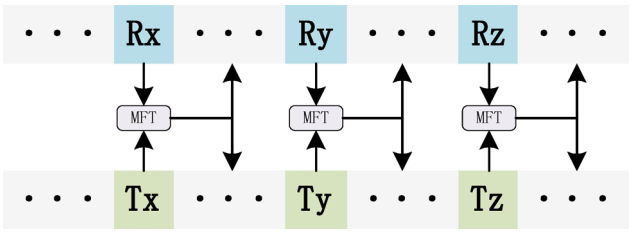


Fig. 14. MFT modules in different positions and in different quantities.

1	2	3	4	5	mAP50	mAP75	mAP
√	√	√	√	√	73.4	33.5	38.2
	√	√	√	√	74.1	32.1	37.6
		√	√	√	75.4	34.1	39.4

Table 4. Differences in the performance of fusion modules at different locations on the FLIR dataset.

MFT	DMFF	LLVIP			FLIR		
		mAP50	mAP75	mAP	mAP50	mAP75	mAP
		94.8	69.6	60.4	72.1	31.3	36.8
√		95.2	70.8	61.5	74.5	33.0	38.4
	√	96.1	71.2	60.9	72.6	32.9	37.5
√	√	96.4	71.5	62.7	75.4	34.1	39.4

Table 5. Ablation studies of MFT and DMFF modules using mAP50, mAP75 and mAP as evaluation metrics.

compared to the baseline. Overall, the experimental results exhibit consistent trends, particularly in the mAP metrics. These results clearly demonstrate the effectiveness of these modules.

Comparison with different input modalities

To demonstrate the overall effectiveness of our proposed method, we configured separate input modes for comparison and conducted tests on the FLIR dataset. Table 6 presents the evaluation of the experimental results, including efficiency (i.e., network parameters) and effectiveness (i.e., mAP50, mAP75, and mAP). Our method significantly enhances the performance of multispectral object detection compared to unimodal and conventional bimodal inputs.

Input	Method	Param(M)	mAP50	mAP75	mAP
RGB	CSPDarknet53	2.59	64.5	23.4	30.6
Thermal	CSPDarknet53	2.59	72.6	33.5	38.4
RGB + T	Two Stream	3.95	72.1	31.3	36.8
RGB + T	Ours	12.41	75.4	34.1	39.4

Table 6. R visible modal input, T denotes thermal modal input, R + T denotes bimodal input.

Conclusions

We propose an innovative cross-modal feature fusion framework aimed at overcoming the drawbacks of CNN-based multispectral fusion techniques, particularly their constrained receptive domain that focuses primarily on local feature interactions. Specifically, we introduce a Transformer-based self-attentive fusion module that unifies intra- and inter-modal information, effectively addressing existing limitations. This framework enhances the model's ability to elucidate relationships between different modalities, thereby improving the comprehensiveness and accuracy of feature fusion in multispectral object detection tasks. Additionally, we conducted numerous ablation experiments to demonstrate our method's effectiveness, achieving 65.1% and 77.3% accuracy on the challenging LLVIP and FLIR datasets, respectively, surpassing current state-of-the-art techniques. Moving forward, we will explore a streamlined and efficient cross-modal feature fusion framework in-depth to meet the multimodal task requirements across various domains. Furthermore, we intend to extend our approach to broader application areas, encompassing object detection, behavioural analysis, and multimodal tasks like environment perception, to address diverse challenges and requirements. We aim to contribute further to the advancement of multimodal data processing through ongoing research and practical applications, promoting the adoption and application of related technologies.

Data availability

The datasets used during this study are publicly available in the [LLVIP] and [FLIR] repository at [https://bupt-ai-cz.github.io/LLVIP/] and [https://www.flir.com/oem/adas/adas-dataset-form/].

