IET Cyber-Systems and Robotics

WILEY

ZHEJIANG UNIVERSITY PRESS | IET The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

RMF-ED: Real-Time Multimodal Fusion for Enhanced Target Detection in Low-Light Environments

Yuhong Wu[1,2] | Jinkai Cui[3] | Kuoye Niu[2] | Yanlong Lu[2] | Lijun Cheng[3] | Shengze Cai[2] | Chao Xu[2]

[1]School of Design, Hong Kong Polytechnic University, Hong Kong, China | [2]College of Control Science and Engineering, Zhejiang University, Hangzhou, China | [3]Sichuan Aerospace System Engineering Institute, Chengdu, China

Correspondence: Kuoye Niu (kuoyeniu@foxmail.com)

Received: 7 January 2025 | **Revised:** 21 February 2025 | **Accepted:** 9 March 2025

Funding: This work was supported by the Zhejiang Province Key Research and Development Program (No. 2024C03276(SD2)).

Keywords: convolutional neural network (CNN) | deep learning | unmanned aerial vehicle (UAV) | visual surveillance

ABSTRACT

Accurate target detection in low-light environments is crucial for unmanned aerial vehicles (UAVs) and autonomous driving applications. In this study, the authors introduce a real-time multimodal fusion for enhanced detection (RMF-ED), a novel framework designed to overcome the limitations of low-light target detection. By leveraging the complementary capabilities of near-infrared (NIR) cameras and light detection and ranging (LiDAR) sensors, RMF-ED enhances detection performance. An advanced NIR generative adversarial network (NIR-GAN) model was developed to address the lack of annotated NIR datasets, integrating structural similarity index measure (SSIM) loss and L1 loss functions. This approach enables the generation of high-quality NIR images from RGB datasets, bridging a critical gap in training data. Furthermore, the multimodal fusion algorithm integrates RGB images, NIR images, and LiDAR point clouds, ensuring consistency and accuracy in proposal fusion. Experimental results on the KITTI dataset demonstrate that RMF-ED achieves performance comparable to or exceeding state-of-the-art fusion algorithms, with a computational time of only 21 ms. These features make RMF-ED an efficient and versatile solution for real-time applications in low-light environments.

1 | Introduction

Target detection is a fundamental task in computer vision, involving identifying and localising objects of interest within an image. This task poses significant challenges due to objects' varying appearances, shapes, and postures, compounded by issues such as illumination and occlusion during imaging. Traditional methods relying on manual features have given way to deep learning techniques, particularly convolutional neural network (CNN), which have demonstrated superior capabilities in feature learning and representation [1–4].

In recent years, advancements in deep learning technologies, coupled with the growing availability of large-scale datasets, have substantially enhanced target detection accuracy. With the increasing computational capabilities of hardware platforms, target detection has become a pivotal research area, especially in autonomous driving and other mobile applications where precise and efficient detection is paramount [5, 6]. In such areas, the growing demand for long-term tasks that span both day and night and emergency searches in entirely dark environments has directed researchers' attention to target detection in low-illumination environments. Performing visual tasks in such conditions is challenging due to insufficient colour features in short-exposure images and noise introduced by brightness enhancement. Long-exposure images also suffer from motion blur, further complicating the task.

Researchers have increasingly turned to infrared (IR) images to address these issues. A commonly used kind of IR image is also called thermal images, which often fall short in accuracy and precision because of their low resolution and lack of contrast. Researchers have increasingly turned to IR imaging to address these issues. IR sensors detect thermal radiation emitted by objects, converting temperature variations into visible contrast through photonic detectors—a passive imaging modality particularly effective in complete darkness. A commonly used kind of IR image is also called thermal images, which often fall short in accuracy and precision because of their low resolution and lack of contrast. Notably, near-infrared (NIR) systems utilising active illumination can overcome these limitations by capturing reflected light patterns, achieving enhanced spatial resolution while preserving thermal sensitivity advantages.

