Frontiers in Physics

TYPE Original Research
PUBLISHED 06 June 2024
DOI 10.3389/fphy.2024.1398678

OPEN ACCESS

EDITED BY
Wenfeng Xia,
King's College London, United Kingdom

REVIEWED BY
Shahaf Asban,
University of South Florida, United States
Muhammad Asim,
Guangdong University of Technology, China

*CORRESPONDENCE
Ning Zhang,
jojenmn@outlook.com

RECEIVED 10 March 2024
ACCEPTED 10 May 2024
PUBLISHED 06 June 2024

CITATION
Zhang N and Zhu W (2024), Enhancing target detection accuracy through cross-modal spatial perception and dual-modality fusion. *Front. Phys.* 12:1398678.
doi: 10.3389/fphy.2024.1398678

Enhancing target detection accuracy through cross-modal spatial perception and dual-modality fusion

Ning Zhang[1,2,3]* and Wenqing Zhu[1,2,3]

[1]Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai, China, [2]University of Chinese Academy of Sciences, Beijing, China, [3]Key Laboratory of Infrared System Detection and Imaging Technology, Chinese Academy of Sciences, Shanghai, China

The disparity between human and machine perception of spatial information presents a challenge for machines to accurately sense their surroundings and improve target detection performance. Cross-modal data fusion emerges as a potential solution to enhance the perceptual capabilities of systems. This article introduces a novel spatial perception method that integrates dual-modality feature fusion and coupled attention mechanisms to validate the improvement in detection performance through cross-modal information fusion. The proposed approach incorporates cross-modal feature extraction through a multi-scale feature extraction structure employing a dual-flow architecture. Additionally, a transformer is integrated for feature fusion, while the information perception of the detection system is optimized through the utilization of a linear combination of loss functions. Experimental results demonstrate the superiority of our algorithm over single-modality target detection using visible images, exhibiting an average accuracy improvement of 30.4%. Furthermore, our algorithm outperforms single-modality infrared image detection by 3.0% and comparative multimodal target detection algorithms by 3.5%. These results validate the effectiveness of our proposed algorithm in fusing dual-band features, significantly enhancing target detection accuracy. The adaptability and robustness of our approach are showcased through these results.

KEYWORDS

spatial perception, cross-modal data fusion, dual-modality feature fusion, target detection performance, multi-scale feature extraction, dual-band feature fusion

1 Introduction

Target detection is a common perception task in the field of remote sensing, and usually we use algorithms to extract human-eye vision-friendly feature information in order to achieve target detection. However, human visual perceptual information and detection system perceptual information are not uniform. Human vision-friendly feature information in any single modality can help humans to analyze the content in an image, but this does not fully reflect the true target perception capability of the detection system [1–3]. Due to the presence of a large number of visually unfriendly features, further exploitation of the perceptual capabilities of the system is possible, and cross-modal feature fusion is the key to exploit the target perception potential of the detection system [4–6]. Deep learning has emerged as a widely adopted methodology for addressing single-band target detection tasks,

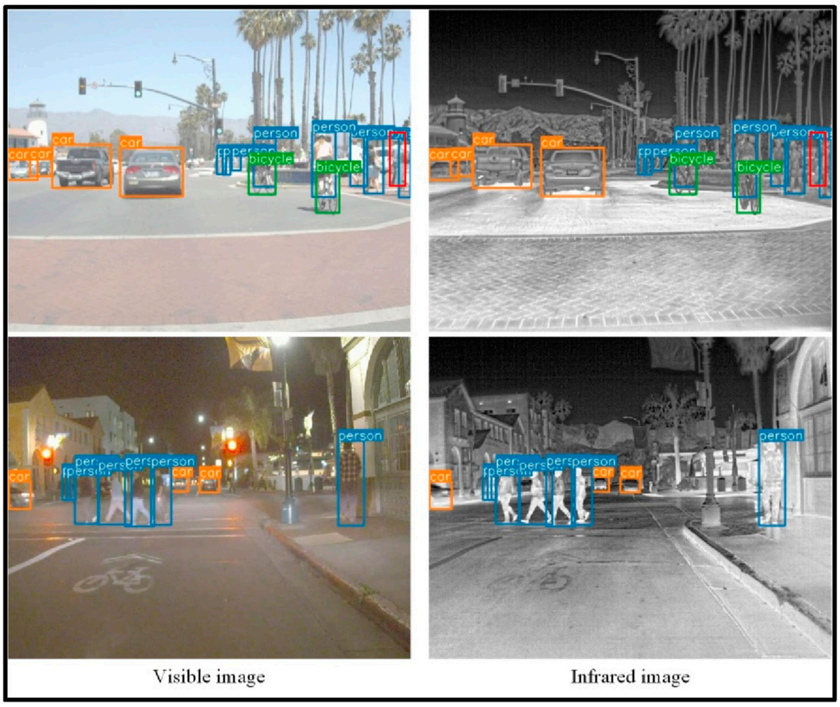

FIGURE 1
Example of infrared and visible dual-band object detection.

and it can be classified into two main categories: one-stage detection and two-stage detection algorithms. The conventional two-stage detectors encompass prominent models such as R-CNN [1], Fast R-CNN [2], Faster R-CNN [3], and Mask R-CNN [4]. In the two-stage detection paradigm, the initial step involves the generation of candidate regions, referred to as region proposals. These candidate regions are subsequently mapped onto the feature map to extract the corresponding feature matrices. These matrices serve as the foundation for conducting classification and regression tasks, ultimately enabling the determination of precise bounding box coordinates.

In the realm of fusion models, particularly in the context of autonomous driving, there have been studies exploring the combination of You Only Look Once (YOLO) and radar sources [5, 6]. These fusion models aim to leverage the strengths of both YOLO, a popular object detection algorithm, and radar sensors to enhance target detection and perception in autonomous vehicles. By fusing the visual information from YOLO with the radar data, these models can potentially improve the robustness and accuracy of object detection, especially in challenging environmental conditions such as low visibility or occlusions. The combination of YOLO and radar sources provides complementary information, with YOLO focusing on visual cues and radar sensors providing valuable depth and motion information. By comparing the proposed novel spatial perception method with these fusion models, it becomes possible to assess the advantages, limitations, and performance improvements achieved by incorporating dual-modality feature fusion and coupled attention mechanisms in the context of target detection and autonomous driving scenarios. Prominent one-stage detectors in the field of target detection include SSD [5], the YOLO series (comprising YOLOv1 [6], YOLOv2 [7], YOLOv3, YOLOv4, and YOLOv5), and Retina-Net [8]. Unlike two-stage detectors, one-stage detection algorithms eliminate the need for generating candidate frames, instead relying on a direct regression method to detect frame location information and category confidence in a single step. Infrared and visible dual-band target detection represents a new focal point in the research domain, as it enables the integration of information from different spectral bands to enhance system robustness. Figure 1 illustrates an example of target detection. Hwang et al. [9] introduced the KAIST multispectral pedestrian benchmark dataset, the first large-scale infrared-visible dual-band dataset for pedestrian detection. This dataset was specifically designed to improve system robustness through the integration of information from diverse spectral bands [9].

Based on this dataset, Wagner et al. were the first to apply deep learning to the field of multispectral pedestrian detection [10]. They also proposed early fusion and late fusion frameworks, and comparison results showed that late fusion outperformed early fusion and the traditional ACF algorithm [11]. Konig et al. proposed a multispectral fusion algorithm based on Faster R-CNN [3]. They also proposed a dual-stream network architecture (RPN) for multispectral pedestrian detection based on Faster R-CNN and compared the effects of different stages of feature fusion, showing that mid-stage fusion outperformed early or late fusion [12]. The distance statistics of single classes, such as pedestrians, in the visible range alone, infrared, and cross-distance, provide valuable insights into the statistical distinctness between these representations. By comparing the distances between samples of the same class in different modalities, we can assess the degree of separation and similarity among the representations. The distance statistics allow us to quantitatively measure the dissimilarity or distinctness between these modalities. The analysis revealed that the distances between pedestrian samples in the visible range alone and infrared were significantly different. This suggests that the two modalities capture distinct information about pedestrians, with each modality providing complementary cues. The visible range primarily captures visual appearance and shape information, while the infrared modality emphasizes thermal signatures and motion patterns. Furthermore, the cross-distance between the fused representation and the individual modalities was examined. The results demonstrated that the cross-distance was significantly smaller compared to the distances within each modality. This indicates that the fused representation successfully combines the distinctive information from both modalities, resulting in a representation that is closer to each individual modality compared to the distances observed within each modality alone. The accuracy of dual-band target detection is influenced by the fusion stage, and effectively fusing the dual-band features poses a significant challenge in research. Li et al. [13] proposed the illumination-aware Faster R-CNN network (IAF R-CNN), while Guan et al. [14] introduced an illumination-aware neural network model. Both approaches involve training separate networks to estimate illumination information, enabling adaptive fusion. The IAF R-CNN method achieves a balance in prediction results between infrared and visible images based on the network's predicted light values. On the other hand, Guan's model employs weights to balance the detection outcomes of the subnetworks designated for daytime and nighttime light conditions. While these approaches enable the fusion of detection results, they fail to inherently increase the information content, resulting in limited model performance.