Received: 13 November 2024; Accepted: 21 May 2025
Published online: 29 May 2025

References

1. Bilal, M. et al. A low-complexity pedestrian detection framework for smart video surveillance systems[J]. *IEEE Trans. Circuits Syst. Video Technol.* **27** (10), 2260–2273 (2016).
2. Zhang, S. et al. Pedestrian search in surveillance videos by learning discriminative deep features[J]. *Neurocomputing* **283**, 120–128 (2018).
3. Führ, G. & Jung, C. R. Camera self-calibration based on nonlinear optimization and applications in surveillance systems[J]. *IEEE Trans. Circ. Syst. Video Technol.* **27** (5), 1132–1142 (2015).
4. Geiger, A., Lenz, P. & Urtasun, R. Are we ready for autonomous driving? The kitti vision benchmark suite[C]. *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3354–3361. (2012).
5. Jeong, M., Ko, B. C. & Nam, J. Y. Early detection of sudden pedestrian crossing for safe driving during summer nights[J]. *IEEE Trans. Circuits Syst. Video Technol.* **27** (6), 1368–1380 (2016).
6. Hwang, S. et al. Multispectral pedestrian detection: Benchmark dataset and baseline[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1037–1045. (2015).
7. Lee, W. J. & Lee, S. W. Improved Spatiotemporal noise reduction for very low-light environments[J]. *IEEE Trans. Circ. Syst. II Express Briefs*. **63** (9), 888–892 (2016).
8. Song, W. et al. Context-interactive CNN for person re-identification[J]. *IEEE Trans. Image Process.* **29**, 2860–2874 (2019).
9. Li, C. et al. Illumination-aware faster R-CNN for robust multispectral pedestrian detection[J]. *Pattern Recogn.* **85**, 161–171 (2019).
10. Zhang, L. et al. Weakly aligned cross-modal learning for multispectral pedestrian detection[C]. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5127–5137. (2019).
11. Zhou, K., Chen, L. & Cao, X. Improving multispectral pedestrian detection by addressing modality imbalance problems[C]. *Computer Vision–ECCV: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer International Publishing, 2020: 787–803. (2020).
12. Li, H. & Wu X. J. DenseFuse: a fusion approach to infrared and visible images[J]. *IEEE Trans. Image Process.* **28** (5), 2614–2623 (2018).
13. Deng, J. et al. Imagenet: A large-scale hierarchical image database[C]. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 248–255. (2009).
14. Ma, J. et al. STDFusionNet: an infrared and visible image fusion network based on salient target detection[J]. *IEEE Trans. Instrum. Meas.* **70**, 1–13 (2021).
15. Han, D. et al. Multi-exposure image fusion via deep perceptual enhancement[J]. *Inform. Fusion*. **79**, 248–262 (2022).
16. Deng, X. & Dragotti, P. L. Deep convolutional neural network for multi-modal image restoration and fusion[J]. *IEEE Trans. Pattern Anal. Mach. Intell.* **43** (10), 3333–3348 (2020).
17. Liu, Y. et al. Multi-focus image fusion with a deep convolutional neural network[J]. *Inform. Fusion*. **36**, 191–207 (2017).
18. Ma, J. et al. DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion[J]. *IEEE Trans. Image Process.* **29**, 4980–4995 (2020).
19. Xu, H., Ma, J. & Zhang, X. P. MEF-GAN: multi-exposure image fusion via generative adversarial networks[J]. *IEEE Trans. Image Process.* **29**, 7203–7216 (2020).
20. Qingyun, F., Dapeng, H. & Zhaokui, W. Cross-modality fusion transformer for multispectral object detection. (2021). arXiv preprint. arxiv:2111.00273.
21. Ding, J. et al. Modal-invariant progressive representation for multimodal image registration[J]. *Inform. Fusion*. **117**, 102903 (2025).
22. Jiang, J. et al. Multi-focus image fusion method based on adaptive weighting and interactive information modulation[J]. *Multimedia Syst.* **30** (5), 290 (2024).