Light detection and ranging (LiDAR) is another commonly used sensor in detection tasks. It is capable of operating independently of illumination conditions. LiDAR creates detailed 3D maps of the surroundings, allowing precise detection of objects such as vehicles, pedestrians, and other entities. Despite its advantages, LiDAR's accuracy can be compromised by adverse weather conditions such as rain and fog, and it generally exhibits lower detection accuracy for small targets such as pedestrians due to its relatively weak resolution.

To overcome the limitations of single-sensor systems, multi-sensor fusion, also known as multi-source heterogeneous information fusion (MSHIF), has emerged as an effective solution for low-illumination detection tasks [7]. MSHIF integrates data from different sensors, avoiding the perception limitations and uncertainties inherent in single-sensor systems. This comprehensive approach enhances the system's external perception capabilities. However, current multi-sensor fusion algorithms typically cannot meet the standards of 30 FPS (frames per second) rates while entirely using each information source to obtain more assured predictions.

This research proposes a multimodal fusion approach that utilises RGB, NIR, and LiDAR data. Such a real-time multimodal fusion for enhanced detection (RMF-ED) framework addresses these challenges through innovative data augmentation and integration strategies, ensuring accurate and efficient detection under challenging illumination conditions.

As shown in Figure 1, the RMF-ED framework integrates LiDAR and image branches through a multimodal fusion architecture, enabling robust detection in spatial and depth dimensions. The main contributions of this work are as follows:

- **NIR-GAN**: We trained an improved generative adversarial network (GAN) to generate high-quality NIR images from their RGB versions by introducing the structural similarity index measure (SSIM) loss function, which better preserves image structures. This allows for easily converting popular object detection datasets into NIR images for training.
- **Learning-based Late Fusion System**: We propose a late-fusion strategy to exploit proposals from the LiDAR and camera information branches through consistency check, information fusion, and model reasoning. Apart from refined targets, improved bounding box predictions and confidence scores are also obtained to meet more accurate performances.
- **Low-illumination Detection**: By effectively leveraging the advantages of NIR cameras and LiDAR, this work achieves satisfactory performance in low-illumination conditions, showing great potential to support nighttime visual tasks.
- **Speed and Versatility**: Our algorithm reaches real-time detection performance with only 5 ms of additional inference time and minimal additional memory, making it a suitable choice for deployment on edge computing devices.

Through these contributions, this work demonstrates the potential for achieving real-time, robust target detection in low-