Enhancing target detection accuracy through cross-modal spatial perception and dual-modality fusion is a promising approach in the field of machine perception. The disparity between human and machine perception of spatial information has been a challenge for machines to effectively sense their surroundings and improve target detection performance. However, cross-modal data fusion provides a viable solution to enhance the perceptual capabilities of systems. In this regard, a novel spatial perception method is introduced in this context, which integrates dual-modality feature fusion and coupled attention mechanisms [14–16]. By fusing information from multiple modalities, the proposed approach aims to validate the enhancement of detection performance through cross-modal information fusion. The approach incorporates cross-modal feature extraction using a multi-scale feature extraction structure with a dual-flow architecture [17–19]. Additionally, a transformer is utilized for feature fusion, optimizing the information perception of the detection system. The algorithm's effectiveness is demonstrated through experimental results, showcasing its superiority over single-modality target detection methods [20–24]. The proposed algorithm outperforms single-modality target detection using visible images by an average accuracy improvement of 30.4% and single-modality infrared image detection by 3.0%. Furthermore, it surpasses comparative multimodal target detection algorithms by 3.5%. These promising results validate the effectiveness of the proposed algorithm in fusing dual-band features, significantly enhancing target detection accuracy. This approach exhibits adaptability and robustness, highlighting its potential for advancing target detection in various domains [25–32].

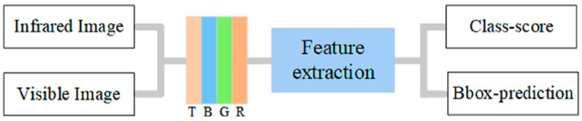

FIGURE 2
The early fusion framework offers several advantages in target detection tasks.

In contrast to scene division methods, researchers have proposed feature fusion approaches to adapt target detection to various scenes. Zhang et al. introduced the Cross-modal Interactive Attention Network (CIAN) [15], which employs channel attention to weight feature maps extracted from different spectral images, thereby achieving feature fusion. Additionally, Zhang et al. proposed the Cyclic Fuse-and-Refine (CFR) module [16] to balance the complementarity and consistency of dual-band features. Incorporating any network using this module leads to improved detection performance. Furthermore, Zhang et al. proposed the Guided Attention Feature Fusion (GAFF) [17] method, which utilizes inter-modal and intra-modal attention modules to guide dynamic weighting and fusion of features, enhancing detection accuracy. The imaging methods for infrared and visible detectors yield distinct visual perceptions, and features extracted from different spectral bands often exhibit significant inconsistency and suffer from alignment accuracy issues, making targets prone to missed detection or misidentification.

Effectively integrating valuable information from different bands and enhancing target detection accuracy and system robustness are the primary considerations in dual-band target detection algorithms. Therefore, this paper presents a dual-band target detection algorithm based on a linear transformer and channel attention. The algorithm utilizes a dual-flow architecture with YOLOv5 as the backbone to separately extract infrared (IR) and visible features. The feature fusion module, based on a linear transformer and channel attention, adaptively learns the interrelationship between IR and visible light without the need for manually designing fusion rules. This adaptive approach enhances the algorithm's adaptability and robustness in variable and complex scenes, resulting in an overall improvement in target detection. The algorithm can effectively enhance target detection accuracy and can be employed to improve detection outcomes. Additionally, various feature interaction methods are designed, and the detection results are analyzed to investigate the impact of different cross-modal interaction approaches.

2 Materials and methods

2.1 Related works

Dual-band fusion target detection methods can be classified into Early Fusion, Halfway Fusion and Late Fusion according to the different stages of fusion. Early Fusion uses infrared images as an expansion channel of visible images, and infrared and visible images

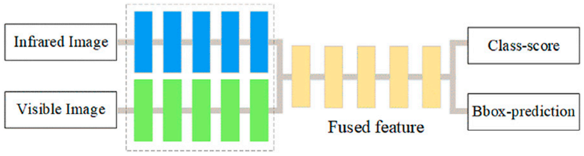

FIGURE 3
The concept of the halfway fusion framework, an alternative approach to integrating information from different sensor modalities.

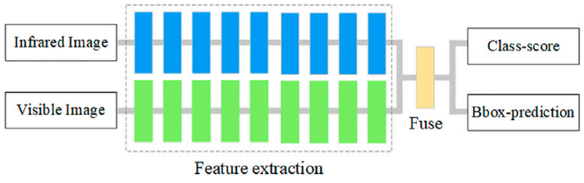

FIGURE 4
Late fusion framework, showcasing the sequential flow of information from the individual sensor modalities through their respective processing modules.

are input to the target detection network after cascading in the channel dimension, and feature fusion is achieved when the network extracts features, and the network structure does not need to be changed, which can be represented as Figure 2. Secondly, the visible image has three channels and the infrared image has only one effective channel, which is directly cascaded in the channel dimension as network input and the features obtained during the convolution calculation are not balanced. In addition, pre-trained models are generally used to initialize the weights during model training, and almost all pre-trained models are trained on the visible dataset, which has a weak representation of infrared features and cannot make full use of infrared features.

Halfway fusion is a dual-stream architecture that first extracts the features of IR and visible images separately, and then fuses the dual-band features in the middle of the network before further extracting the high-level semantic information of the fused features and finally making decisions, as shown in Figure 3. The dual-stream architecture is more flexible, does not require high resolution consistency of the input images, and the mid-level features balance the strengths and weaknesses of the bottom and top-level features, preventing the network from focusing too much on extracting detailed information or abstract features, and making it easier to learn the correlation between IR and visible features.

Post-fusion, which can also be called decision-level fusion, usually uses two identical sub-networks to extract the features of IR and visible images separately, and fuses the high-level semantic features to get the detection results before the final decision, as shown in the framework diagram in Figure 4. The architecture of the post-fusion method is flexible, the features in the fusion stage are more abstract, and the requirements for the input data form and target form are low, e.g., it can be a point-plane target, but the fusion effect almost depends on the feature expression ability of a single sub-network, and the detection effect of dual-band fusion cannot be enhanced by complementary information.

The algorithm in this paper uses the FLIR dataset for algorithm validation, which is dominated by face targets for infrared and visible targets, and most of the images are aligned at pixel level, and a small portion of the images are not aligned, but within an acceptable range. The existence of alignment errors in the dataset will have a certain impact on the effect of the pre-fusion algorithm. Taking into account the form of the data, the advantages and disadvantages of each stage of fusion, this paper selects the medium-term fusion method as the theoretical basis to build a dual-stream architecture for dual-band target detection.

2.2 Model framework

As shown in Figure 5., our approach is consisted of a dual-stream architecture, ex-tracting infrared and visible image features, and integrating multimodal features in the middle of the model using the feature fusion module based on the linear transformer and channel attention proposed in this paper, which is passed to the single-stream network to enhance cross-modal feature interaction. The model structure is shown in Figure 5. The model can be divided into P_1, P_2 and P_3 phases in the feature extraction part, outputting three scales of infrared and visible features. The P_2 stage includes a convolution module and nine tandem C3 modules; the P_3 stage contains a convolution module, an SPP module and three tandem C3 modules. After acquiring the IR and visible features in the P_1, P_2 and P_3 phases, the proposed fusion module is used to integrate the features and pass them into the unimodal network to achieve cross-modal feature interaction. The artificial neural network (ANN) has gained significant popularity as a versatile computational model with diverse applications. One such application involves the utilization of a clustering network-based intelligent power line inspection system, which has been extensively explored by researchers. Additionally, the investigation of the bifurcation phenomena pertaining to optimal solutions in constrained optimization problems has been undertaken, with a particular focus on its implications within the field of mathematics [33, 34].