23. Zhai, H. et al. MSI-DTrans: a multi-focus image fusion using multilayer semantic interaction and dynamic transformer[J]. *Displays* **85**, 102837 (2024).
24. Peng, P. et al. HAFNet: hierarchical attentive fusion network for multispectral pedestrian detection[J]. *Remote Sens.* **15**(8), 2041 (2023).
25. Zhang X, Zhang X, Wang J, et al. TFDet: Target-aware fusion for RGB-T pedestrian detection[J]. *IEEE Transactions on Neural Networks and Learning Systems* (2024).
26. Ouyang, Y. et al. FusionGCN: multi-focus image fusion using superpixel features generation GCN and pixel-level feature reconstruction CNN[J]. *Expert Syst. Appl.* **262**, 125665 (2025).
27. Bao, C. et al. Dual-YOLO architecture from infrared and visible images for object detection[J]. *Sensors* **23** (6), 2934 (2023).
28. Yang, X. et al. BAANet: Learning bi-directional adaptive attention gates for multispectral pedestrian detection[C]. *2022 international Conference on Robotics and Automation (ICRA)*. IEEE, 2920–2926. (2022).
29. Wang, Q. et al. Improving rgb-infrared pedestrian detection by reducing cross-modality redundancy[C]. *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 526–530. (2022).
30. Li, R. et al. *Multiscale Cross-modal Homogeneity Enhancement and Confidence-aware Fusion for Multispectral Pedestrian Detection[J]* (IEEE Transactions on Multimedia, 2023).
31. Cao, Y. et al. Locality guided cross-modal feature aggregation and pixel-level fusion for multispectral pedestrian detection[J]. *Inform. Fusion*. **88**, 1–11 (2022).
32. Ding, J. et al. LG-Diff: learning to follow local class-regional guidance for nearshore image cross-modality high-quality translation[J]. *Inform. Fusion*. **117**, 102870 (2025).
33. Carion, N. et al. End-to-end object detection with transformers[C]. *European Conference on Computer Vision*. Cham: Springer International Publishing, pp. 213–229. (2020).
34. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint. arxiv:2010.11929 (2020).
35. Esser, P., Rombach, R. & Ommer, B. Taming transformers for high-resolution image synthesis[C]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: 12873–12883. (2021).
36. Shen, J. et al. ICAFusion: iterative cross-attention guided feature fusion for multispectral object detection[J]. *Pattern Recogn.* **145**, 109913 (2024).
37. Ding, J. et al. *Novel Pipeline Integrating Cross-Modality and Motion Model for Nearshore Multi-object Tracking in Optical Video surveillance[J]* (IEEE Transactions on Intelligent Transportation Systems, 2024).
38. Zang, Y. et al. Transformer fusion and histogram layer multispectral pedestrian detection network[J]. *Signal. Image Video Process.* **17** (7), 3545–3553 (2023).
39. Yu, D. et al. Mixed pooling for convolutional neural networks[C]. *Rough Sets and Knowledge Technology: 9th International Conference, RSKT 2014,* Shanghai, China, October 24–26, 2014, Proceedings 9. Springer International Publishing, 364–375. (2014).
40. Zhao, T., Yuan, M. & Wei, X. Removal and selection: improving RGB-infrared object detection via Coarse-to-Fine Fusion. arXiv preprint. arxiv:2401.10731 (2024).
41. Liu, J. et al. Multispectral deep neural networks for pedestrian detection. arXiv preprint. arxiv:1611.02644 (2016).
42. Zhang, H. et al. Guided attentive feature fusion for multispectral pedestrian detection[C]. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 72–80. (2021).
43. Chen, Y. T. et al. Multimodal object detection via probabilistic ensembling[C]. *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, pp. 139–158. (2022).
44. Cao, Y. et al. Multimodal object detection by channel switching and spatial attention[C]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 403–411. (2023).
45. Ma, J. et al. FusionGAN: A generative adversarial network for infrared and visible image fusion[J]. *Inform. Fusion*. **48**, 11–26 (2019).
46. Ma, J. et al. GANMcC: a generative adversarial network with multiclassification constraints for infrared and visible image fusion[J]. *IEEE Trans. Instrum. Meas.* **70**, 1–14 (2020).