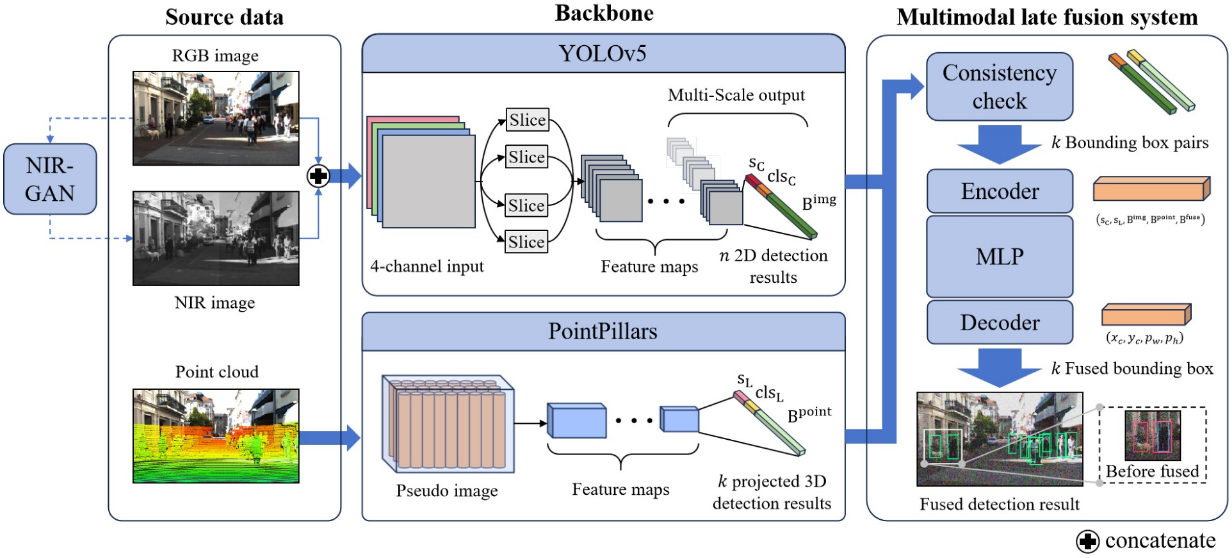

FIGURE 1 | Schematics of the proposed multimodal fusion model.

illumination environments, addressing a critical need in autonomous driving and other applications requiring reliable nighttime operation.

2 | Related Works

2.1 | Visual Detection

Low-illumination environments pose significant visual detection challenges due to low brightness, contrast, and increased image noise. Despite advancements in deep learning, state-of-the-art object detectors often underperform in these conditions. This is primarily due to the dim edges, low-contrast details in low-light images, and reduced divergence in the red, green, and blue channels.

Research addressing low-light environments primarily emphasises image enhancement techniques that improve visual quality. Classic methods such as adaptive histogram equalisation [8] and CNN-based approaches with bright channel priors [9] laid the foundation for low-light enhancement. For instance, Cai et al. [10] introduced Retinexformer, leveraging decomposition and illumination adjustment based on Retinex theory [11]. Similarly, Jiang et al. [12] developed EnlightenGAN, while deep convolutional networks like MSR-Net [13] further advanced the field. However, these methods often overlook detailed feature recovery and are computationally intensive, limiting their utility for complex downstream tasks such as object detection.

Updating camera equipment is another natural solution. Passive IR or thermal imaging distinguishes objects by their heat signatures. Due to the different image features and low resolution, target detection tasks based on thermal images often fail to achieve satisfactory accuracy and precision. NIR cameras, which work similarly to RGB cameras in the near and short-wave IR bands, offer higher resolution and detailed texture information. They have been widely used in illumination-invariant face recognition [14], 3D reconstruction [15], and scene parsing [16]. In these researches, NIR cameras compensate for missing details in RGB images under low-illumination environments.

2.2 | LiDAR Detection

LiDAR technology provides distinct advantages in low-illumination environments. Unlike cameras, LiDAR is unaffected by lighting conditions and captures data in a privacy-friendly manner, making it suitable for complex outdoor environments [17]. LiDAR sensors create detailed 3D maps of surroundings, enabling precise detection of objects such as vehicles and pedestrians.

LiDAR object detection algorithms can be categorised into point-based, voxel-based, and projection-based methods. Like the one proposed by Charles et al. [18], point-based methods directly process raw 3D position and reflection data. However, the quaternion representation of points can lead to redundancy and speed drawbacks. Voxel-based algorithms convert point clouds into regular grids (voxels), significantly reducing computational load and leveraging 3D convolutional techniques for efficient processing. As a variety, PointPillars [19] enables interference at over 60 FPS with passable accuracy. Projection-based methods project point clouds onto 2D planes like CPM (camera plane maps)or BEV (bird's-eye view) images, enabling the use of 2D CNNs for detection [20, 21]. Jason et al. [22] proposed a method to project point clouds into several aggregated views, which helps restore features of pixels in CPM corrupted by LiDAR's lower resolution.

2.3 | Generative Adversarial Networks

GANs have emerged as a significant framework in machine learning, particularly in generative modelling. Introduced by Goodfellow et al. in 2014 [23], GANs consist of two key components: a generator network that creates synthetic data and a discriminator network that distinguishes real from synthetic data. This dual network structure, typically implemented with deep neural networks, allows GANs to capture complex patterns and produce realistic outputs. The architecture of GAN is illustrated in Figure 2. The generator learns to synthesise realistic images through adversarial training, while the discriminator evaluates their authenticity. This framework enables robust domain translation capabilities without requiring paired training data.