The core modules are the focus module, the C3 module, the convolution module, the SPP module and the proposed fusion module. The focus module slices the input image to achieve lossless 2-fold image down sampling. The convolutional module consists of a convolutional layer, a batch normalization layer and a SiLU activation layer and is the basic structural unit of the convolutional neural network. The C3 module establishes a parallel two-branch structure, one consisting of a convolutional module and a bottleneck module, and the other with only a convolutional module, which mainly implements deep feature learning, deepens the network and is less prone to gradient disappearance. The SPP module maximizes the pooling of the feature maps using four different scales from $\{1 \times 1, 5 \times 5, 9 \times 9, 13 \times 13\}$ and then convolves the different scales by channel dimension concatenation, which increases the perceptual field of the network compared to that of single-scale max pooling. The fusion module integrates infrared and visible features, introduces global information, increases the perceptual field, models the spatial location correlation, and recalibrates the importance of channel features. The importance of channel features is recalibrated

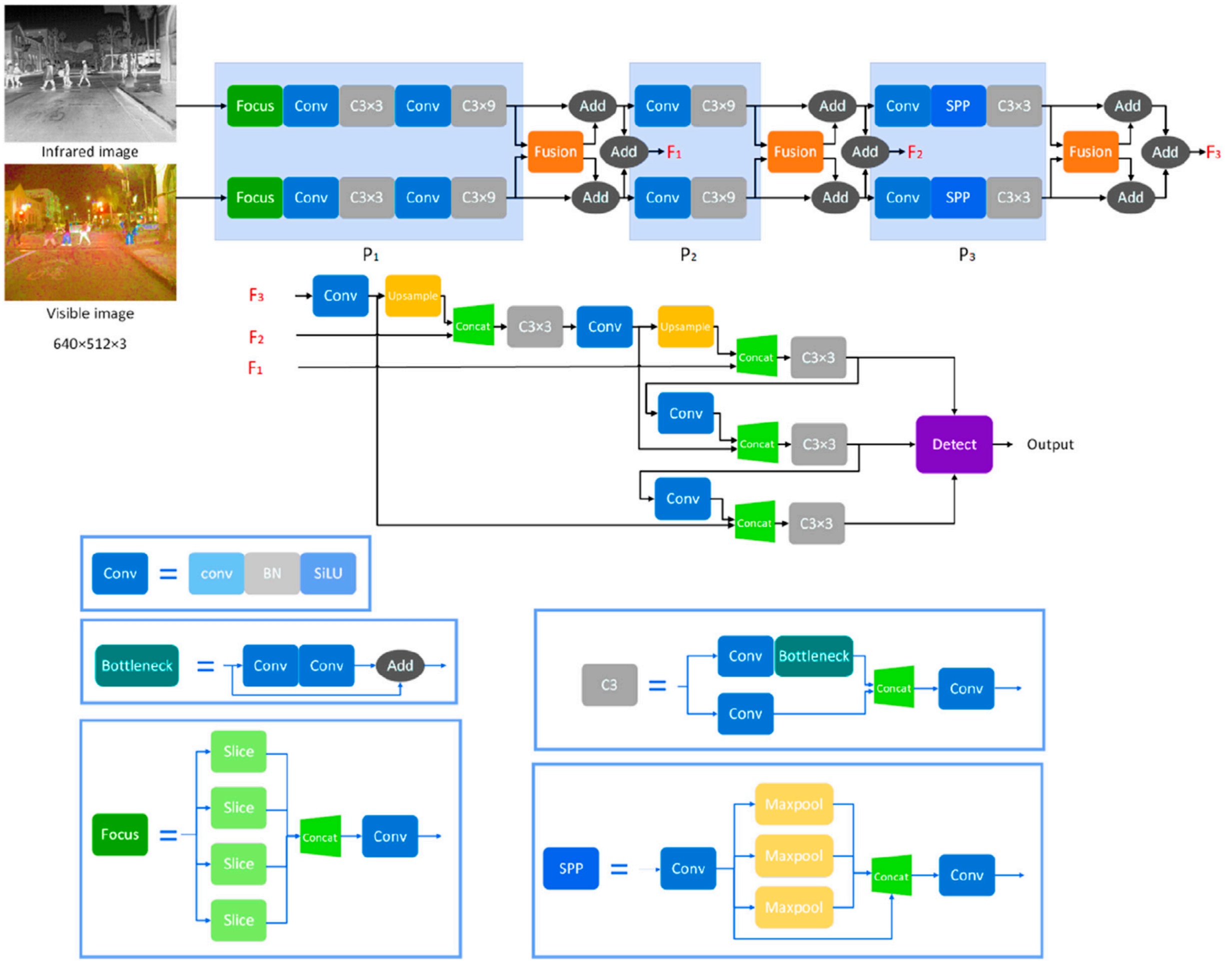
Infrared image
Visible image
640×512×3
Focus
Conv
C3×3
Conv
C3×9
Add
Fusion
Add
F₁
P₁
P₂
P₃
SPP
F₂
F₃
Upsample
Concat
Detect
Output
Conv = conv BN SiLU
Bottleneck = Conv Conv Add
C3 = Conv Bottleneck Conv Concat Conv
Focus = Slice Slice Slice Slice Concat Conv
SPP = Conv Maxpool Maxpool Maxpool Concat Conv

FIGURE 5
Dual-stream architecture with feature fusion module.

to enhance useful channel features while suppressing irrelevant features.

The conducted investigations demonstrate the extensive scope of research topics within the field of monitoring and image processing technologies [35–39]. They encompass various techniques, such as removing reflections from metallic objects while preserving optical conditions, eliminating highlights from black-and-white images using precision GANs, predicting viewports in 360-degree video multicast, optimizing volumetric video streams, developing fusion networks for traffic detection, detecting infrared-visible objects, and tracking maneuvering targets [40–44]. The investigations also delve into the effects of spatial scale on layout learning and drainage behavior, semantic and sample segmentation in coastal urban areas, achieving human-like assembly through variable acceptance control, identifying and tracking drainage pipeline defects, analyzing anomalies in sensor data for autonomous vehicles, developing autonomous pipeline navigation for bio-robots, utilizing structured illumination microscopy, studying flame retardants, precise multi-view stereo reconstruction, stitching 3D point clouds for aero-engine blade measurement, imitating tool-based clothing folding through visual observations, developing a path planning method that emulates human-like behavior, and predicting individual future paths by considering temporal and spatial intervals [45–49]. These resources exemplify a diverse range of research efforts across various fields [50–54].

2.3 Feature fusion module

After extracting the dual-band features separately through the YOLOv5 network, IR and visible feature interactions are enhanced by the proposed linear transformer and channel attention-based feature fusion module. The feature fusion module consists of a fusion transformer module and a fusion attention module, and the structure is shown in Figure 6. The infrared features, which are assumed to be $F_{IR} \in \mathbb{R}^{H \times W \times C}$, and the visible features, which are assumed to be $F_{Vis} \in \mathbb{R}^{H \times W \times C}$, are first spatial position encoded (positional encoding) [18]. In this paper, a sine and cosine function is used to represent the relative position of the spatial coordinates. Assuming that the feature map channel dimension is C, the position encoding of $d_{model} = C$ and the specific spatial location p is given by Eq. 1.

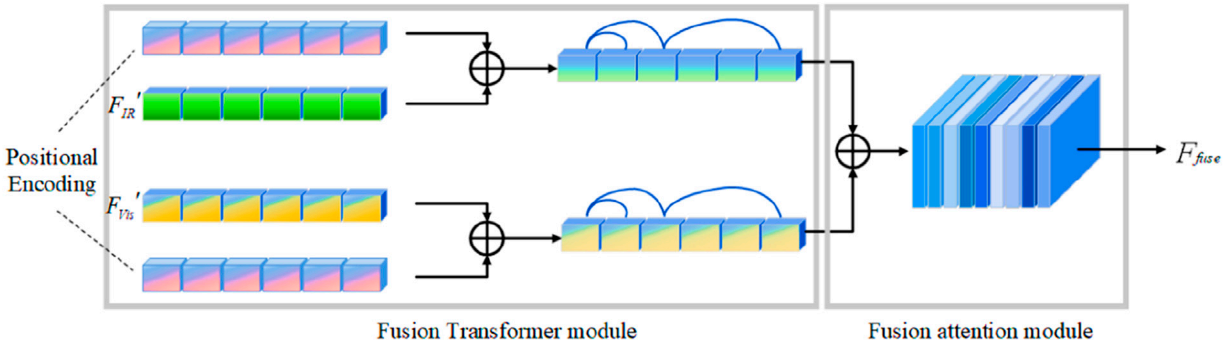

FIGURE 6
Schematic diagram of the feature fusion module.

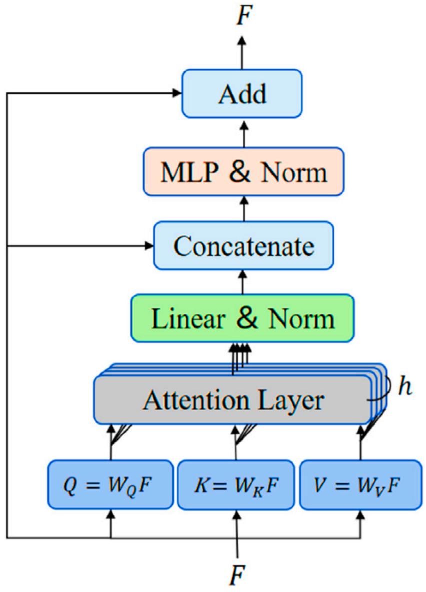

FIGURE 7
Encoder layer, Structure of the Transformer Encoder Layers.

$$\begin{cases} PE(4i, p) = \sin\left(\frac{p_x}{10000^{2i/d_{model}}}\right) \\ PE(4i+1, p) = \cos\left(\frac{p_x}{10000^{2i/d_{model}}}\right) \\ PE(4i+2, p) = \sin\left(\frac{p_y}{10000^{2i/d_{model}}}\right) \\ PE(4i+3, p) = \cos\left(\frac{p_y}{10000^{2i/d_{model}}}\right) \end{cases} \tag{1}$$

where $i = [0, 2, 4, ..., d_{model}/2]$ and p_x, p_y are the spatial positions of the points p in the feature map. The position codes are added to the infrared feature F_{IR}' and the visible feature F_{Vis}' to obtain the features F_{IR-pos} and $F_{Vis-pos}$ that introduce the position information.