47. Li, H., Wu, X. J. & Durrani, T. NestFuse: an infrared and visible image fusion architecture based on nest connection and spatial/channel attention models[J]. *IEEE Trans. Instrum. Meas.* **69** (12), 9645–9656 (2020).
48. Zhang, H. & Ma, J. SDNet: a versatile squeeze-and-decomposition network for real-time image fusion[J]. *Int. J. Comput. Vision*. **129** (10), 2761–2785 (2021).
49. Xu, H. et al. U2Fusion: a unified unsupervised image fusion network[J]. *IEEE Trans. Pattern Anal. Mach. Intell.* **44** (1), 502–518 (2020).
50. Tang, L. et al. DIVFusion: darkness-free infrared and visible image fusion[J]. *Inform. Fusion*. **91**, 477–493 (2023).
51. Kieu, M., Bagdanov, A. D. & Bertini, M. Bottom-up and Layerwise domain adaptation for pedestrian detection in thermal images[J]. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)*. **17** (1), 1–19 (2021).
52. Cao, Y. et al. Every feature counts: An improved one-stage detector in thermal imagery[C]. *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*. IEEE, pp. 1965–1969. (2019).
53. Devaguptapu, C. et al. M,. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery[C]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0. (2019).
54. Liu, W. et al. Ssd: Single shot multibox detector[C]. *Computer Vision–ECCV 2016: 14th European Conference*, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, pp. 21–37. (2016).
55. Lin, T. Y. et al. Focal loss for dense object detection[C]. *Proceedings of the IEEE International Conference on Computer Vision*. 2980–2988. (2017).
56. Tian, Z. et al. Fully convolutional one-stage 3d object detection on lidar range images[J]. *Adv. Neural. Inf. Process. Syst.* **35**, 34899–34911 (2022).
57. Zhang, S. et al. Single-shot refinement neural network for object detection[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4203–4212. (2018).
58. Jiang, X. et al. IARet: a lightweight multiscale infrared aerocraft recognition algorithm[J]. *Arab. J. Sci. Eng.* **47** (2), 2289–2303 (2022).
59. Li, Q. et al. *Confidence-aware Fusion Using dempster-shafer Theory for Multispectral Pedestrian detection[J]* (IEEE Transactions on Multimedia, 2022).
60. Luo, F. et al. Thermal infrared image colorization for nighttime driving scenes with top-down guided attention[J]. *IEEE Trans. Intell. Transp. Syst.* **23** (9), 15808–15823 (2022).
61. Chen, Q. et al. You only look one-level feature[C]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13039–13048. (2021).
62. Zhang, H. et al. Multispectral fusion for object detection with cyclic fuse-and-refine blocks[C]. *2020 IEEE International conference on image processing (ICIP)*. IEEE, 276–280. (2020).

Acknowledgements

This research was supported by the following projects: (1) Inner Mongolia Science and Technology Depart-

ment, Inner Mongolia Natural Science Foundation Program, 2024QN06012. (2) Inner Mongolia Autonomous Region Department of Education, First-class Discipline Research Special Project, YLXKZX-NKD-014. (3) Inner Mongolia Autonomous Region Department of Education, Scientific Research Project of Higher Education Institutions in Inner Mongolia Autonomous Region, NJZY23081. (4) Inner Mongolia Science and Technology Department, 2022 Third Batch of Regional Key R&D and Achievement Transformation Program Projects (Social Public Welfare), 2022YFSH0044. (5) Project of Basic Research Funds for Colleges and Universities directly under Inner Mongolia,2023QNJS198.

Author contributions

Conceptualization, G.R. and G.L.; methodology, G.L.; software, G.R.; validation, G.L., G.R. and J.W.; formal analysis, J.W.; investigation, M.Z.; resources, Z.Y.; data curation, B.J.; writing—original draft preparation, H.G.; writing—review and editing, Q.G.; visualization, G.L.; supervision, G.L.; project administration, G.R.; funding acquisition, G.R. All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to G.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.