Original GANs suffered from issues such as a strong discriminator but weak generator, mainly caused by imperfect Jensen–Shannon (J-S) divergence and mode collapse [24]. Conditional GANs, such as pix2pix [25], enhanced image generation by incorporating additional information as guidance and replacing multi-linear perceptions with CNNs. CycleGAN [26] extended these capabilities by enabling domain translation without paired images and simultaneously training generators for two diverse domains. This research uses cycleGAN's characteristics to generate NIR images from the RGB domain. Considering the importance of the detailed quality of an image in downstream detection tasks, SSIM loss [27] is introduced as part of the loss function to simulate NIR camera outputs better.

2.4 | Multi-Sensor Fusion

Single-modal data perception has inherent drawbacks. For instance, achieving precise image processing outcomes at

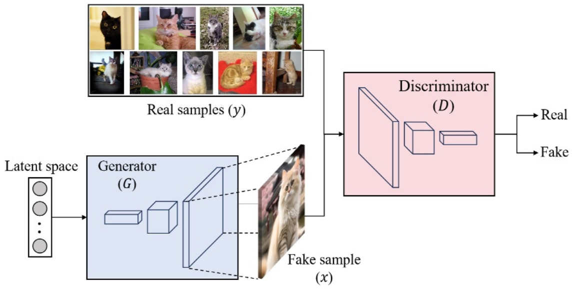

FIGURE 2 | Structure of GAN for generating images, where G and D stand for generator and discriminator, respectively.

nighttime is challenging. Also, LiDAR's mechanical structure results in varying resolutions at different distances and is susceptible to extreme weather conditions such as fog and heavy rain. Although LiDAR and camera data are defective in various aspects individually, their combination leverages the complementary strengths of each modality, leading to improved perception performance.

Multimodal fusion methods for perception tasks have advanced rapidly. Mainstream approaches typically employ RGB cameras and LiDAR as the primary fusion sensors, with IR cameras and millimetre wave radar used to address specific conditions like fog, heavy rain, and darkness [7].

Multimodal data fusion methods can be categorised into three paradigms: data-level fusion (early fusion), feature-level fusion (deep fusion), and object-level fusion (late fusion).

Data-level or early fusion combines raw sensor data from different modalities through spatial alignment. This method integrates semantic features from image data with various formats of LiDAR information, such as raw point clouds, voxelised tensors, and 2D mapping images, resulting in improved object detection performance.

Feature-level deep fusion combines cross-modal data in the feature space through concatenation or element-wise multiplication. VirConvNet [28], for example, fuses voxelised LiDAR data with pseudo points generated in images to design a new operator virtual sparse convolution (VirConv) and reaches high performances.

Object-level fusion combines the detection results from different modalities at the decision level. This approach simplifies implementation and allows for independent optimisation of each sensor's detection algorithm. CLOCs (camera-LiDAR object candidates) [29] and Fast-CLOCs [30] refine 3D region proposals from LiDAR by incorporating 2D proposals from camera data, using statistical features like confidence scores, distances, and IoU (intersection over union) to optimise the final proposals.

Integrating multiple sensors through data-level, feature-level, and object-level fusion methods offers significant potential to overcome the limitations of individual sensors in low-illumination target detection. Future research should enhance these fusion-based detection systems' real-time performance and accuracy, particularly in challenging environmental conditions, to improve their applicability in practical scenarios such as autonomous driving and surveillance.

3 | Research Methodology

This research aims to enhance target detection in low-illumination environments by integrating RGB images, NIR images, and point cloud data through a multi-sensor fusion framework.

The RMF-ED methodology comprises three primary components: NIR image generation, backbone architectures, and the multimodal late fusion system. Each component is designed to seamlessly integrate into the RMF-ED framework, ensuring optimised performance in low-light conditions.

An NIR-GAN model was trained to generate an NIR dataset for detection tasks directly transferred from existing RGB datasets in NIR image generation. In the image and point cloud data branches, YOLO and PointPillars were selected as backbones, and raw proposals were output on each branch. In a multisensory fusion system, a late fusion system including consistency check, fusion model, and post-process was proposed to obtain refined results with new locations and confidence scores from raw proposals.