The transformer encoder is made up of sequentially connected encoder layers and differs from convolutional neural networks in that it does not use convolutional kernels for feature extraction. The structure of the encoder layers in this paper is shown in Figure 7. The input features F are mapped into query vectors Q, key vectors K and value vectors V using matrices W_Q, W_K and W_V, respectively. The core of the encoder layer lies in the computation of the attention layer. The most basic attention layer is based on the query vector Q and the key vector K dot product. Then, the value vector V dot product is used to obtain the attention weight, and the self-attention SA can be expressed as Eq. 2.

$$SA(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right)V \tag{2}$$

where D is the feature dimension and $\frac{1}{\sqrt{D}}$ is the scaling factor to prevent the gradient of the softmax activation function from converging to zero due to the large values obtained by the matrix dot product. However, in the transformer, the inner product calculation is very resource intensive, and in this paper, a linear transformer model is used [19]. Replacing the exponential kernel of the softmax function with $\phi(Q) \cdot \phi(K)^T$, where $\phi(\cdot) = \text{elu}(\cdot) + 1$, the computational complexity is reduced from $O(N^2)$ to $O(N)$ while the performance remains almost constant. After the attention layer computation, a linear function (Linear) is used for the mapping, followed by layer normalization to normalize the individual samples and link them to the input features F. The network is then deepened by a multilayer perceptron (MLP) and layer normalization, and the output features are summed with the input F. The features F_{IR-pos} and $F_{Vis-pos}$ are input to the transformer encoder layer for computation to obtain F'_{IR-pos} and $F'_{Vis-pos}$, respectively, introducing the global information of the unimodal features.

The fused attention module focuses on using channel attention to correct for the importance of channel dimensions, enhancing detection task specificity features and suppressing irrelevant features. The output of the encoder layers F'_{IR-pos} and $F'_{Vis-pos}$ are summed over the feature elements to obtain F'_{fuse}. First, the vector of $C \times 1 \times 1$ is obtained by the global average pooling of F'_{fuse}. Next, a nonlinear mapping is performed using the fully connected layer and the ReLU activation function. Then, the size of $1 \times 1 \times C$ is generated using the fully connected layer and the Sigmoid function, and the value is normalized between $[0, 1]$ to obtain the vector of channel weights S and finally multiplied by F'_{fuse} to obtain $\widehat{F}'_{fuse}$, achieving a flexible correction of the fused features. The process can be represented by Eq. 3.

$$F_{fuse} = \varphi\left(W_2\left(\phi\left(W_1\left(GAP\left(F'_{fuse}\right)\right)\right)\right)\right)\cdot F'_{fuse} \tag{3}$$

where GAP denotes global average pooling, W_1 and W_2 denote the fully connected layer weights, ϕ denotes the ReLU activation function, and φ denotes the Sigmoid activation function. Finally, the computed fused features F_{fuse} are passed to the unimodal network for cross-modal feature interaction.

2.4 Loss function

We use coupled loss function to optimized feature fusion facilitating target perception, which consists of a confidence loss, a classification loss and a localization loss. The confidence loss $\mathcal{L}_{obj}$ is used to determine the probability that a target exists within the bounding box of the regression, the classification loss $\mathcal{L}_{class}$ optimizes the category prediction task, and the localization loss $\mathcal{L}_{bbox}$ is used as the loss for the bounding box regression. Both the confidence loss and the categorical loss are binary cross-entropy losses. The confidence loss $\mathcal{L}_{obj}$ can be expressed as Eq. 4.

$$\mathcal{L}_{obj} = \sum_{i=0}^{K^2}\sum_{j=0}^{M}\left[\widehat{A}_i^j \log\left(A_i^j\right) + \left(1-\widehat{A}_i^j\right)\log\left(1-A_i^j\right)\right] \tag{4}$$

$$A_i^j = V_{i,j} \times IoU_{pre}^{gt} \tag{5}$$

In Eq. 4, K represents the number of grids, M represents the number of candidate boxes for each grid, and $\widehat{A}_i^j$ and A_i^j represent the prediction confidence and the true confidence for i, j candidate boxes, respectively. In Eq. 5, $V_{i,j}$ is one when the grid is i and the candidate box is the target in j and 0 when $V_{i,j}$ is the opposite. IoU_{pre}^{gt} represents the intersection ratio of the predicted box to the true box. The classification loss $\mathcal{L}_{class}$ can be expressed as in Eq. 6.

$$\mathcal{L}_{class} = \sum_{i=0}^{K^2} V_{i,j}^{obj} \sum_{C\epsilon class}\left[\widehat{p}_i(c)\log\left(p_i(c)\right) + \left(1-\widehat{p}_i(c)\right)\log\left(1-p_i(c)\right)\right] \tag{6}$$

where $p_i(c)$ and $\widehat{p}_i(c)$ indicate the true and predicted probabilities of whether the target in the i grid is in category c, respectively, C indicates the total number of categories, and K indicates the number of grids. $V_{i,j}^{obj}$ is one when the anchor box at (i, j) contains a target and 0 otherwise. The localization loss $\mathcal{L}_{bbox}$ uses the complete IoU (CIoU) loss to optimize the detection frame regression task. In addition to considering the cross-merge ratio IoU and centroid distance, the CIoU loss is added to the distance IoU (DIoU) [20] loss function with the addition of an influence factor u to measure the consistency of the relative proportions of the two rectangular boxes, as in Eq. 7.

$$\mathcal{L}_{bbox} = \mathcal{L}_{CIoU}^{=} = 1 - IoU + \frac{\rho^2\left(b_{det} - b_{gt}\right)}{d^2} + \alpha u \tag{7}$$

where b_{det} denotes the centroid of the prediction box, b_{gt} denotes the centroid of the label, ρ denotes the Euclidean distance between the two centroids, d denotes the diagonal distance between the prediction box and the concatenation of the label boxes, and u and α are calculated as shown in Eqs 8, 9.

$$u = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{8}$$

$$\alpha = \frac{u}{(1 - IoU) + u} \tag{9}$$

where w^{gt}, h^{gt}, w and h are the lengths and heights of the label and prediction frames, respectively. Finally, the hyperparameters λ_{obj}, λ_{class} and λ_{box} are introduced to balance the partial loss functions, and the final loss function for the target detection $\mathcal{L}_{detection}$ can be expressed as Eq. 10.

$$\mathcal{L}_{detection} = \lambda_{obj}\mathcal{L}_{obj} + \lambda_{class}\mathcal{L}_{class} + \lambda_{bbox}\mathcal{L}_{bbox} \tag{10}$$

2.5 Different feature fusion models

To investigate the effect of cross-modal interactions on detection, we use a summation approach to fuse IR and visible features while retaining the base framework, and the framework of the different feature interaction models is shown in Figure 8. Fuse model-1 is the basic framework in this paper and uses a dual-stream architecture to extract IR and visible features separately, introducing features extracted by another subnetwork at the connection of P_1 and P_2 and P_2 and P_3, respectively, as shown in Figure 8A.

The model structure is shown in Figure 8B. Fuse model-3 extracts the features of the source image in the P_1 part of the network and then shares the fused features with the P_2 and P_3 parts of the network to further extract the deeper features, as shown in Figure 8C.

3 Experiments and results

3.1 Dataset with experimental details

In this paper, we use the FLIR public dataset to train and validate the algorithm model. The FLIR dataset is an autonomous driving scene dataset containing road target images captured under complex daytime and nighttime conditions, with 4,129 pairs of infrared and visible images in the training set and 1,010 pairs in the test set. In this paper, three types of targets, namely, pedestrians (person), vehicles (car) and bicycles (bicycle), were selected, and the target attributes of the training set were counted, as shown in Figure 9. Figure 9 shows the (a) target number distribution, (b) target centroid distribution, and (c) target size distribution. Figure 9 shows the number of targets in the three categories varies widely and the categories are seriously unbalanced, with vehicle targets far outnumbering pedestrian and bicycle targets. The statistics on the distribution of target centroid (x, y) coordinates show that the y-coordinates of most target centers are approximately 0.5, and the x-coordinates are between 0.1 and 0.7, which is in line with the characteristics of autonomous driving scenarios; the height and width of most targets are mainly within 0.08, i.e., the target size is within 50 × 40 pixels with predominantly face targets.