3.1 | NIR Image Generation

Data-driven approaches thrive on the availability of large, high-quality training datasets. While object detection in RGB images benefits from extensive and well-curated datasets for both 2D and 3D tasks [31], the field of NIR image detection is hindered by a scarcity of labelled datasets. This limitation presents a significant challenge in developing robust detection systems that perform effectively in NIR domains. However, existing NIR image datasets predominantly focus on outdoor scenes, such as landscapes, fruits, and vegetation, as referenced in [32, 33]. These datasets lack annotations for pedestrian bounding boxes, which poses a significant challenge for developing robust pedestrian detection models in the NIR domain.

The primary objective of this section is to obtain sufficient labeled pedestrian data in NIR images. To achieve this, we employed domain translation techniques to convert RGB images into the NIR domain using cycleGAN, which has shown exceptional capabilities in style transfer and the generation of realistic synthetic images [26]. Leveraging these capabilities, we can effectively translate existing labeled RGB datasets into the NIR domain, facilitating downstream tasks such as detection and re-identification with enhanced accuracy and reliability.

We selected cycleGAN as the framework because it can operate without paired images while enhancing performance across various data domains (Figure 3). The architecture of cycleGAN comprises two generator networks ($G_{X\to Y}$ and $G_{Y\to X}$) and two

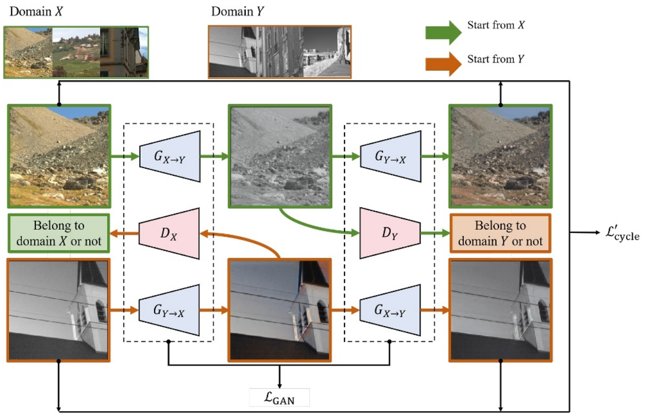
Domain X
Domain Y
Start from X
Start from Y
$G_{X\to Y}$
$G_{Y\to X}$
Belong to domain X or not
D_X
D_Y
Belong to domain Y or not
$\mathcal{L}'_{\text{cycle}}$
$G_{Y\to X}$
$G_{X\to Y}$
$\mathcal{L}_{\text{GAN}}$

FIGURE 3 | Structure of the cycleGAN utilised for NIR image generation.

discriminator networks (D_X and D_Y). Given two sets of images from RGB domain X and NIR domain Y, the generator $G_{X\rightarrow Y}$ takes an RGB image from domain X and produces a corresponding NIR image in domain Y, and vice versa for $G_{Y\rightarrow X}$. Subsequently, the discriminator D_X is trained to distinguish between real and fake RGB images in domain X, while D_Y performs the analogous task for NIR images in domain Y.

The cycle consistency loss in cycleGAN, denoted as L_{cycle}, was employed to ensure image consistency and preservation of important features after the $X \rightarrow Y \rightarrow X$ translation loop. Which can be mathematically expressed as follows:

$$\begin{aligned} L_{\text{cycle}} = \mathbb{E}_{\boldsymbol{x}\in X}\left[\|G_{Y\rightarrow X}(G_{X\rightarrow Y}(\boldsymbol{x})) - \boldsymbol{x}\|_1\right] \\ + \mathbb{E}_{\boldsymbol{y}\in Y}\left[\|G_{X\rightarrow Y}(G_{Y\rightarrow X}(\boldsymbol{y})) - \boldsymbol{y}\|_1\right]. \end{aligned} \tag{1}$$

where L_1 loss was used in the original cycleGAN to evaluate and minimise the differences between the original and generated images.

In this work, downstream object detection tasks exhibited a high sensitivity to image structures, particularly at the edges. Although L_1 loss ensured the overall quality, such as the brightness and colour of the generated images, it neglected the detailed edge information, leading to artefacts and blurring. To preserve edge details areas while maintaining brightness and colour, we employed a combination of SSIM loss L_{SSIM} and L_1 [34].

$$L'_{\text{cycle}} = \alpha L_{\text{SSIM}} + (1-\alpha)L_1, \tag{2}$$

where α is a hyperparameter to adjust the proportion of the two terms. To balance the loss values of the two terms, we conducted experiments on several images and set α to 0.8.

In general, the objective of this improved cycleGAN can be formulated as a min-max optimisation problem:

$$\arg \min_{G_{X\rightarrow Y}, G_{Y\rightarrow X}} \max_{D_X, D_Y} L_{\text{GAN}} + \lambda L'_{\text{cycle}} \tag{3}$$

where the weighting parameter λ controls the trade-off between adversarial training (L_{GAN}) and cycle-consistency regularisation (L'_{cycle}).

3.2 | Backbones

RMF-ED incorporates two distinct detection backbones optimised for multimodal data: YOLOv5 for the image data and PointPillars for LiDAR data. These backbones were chosen for their balance of accuracy and real-time processing capabilities.

1. Image Data Branch: YOLOv5 processes RGB and generates NIR images. An early fusion strategy combines RGB and NIR images into a single 4-channel input, maximising the utility of both modalities. This approach significantly improves the quality of raw detection proposals compared to RGB images alone.
2. LiDAR Data Branch: PointPillars was selected as the LiDAR backbone due to its efficient processing capabilities, achieving an inference time of just 16 ms per frame. PointPillars converts raw point clouds into 3D proposals, which are then projected onto the 2D camera plane for alignment with YOLOv5 outputs.

Combining these backbones ensures robust detection in both spatial and depth dimensions, laying a strong foundation for the subsequent fusion process.

3.2.1 | Image Data Branch

Object detection is fundamental in computer vision, enabling machines to identify and localise objects within images or video streams. After evaluating various prominent one-stage and two-stage detection frameworks, we excluded two-stage algorithms due to the need for instantaneous processing. Ultimately, we selected YOLOv5 as our image detection framework. YOLOv5 is noted for its outstanding accuracy and efficiency, leveraging hierarchical feature extraction techniques such as spatial pyramid pooling [35] to enhance multi-scale object detection. This makes it an optimal choice for real-time computer vision applications. Details on the architecture and implementation of YOLOv5 can be found in ref. [36].

We employed an early fusion strategy to integrate image information from RGB and NIR domains in our image detection branch. The early fusion strategy operates on original resources to maximise the utility of available information. First, 3-channel RGB test images $\in \mathbb{R}^{W\times H\times 3}$ were put into pretrained NIR-GAN to generate 1-channel NIR images $\in \mathbb{R}^{W\times H\times 1}$. Subsequently, they were concatenated to form 4-channel fused images $\in \mathbb{R}^{W\times H\times 4}$, which were then fed into the modified detector to obtain raw proposals. Predictions of ordinary 3-channel RGB images established the baseline.

It is important to note that the dataset used to train the GAN model predominantly consists of images captured under normal lighting conditions, including shadows under sunlight. Most low-illumination conditions, which typically occur at night or in dim indoor environments, were not represented in the nirscene1 dataset. This discrepancy could lead to a distribution variance between the training data and the on-scene task images. To address this issue, we analysed the average pixel illumination and variance metrics of the nirscene1 dataset. During object detection tasks, the illumination of RGB images was translated and aligned with nirscene1 by applying gamma correction. We believe this adjustment can mitigate the mismatch in the illumination domain and restore details in dark areas, thereby enhancing detection performance in low-light conditions.