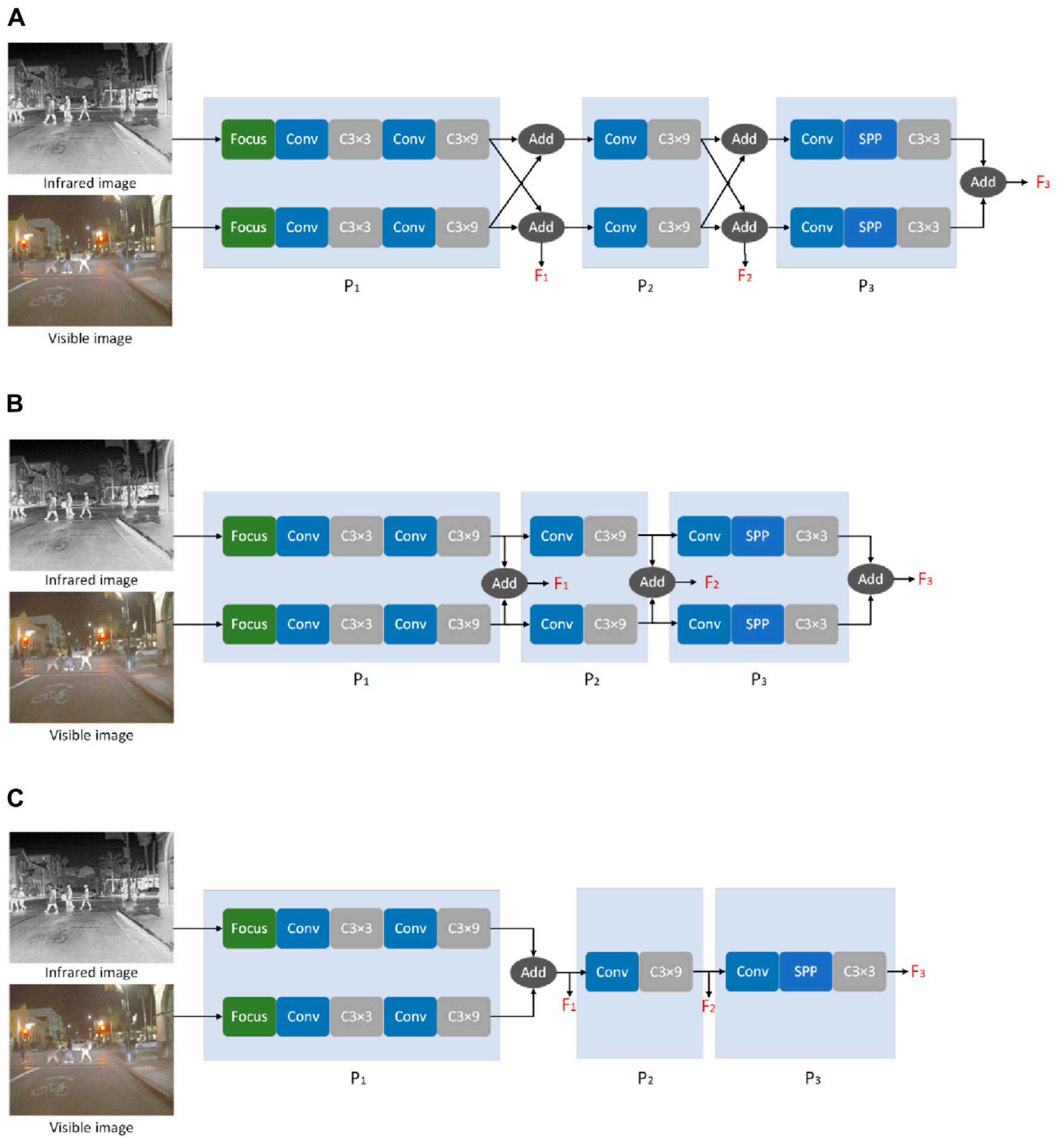

FIGURE 8
Frameworks for investigating cross-modal interactions. **(A)** Subnetwork at connections, **(B)** model structure, **(C)** network to further extracted the deeper features.

The study conducted an analysis of distance statistics to investigate the distinctness and complementary nature of different modalities in capturing information about pedestrians. The researchers focused on comparing the distances between pedestrian samples in the visible range and infrared. The results demonstrated significant differences between the distances observed in the visible range alone and those obtained in the infrared modality. This indicates that the two modalities capture unique information about pedestrians, with the visible range emphasizing visual appearance and shape, while the infrared modality highlights thermal signatures and motion patterns. Furthermore, the study examined the cross-distance between the fused representation and the individual modalities. The findings revealed that the cross-distance was significantly smaller compared to the distances within each modality. This indicates that the fused representation successfully integrates the distinct information from both modalities, resulting in an embedding that is closer to each individual modality. Before model training, in addition to conventional data enhancement methods such as the random cropping, horizontal flipping, vertical flipping and proportional scaling of images, the mosaic data enhancement method was used to randomly stitch 4 images to prevent overfitting. The experimental environment used an I9-109000k CPU, 32 GB RAM, NVIDIA 3090 TI video card and 24 GB video memory. The optimizer used in YOLOv5 is stochastic gradient descent

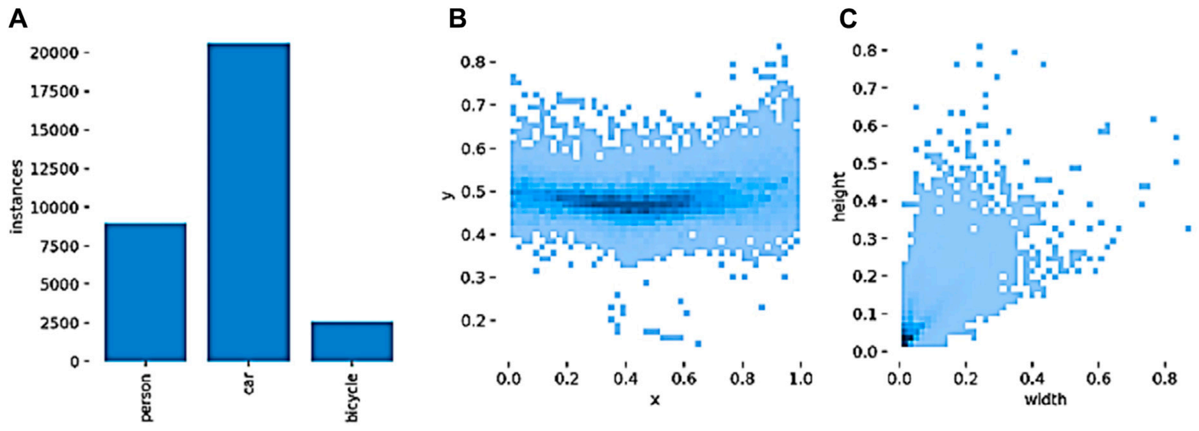

FIGURE 9
(A) Object attribute statistics, **(B)** Target Distribution and **(C)** Attributes in the FLIR Dataset.

(SGD) [21]. In YOLOv5, the default parameters were used, and anchor box sizes of [10, 16, 30, 55] were used as the detection frames for small, medium and large targets. The values of λ_{obj}, λ_{class} and λ_{bbox} in the loss function were 1, 0.5 and 0.05, respectively.

3.2 Evaluation indicators

In this paper, we used the mean average precision (mAP) to evaluate the detection results, and the definition is described below. The intersection over union (IoU) refers to the ratio of the intersection of the real frame and the predicted frame to the merged set.

$$IoU = \frac{A_p \cap A_{gt}}{A_p \cup A_{gt}} \tag{11}$$

For the target detection problem, suppose the threshold is τ. When $IoU \geq \tau$, if a target is identified and classified correctly, it is considered a positive case; otherwise, it is a negative case. Thus, according to the true result and the predicted result, the following cases can be classified: true positive (TP), true positive case and correct prediction; false-positive (FP), true negative case but predicted positive case; true negative (TN), true negative case and correct prediction; and false-negative (FN), true positive case but predicted negative case. Precision and recall are calculated as in Eqs 12, 13, respectively.

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

When calculating the metrics, the IoU is fixed at a threshold of τ, and the class with the highest probability is taken for category discrimination. Other parameters remain unchanged, the prediction results are arranged in descending order of confidence, and different thresholds are taken to calculate the precision-recall curve. The area under the calculated P-R curve is the average precision (AP), and mAP is the average AP value for all categories. The two main indicators involved in this paper are mAP50 and mAP. mAP50 is the average of all AP categories when the IoU threshold τ is 0.5. mAP uses the definition of COCO, calculates the mAP value for each interval of 0.05 between the IoU threshold τ and $[0.5, \ 0.95]$, and finally takes the average value to obtain the final mAP. The formula is shown in Eq. 14.

$$mAP = \frac{mAP_{0.50} + mAP_{0.55} + \ldots + mAP_{0.95}}{10} \tag{14}$$

where $mAP_{0.50}$ is the average accuracy at the IoU threshold τ and so on.