3.2.2 | Point Cloud Data Branch

With the rise of autonomous driving and robotics, accurate and efficient object detection in 3D point clouds has become a critical research area. Due to the vast amount of data involved, 3D detection algorithms often lag in efficiency compared to image-based ones. The inference time of state-of-the-art LiDAR detection algorithms, based on point voxel and image data, shows a potential of around 30 ms. In contrast, PointPillars has achieved remarkable inference speed with only 16 ms latency, and subsequent research efforts have widely adopted its architecture. These make PointPillars the ideal detection backbone regarding the necessity of real-time requirements. The architecture and implementation details of PointPillars are thoroughly discussed in [31].

3.2.3 | 3D to 2D Projection

To enable seamless processing of both the 3D bounding box obtained from PointPillars and the 2D bounding box obtained from YOLOv5, we perform a projection transformation to map the 3D bounding box onto the camera coordinates. Subsequently, we utilise the minimum bounding rectangle to represent the projected 3D bounding box. This ensures spatial consistency and facilitates unified handling of the bounding boxes across different detection methods.

Let $\mathrm{L} \in \mathbb{R}^3$, $\mathrm{C} \in \mathbb{R}^3$, $\mathrm{W} \in \mathbb{R}^3$, and $\mathrm{P} \in \mathbb{R}^2$ be the coordinate systems of LiDAR, camera, world, and pixel, respectively. Thus, $\boldsymbol{B}_\mathrm{L}$, $\boldsymbol{B}_\mathrm{C}$, and $\boldsymbol{B}_\mathrm{P}$ respectively represent the bounding boxes under LiDAR, camera, and pixel coordinate systems. To transform a 3D bounding box $\boldsymbol{B}_\mathrm{L} = (x_\mathrm{L}, y_\mathrm{L}, z_\mathrm{L})$ from L to C, it is first transformed into the world coordinate system and then into the camera coordinate system:

$$\boldsymbol{B}_\mathrm{C} = \boldsymbol{T}_{\mathrm{W}\rightarrow\mathrm{C}} \boldsymbol{T}_{\mathrm{L}\rightarrow\mathrm{W}} \boldsymbol{B}_\mathrm{L} \tag{4}$$

where $\boldsymbol{T}_{\mathrm{W}\rightarrow\mathrm{C}}$ and $\boldsymbol{T}_{\mathrm{L}\rightarrow\mathrm{W}}$ are the transformation matrices in homogeneous coordinates. Objects represented in the camera coordinate system must be transformed to the imaging plane using geometric relations. Subsequently, the transformed objects are mapped to the pixel screen utilising the principles of pinhole imaging, assuming distortion effects are disregarded:

$$\begin{bmatrix} \boldsymbol{B}_\mathrm{P} & 1 \end{bmatrix}^\mathrm{T} = \begin{bmatrix} \frac{1}{\mathrm{d}_x} & 0 & u_0 \\ 0 & \frac{1}{\mathrm{d}_y} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{B}_\mathrm{C} \tag{5}$$

where f is the focal length of the camera (assumed to be equal in vertical and horizontal directions), $\frac{1}{\mathrm{d}_x}$ and $\frac{1}{\mathrm{d}_y}$ are the units of length represented by one pixel, and u_0 and v_0 are the horizontal and vertical pixel offsets between the image centre and the origin of the pixel coordinates.

4 | Multi-Sensory Fusion System

The image data branch and point cloud data branch independently conduct object detection tasks on image and point cloud data, respectively, with each branch generating a set of bounding boxes as the raw proposals. As shown in Figure 4, the framework incorporates a multimodal detection fusion system to enhance the robustness and accuracy of the prediction results. Initially, the system verifies multimodal data's semantic and geometric consistency and merges valid data using a fusion encoder. Rather than reusing the original bounding boxes of raw proposals, a fusion model is trained to predict new fusion bounding boxes' locations and confidence scores based on the learned joint distribution. To evaluate the accuracy of the fusion bounding boxes, we drew inspiration from YOLO and developed a novel regression loss function.