3.3 Experimental analysis

To illustrate the difference between human perception and machine perception, we give the multi-scale features extracted by the algorithm and their corresponding fusion results, and the visualization results of the features for machine perception are shown in the Figure 10. In the low-dimensional feature extraction with more detailed features, the higher-order features reflect the semantic features of the image. The fused features are generated from the multiscale features of both bands and finally help in image detection. In the model, the fused features are optimized by the loss function achieved by the detection, i.e., the features on which the detection depends, are completely unfavorable to human observation.

Figure 10 visually illustrates the multi-scale features and fused outcome achieved through our novel spatial perception method. The figure showcases the effectiveness of our approach in integrating dual-modality feature fusion and coupled attention mechanisms, resulting in enhanced target detection performance through the fusion of cross-modal information.

Figure 11 illustrates a qualitative comparison of dual-band object detection outcomes specifically in nighttime scenes, serving as a visual representation of the proposed spatial perception method's performance and efficacy in low-light conditions.

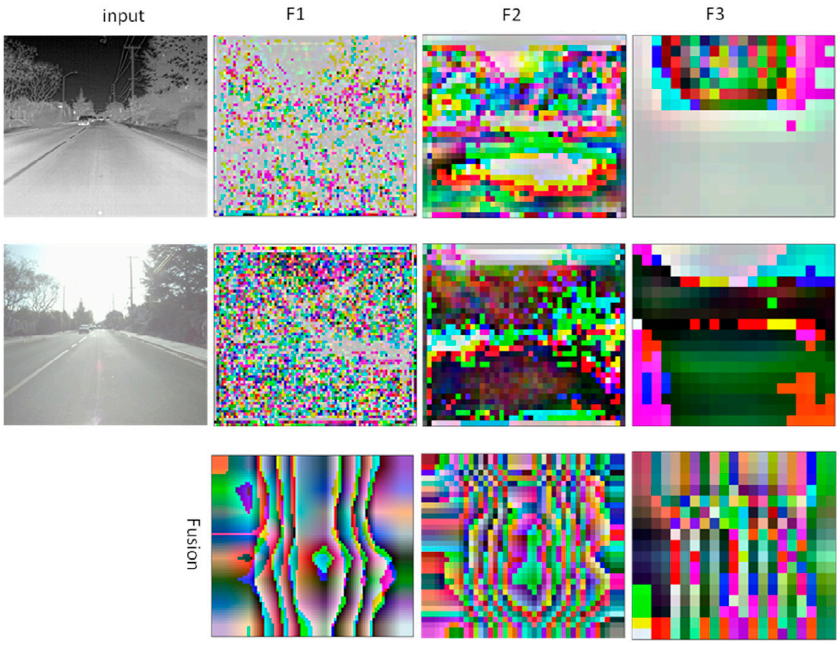

FIGURE 10
Visualization of multi-scale features and fused result.

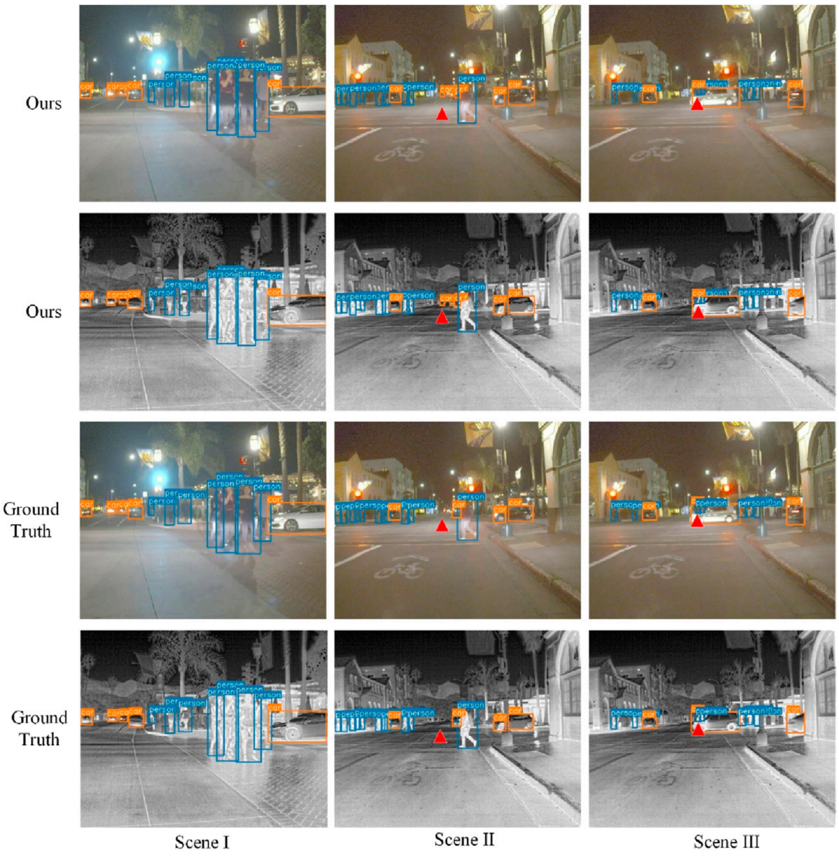

FIGURE 11
Qualitative comparison of dual-band object detection results in nighttime scenes.

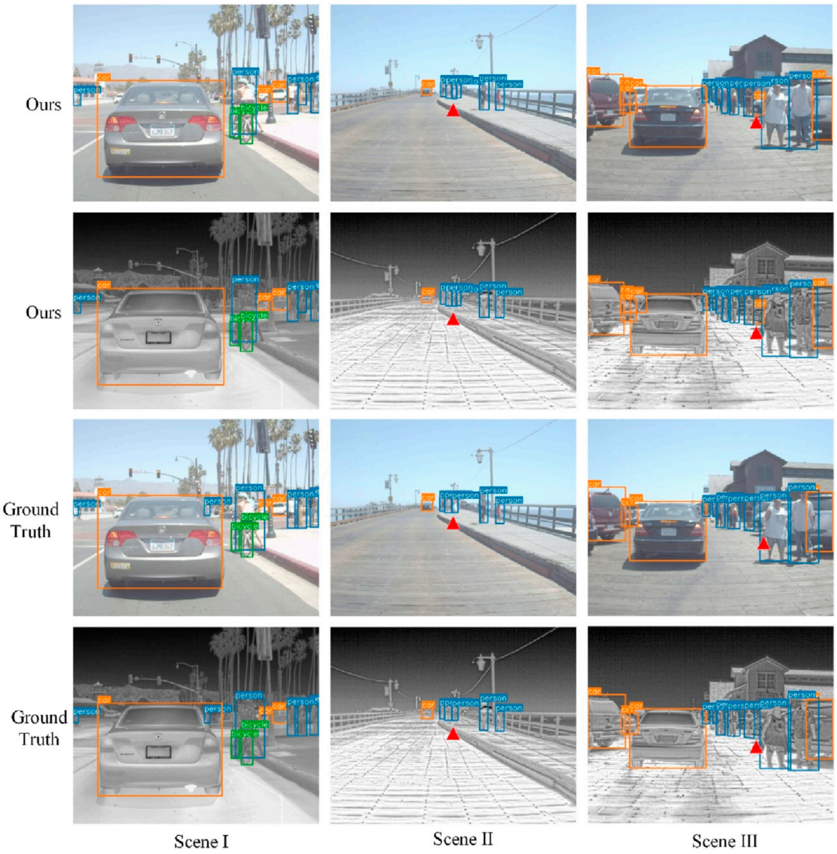

FIGURE 12
Qualitative comparison of dual-band object detection results in daytime scenes.

Figure 11 presents a side-by-side evaluation of the detection results achieved using the proposed algorithm, featuring a collection of nighttime scenes captured using dual-band modalities. Each scene showcases the highlighted and labeled detected objects, along with their respective classes and bounding boxes. Figure 12 compares dual-band object detection results in daytime scenes, highlighting the effectiveness of the proposed spatial perception method. Figure 12 shows a side-by-side analysis of detection outcomes using the algorithm, showcasing labeled objects in dual-band imagery. This visual comparison allows readers to assess the algorithm's performance in accurately identifying and localizing objects in well-lit conditions. The qualitative results support the quantitative findings, emphasizing the method's ability to enhance target detection. Figure 12 also demonstrates the method's adaptability to different lighting conditions, reinforcing its practical applicability in real-world scenarios.

The labels from the method in this paper were compared with the labels in the FLIR ADAS dataset, and the dual-band target detection effects during the day and at night are shown in Figure 11, 12, respectively. Figure 11 shows the detection effect of a night scene. Most targets, such as pedestrians and vehicles, cannot be easily detected by human eyes in the visible image, but they are more prominent in the infrared image. The algorithm in this paper completely detects pedestrians, vehicles and other targets in all three scenes. In Scene II, the algorithm in this paper detects two small vehicle targets that are not labeled (see red marks in the Figure 11), and in Scene III, it can detect most of the pedestrians that are obscured by vehicles. Figure 12 shows the detection effect of the algorithm in daytime scenes. The image quality of the visible image in daytime is significantly improved, while the prominence of the targets in the infrared image is slightly reduced, and the surrounding area of the targets is blurred. The results of this algorithm are more accurate than those based on manual annotation. Subjective observation shows that the algorithm is able to exploit the advantages of infrared and visible images. The detection performance is accurate and stable in daytime and nighttime scenes and is not easily affected by lighting and other factors.

TABLE 1 Quantitative comparison on the FLIR dataset.

Method	Data	Precision	Recall	mAP50	mAP
YOLOv5	RGB	79.4	60.9	67.3	31.9
YOLOv5	IR	80.9	72.5	78.6	40.4
GPT	RGB + IR	81.1	72.2	79.1	40.2
Ours	RGB + IR	81.4	73.2	80.0	41.6

To verify the effectiveness of the dual-band fusion target detection algorithm, the algorithm in this paper was compared with the single-band detection model. The dual-band target detection algorithm GPT [22], YOLOv5, and the model in this paper are all used as pretrained models. YOLOv5 is a unimodal

TABLE 2 Quantitative comparison of the day-night scenes from the FLIR dataset.

Method	Data	Daytime mAP50 mAP		Nighttime mAP50 mAP	
YOLOv5	RGB	72.7	35.1	56.6	23.3
YOLOv5	IR	77.6	40.2	84.2	43.8
GPT	RGB + IR	77.0	39.3	85.3	43.5
Ours	RGB + IR	78.1	40.6	84.9	44.2

TABLE 3 Ablation experiment of Feature fusion module.

Method	Data	Precision	Recall	mAP50	mAP
Fuse model-1	RGB + IR	80.0	67.9	76.3	39.5
Ours	RGB + IR	81.4	73.2	80.0	41.6

TABLE 4 Quantitative comparison of models at different fusion stages.

Method	Data	Precision	Recall	mAP50	mAP
Fuse model-1	RGB + IR	80.0	67.9	76.3	39.5
Fuse model-2	RGB + IR	78.6	72.0	77.6	40.6
Fuse model-3	RGB + IR	78.2	71.2	77.7	40.0

model trained on visible and infrared images together and tested on the visible and infrared images in the test set. The training set and parameters of the GPT algorithm and the algorithm in this paper are the same. In this section, the precision, recall, mAP50, mAP75 and mAP metrics were used for quantitative comparison, and the results are shown in Table 1.

From the Table, it can be obtained that compared with visible single-mode target detection, the algorithm in this paper improves 18.9% and 30.4% in mAP50 and mAP, respectively. Compared with the single-mode infrared detection results, the algorithm in this paper improves 1.8% and 3.0% in mAP50 and mAP metrics, respectively, indicating that multi-band fusion detection can improve the detection effect; compared with GPT multimodal detection algorithm, mAP50 and mAP improve 1.1% and 3.5%, respectively. In addition to the average accuracy, the algorithm in this paper balances up in the accuracy and recall metrics, indicating that the combined multi-band information can reduce the false detection rate.

To verify the effectiveness of the model in daytime and nighttime scenes, the FLIR test set was split into daytime and nighttime datasets and tested separately, with 700 pairs of images in the daytime dataset and 310 pairs of images in the nighttime dataset, and the test results are shown in Table 2. From this table, it can be seen that visible light images are more influenced by light factors, and targets in dark or glare conditions cannot be detected. Thus, the average accuracy of detection at night is significantly lower than that during the day. The contrast of infrared images in the FLIR dataset is relatively low in daytime and the edges are more blurred, so the average accuracy of detection in the infrared band at night is higher than that in daytime, which is in line with the characteristics of infrared and visible light imaging. In addition, because some of the images in the FLIR dataset are not aligned at the pixel level and the labels in the dataset are labeled according to the infrared images, there are problems with missing labels and bounding box shifts, resulting in the detection accuracy index of the visible band and the fusion algorithm being affected to some extent. Taken together, the algorithm proposed in this paper is able to perform in a stable manner for a variety of complex scenes, regardless of illumination.

To verify the impact of the proposed feature fusion module on target detection, the model was retrained by replacing the feature fusion module with a summation layer (see Figure 4A for the model structure) while keeping the training data, basic network model and parameters unchanged, and the test results on the FLIR test set are shown in Table 3. The model including the feature fusion module can improve 1.75%, 7.80%, 4.84% and 5.32% in Precision, Recall, mAP50 and mAP indexes, respectively, indicating that the feature fusion module proposed in this paper can effectively fuse important complementary features of IR and visible images and significantly improve the detection performance compared with simple summation for fusion.

This paper introduces three fusion models, namely, Fuse model-1, Fuse model-2, and Fuse model-3, designed to investigate the impact of different fusion methods on target detection accuracy. The experimental results, presented in Table 4, focus on the FLIR test set, with only the network model being modified while keeping other factors unchanged. The optimal values of the three models are indicated in bold. Fuse model-1 achieves the highest accuracy rate, Fuse model-2 exhibits the highest recall and average precision, and Fuse model-3 demonstrates the highest average precision at an Intersection over Union (IoU) threshold of 0.5.

Both Fuse model-2 and Fuse model-3 perform comparably. However, Fuse model-1 falls short, indicating a significant disparity between infrared and visible features. The simple feature summation fusion employed in the dual-stream architecture fails to achieve the desired effect of feature complementarity. To address this limitation, the proposed algorithm in this paper improves upon the framework of Fuse model-1, resulting in a significant enhancement in performance. These findings suggest that the proposed feature fusion module effectively integrates complementary information from infrared and visible bands, thereby enhancing target detection performance.

4 Conclusion

The target detection accuracy of visible or infrared single-band images depends on the imaging quality, and dual-band images can provide complementary information to enhance the stability and reliability of the detection system. To verify the effectiveness of the dual-band information in improving the detection of the target, this paper proposes a dual-band target detection framework based on feature fusion applied to infrared and visible target detection. In our method, infrared and visible features are extracted separately through a dual-stream architecture and integrated through a feature fusion module based on a linear transformer and channel attention. Taken together, the combination of cross-modal features significantly improves target detection accuracy and provides strong adaptability and robustness in both day and night scenes, exploiting the complementary advantages of visible and IR images. In addition, due to the adaptive nature of the fusion strategy, the algorithm in this paper is expected to be applicable to other types of multimodal remote sensing imaging detection tasks. The study introduces a dual-stream architecture to separately extract features from infrared and visible images. These features are then fused using a linear transformer and a channel attention-based fusion module. This adaptive fusion approach enables the learning of interrelationships between different modalities without the need for manually designed fusion rules. Experimental results on the FLIR dataset demonstrate that the proposed method significantly outperforms single-modality detection baselines, achieving an average accuracy improvement of over 30% compared to visible image detection. The algorithm also demonstrates robust performance in both day and night scenes. Through visualization and quantitative evaluation, this research validates that cross-modal data fusion can indeed enhance a system's perceptual capabilities for tasks such as target detection.

Data availability statement

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

Author contributions

NZ: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Writing–original draft. WZ: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Writing–original draft.

Funding

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

Conflict of interest

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

Publisher's note

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

References

1. Girshick R, Donahue J, Darrell T, Malik J, Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2014). p. 580–7.

2. Girshick R Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision (2015). p. 1440–8.

3. Ren S, He K, Girshick R, Sun J, Faster r-cnn: towards real-time object detection with region proposal networks. *Adv Neural Inf Process Syst* (2015) 28. doi:10.1109/TPAMI.2016.2577031

4. He K, Gkioxari G, Dollár P, Girshick R, Presents the front cover of the proceedings record. In: IEEE international conference on computer vision (2017). p. 2961–9.

5. 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 Proceedings, Part I 14; October 11–14, 2016; Amsterdam, The Netherlands. Springer International Publishing (2016). p. 21–37.

6. Redmon J, Divvala S, Girshick R, Farhadi A, You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016). p. 779–88.

7. Redmon J, Farhadi A, YOLO9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2017). p. 7263–71.

8. Lin TY, Goyal P, Girshick R, He K, Dollár P, Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision (2017). p. 2980–8.

9. 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.

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

11. Dollár P, Appel R, Belongie S, Perona P, Fast feature pyramids for object detection. *IEEE Trans pattern Anal machine intelligence* (2014) 36(8):1532–45. doi:10.1109/tpami.2014.2300479

12. Konig D, Adam M, Jarvers C, Layher G, Neumann H, Teutsch M, Fully convolutional region proposal networks for multispectral person detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops (2017). p. 49–56.

13. 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

14. 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

15. 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

16. Zhang H, Fromont E, Lefevre S, Avignon B, Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In: 2020 IEEE International conference on image processing (ICIP). IEEE 2020 (2020, October). p. 276–80.

17. 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.

18. 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. doi:10.48550/arXiv.1706.03762

19. Katharopoulos A, Vyas A, Pappas N, Fleuret F Transformers are rnns: Fast autoregressive transformers with linear attention. In: Proceedings of the 37th International Conference on Machine Learning, PMLR (2020). Vol. 119, p. 5156–5165.

20. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S Generalized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2019). p. 658–66.

21. Bottou L, Curtis FE, Nocedal J Optimization methods for large-scale machine learning. *SIAM Rev* (2018) 60(2):223–311. doi:10.1137/16m1080173

22. Qingyun F, Dapeng H, Zhaokui W *Cross-modality fusion transformer for multispectral object detection* (2021). Available from: https://arxiv.org/abs/2111.00273 (Accessed 2021).

23. Diwan T, Anirudh G, Tembhurne JV Object detection using YOLO: challenges, architectural successors, datasets and applications. *multimedia Tools Appl* (2023) 82(6): 9243–75. doi:10.1007/s11042-022-13644-y

24. Zhu L, Xie Z, Liu L, Tao B, Tao W Iou-uniform r-cnn: breaking through the limitations of rpn. *Pattern Recognition* (2021) 112:107816. doi:10.1016/j.patcog.2021.107816

25. Tahir NUA, Long Z, Zhang Z, Asim M, Elaffendi M PVswin-YOLOv8s: UAV-based pedestrian and vehicle detection for traffic management in smart cities using improved YOLOv8. *Drones* (2024) 8(3):84. doi:10.3390/drones8030084

26. Tahir NUA, Zhang Z, Asim M, Chen J, Elaffendi M Object detection in autonomous vehicles under adverse weather: a review of traditional and deep learning approaches. *Algorithms* (2024) 17(3):103. doi:10.3390/a17030103

27. Xuan H, Luo L, Zhang Z, Yang J, Yan Y Discriminative cross-modality attention network for temporal inconsistent audio-visual event localization. *IEEE Trans Image Process* (2021) 30:7878–88. doi:10.1109/tip.2021.3106814

28. Xie B, Milam G, Ning B, Cha J, Park CH DXM-TransFuse U-net: dual cross-modal transformer fusion U-net for automated nerve identification. *Comput Med Imaging Graphics* (2022) 99:102090. doi:10.1016/j.compmedimag.2022.102090

29. Liu X, Luo Y, Yan K, Chen J, Lei Z CMC2R: cross-modal collaborative contextual representation for RGBT tracking. *IET Image Process* (2022) 16(5):1500–10. doi:10.1049/ipr2.12427

30. Feng F, Ming Y, Hu N SSLNet: a network for cross-modal sound source localization in visual scenes. *Neurocomputing* (2022) 500:1052–62. doi:10.1016/j.neucom.2022.05.098

31. Cai Y, Sui X, Gu G, Chen Q Learning modality feature fusion via transformer for RGBT-tracking. *Infrared Phys Tech* (2023) 133:104819. doi:10.1016/j.infrared.2023.104819

32. Wang H, Song K, Huang L, Wen H, Yan Y Thermal images-aware guided early fusion network for cross-illumination RGB-T salient object detection. *Eng Appl Artif Intelligence* (2023) 118:105640. doi:10.1016/j.engappai.2022.105640

33. Lv XL, Chiang HD Visual clustering network-based intelligent power lines inspection system. *Eng Appl Artif Intelligence* (2024) 129:107572. doi:10.1016/j.engappai.2023.107572

34. Li T, Wang Z The bifurcation of constrained optimization optimal solutions and its applications. *AIMS Math* (2023) 8(5):12373–97. doi:10.3934/math.2023622

35. Chen J, Song Y, Li D, Lin X, Zhou S, Xu W Specular removal of industrial metal objects without changing lighting configuration. *IEEE Trans Ind Inform* (2023) 20:3144–53. doi:10.1109/tii.2023.3297613

36. Xu H, Li Q, Chen J Highlight removal from a single grayscale image using attentive GAN. *Appl Artif Intelligence* (2022) 36(1):1988441. doi:10.1080/08839514.2021.1988441

37. Li J, Han L, Zhang C, Li Q, Liu Z Spherical convolution empowered viewport prediction in 360 video multicast with limited FoV feedback. *ACM Trans Multimedia Comput Commun Appl* (2023) 19(1):1–23. doi:10.1145/3511603

38. Li J, Zhang C, Liu Z, Hong R, Hu H Optimal volumetric video streaming with hybrid saliency based tiling. *IEEE Trans Multimedia* (2022) 25:2939–53. doi:10.1109/tmm.2022.3153208

39. Chen J, Wang Q, Peng W, Xu H, Li X, Xu W Disparity-based multiscale fusion network for transportation detection. *IEEE Trans Intell Transportation Syst* (2022) 23(10):18855–63. doi:10.1109/tits.2022.3161977

40. Zhang R, Li L, Zhang Q, Zhang J, Xu L, Zhang B, et al. The effect of two facets of physicians' environmental stress on patients' compliance with COVID-19 guidelines: moderating roles of two types of ego network. *IEEE Trans Circuits Syst Video Tech* (2023) 1–25. doi:10.1080/08870446.2023.2295902

41. Di Y, Li R, Tian H, Guo J, Shi B, Wang Z, et al. A maneuvering target tracking based on fastIMM-extended Viterbi algorithm. *Neural Comput Appl* (2023) 1–10. doi:10.1007/s00521-023-09039-1

42. Zhu J, Dang P, Zhang J, Cao Y, Wu J, Li W, et al. The impact of spatial scale on layout learning and individual evacuation behavior in indoor fires: single-scale learning perspectives. *Int J Geographical Inf Sci* (2024) 38(1):77–99. doi:10.1080/13658816.2023.2271956

43. Zhang H, Liu H, Kim C Semantic and instance segmentation in coastal urban spatial perception: a multi-task learning framework with an attention mechanism. *Sustainability* (2024) 16(2):833. doi:10.3390/su16020833

44. Cao X, Huang X, Zhao Y, Sun Z, Li H, Jiang Z, et al. A method of human-like compliant assembly based on variable admittance control for space maintenance. *Cyborg Bionic Syst* (2023) 4:0046. doi:10.34133/cbsystems.0046

45. Ma D, Fang H, Wang N, Lu H, Matthews J, Zhang C Transformer-optimized generation, detection, and tracking network for images with drainage pipeline defects. *Computer-Aided Civil Infrastructure Eng* (2023) 38(15):2109–27. doi:10.1111/mice.12970

46. Zhao X, Fang Y, Min H, Wu X, Wang W, Teixeira R Potential sources of sensor data anomalies for autonomous vehicles: an overview from road vehicle safety perspective. *Expert Syst Appl* (2023) 121358. doi:10.1016/j.eswa.2023.121358

47. Ma S, Chen Y, Yang S, Liu S, Tang L, Li B, et al. The autonomous pipeline navigation of a cockroach bio-robot with enhanced walking stimuli. *Cyborg Bionic Syst* (2023) 4:0067. doi:10.34133/cbsystems.0067

48. Qian J, Cao Y, Bi Y, Wu H, Liu Y, Chen Q, et al. Structured illumination microscopy based on principal component analysis. *ELight* (2023) 3(1):4. doi:10.1186/s43593-022-00035-x

49. Jiang H, Xie Y, Zhu R, Luo Y, Sheng X, Xie D, et al. Construction of polyphosphazene-functionalized Ti3C2TX with high efficient flame retardancy for epoxy and its synergetic mechanisms. *Chem Eng J* (2023) 456:141049. doi:10.1016/j.cej.2022.141049

50. Shi Y, Xi J, Hu D, Cai Z, Xu K RayMVSNet++: learning ray-based 1D implicit fields for accurate multi-view stereo. *IEEE Trans Pattern Anal Machine Intelligence* (2023) 45:13666–82. doi:10.1109/tpami.2023.3296163

51. Dong Y, Xu B, Liao T, Yin C, Tan Z Application of local-feature-based 3-D point cloud stitching method of low-overlap point cloud to aero-engine blade measurement. *IEEE Trans Instrumentation Meas* (2023) 72:1–13. doi:10.1109/tim.2023.3309384

52. Zhou P, Qi J, Duan A, Huo S, Wu Z, Navarro-Alarcon D Imitating tool-based garment folding from a single visual observation using hand-object graph dynamics. *IEEE Trans Ind Inform* (2024) 20:6245–56. doi:10.1109/tii.2023.3342895

53. Zhao J, Song D, Zhu B, Sun Z, Han J, Sun Y A human-like trajectory planning method on a curve based on the driver preview mechanism. *IEEE Trans Intell Transportation Syst* (2023) 24:11682–98. doi:10.1109/tits.2023.3285430

54. Jiang Y, Yang Y, Xu Y, Wang E, Spatial-temporal interval aware individual future trajectory prediction. *IEEE Trans Knowledge Data Eng* (2023) 1–14. doi:10.1109/tkde.2023.3332929

55. Yang Y, Shang X, Li B, Ji H, Lang Y, Detection-free cross-modal retrieval for person identification using videos and radar spectrograms. *IEEE Trans Instrumentation Meas* (2024) 73:1–12. doi:10.1109/tim.2024.3372210