4.1 | Bounding Box Pair Matching

The detection results of each branch are characterised by three essential properties: the spatial location of bounding boxes, class probabilities, and confidence scores. However, a challenge arises as the detection results from different modalities do not align on a one-to-one basis. The Hungarian and Kuhn-Munkres algorithms are commonly employed to find solutions, albeit requiring substantial computational resources and posing difficulties in determining appropriate cost weights. To simplify the matching problem, we propose the utilisation of semantic consistency and geometric consistency as prepossessing techniques for the multimodal data.

Semantic consistency checks the class probabilities of detection proposals from camera and LiDAR (cls_C and cls_L), ensuring that results obtained from different modalities correspond to the same underlying objects. The classification of classes is determined based on the maximum score obtained from the softmax value of class probabilities. In this study, we precisely define the targets of semantic consistency as ‘person’ and ‘pedestrian’ in YOLO and PointPillars, respectively. Consequently, all detection results about other classes are filtered out.

Geometric consistency refers to the location correspondence between objects detected in different modalities. Utilising the

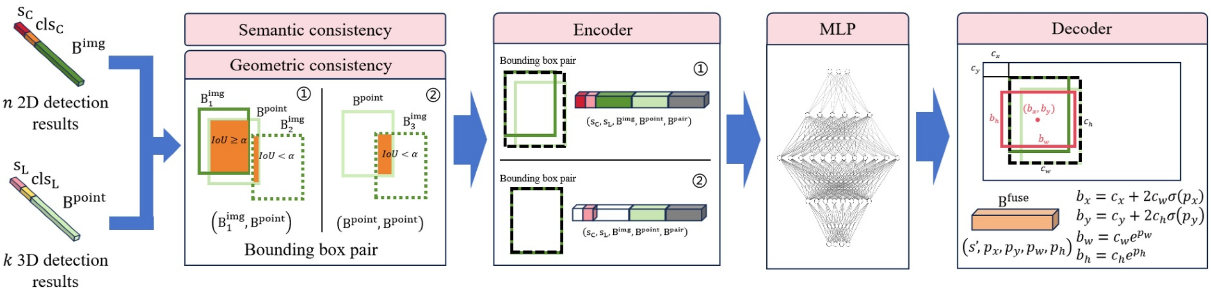

FIGURE 4 | Detailed structure of the fusion model for detection.

IoU metric, we assess the geometric consistency of the bounding boxes between the two modalities. Corresponding objects should exhibit matched bounding boxes that align closely regarding location. Due to its richer original data, point cloud data offers more comprehensive information compared to images. Let $B^{\text{point}} \in \mathbb{R}^4$ and $B^{\text{img}} \in \mathbb{R}^4$ represent the bounding boxes detected from the point cloud and image data, respectively. By aligning the k projected PointPillar proposals $\left\{B_1^{\text{point}}, B_2^{\text{point}}, \ldots, B_k^{\text{point}}\right\}$ with n YOLOv5 proposals $\left\{B_1^{\text{img}}, B_2^{\text{img}}, \ldots, B_n^{\text{img}}\right\}$, we determine the best-matched results between the two modalities.

For each B_j^{point}, we calculate its IoU with every B_i^{img}, $i = 1, 2, \ldots, k$. The B_i^{img} with the highest IoU beneath the threshold is recognised as the most corresponding one to match a bounding box pair P_j. If no B^{img} is above the threshold, the match would be B_j^{point} itself.

$$P_j = \begin{cases} \left(B_j^{\text{point}}, B_j^{\text{point}}\right), \forall B_i^{\text{img}}, \text{IoU}\left(B_i^{\text{img}}, B_j^{\text{point}}\right) < \alpha \\ \left(B_j^{\text{point}}, \arg\max_i \text{IoU}\left(B_j^{\text{point}}, B_i^{\text{img}}\right)\right), \text{otherwise} \end{cases} \tag{6}$$

We refer to the test criteria for the famous PASCAL VOC dataset, which tests the accuracy of target recognition with an IoU threshold of 0.5. Considering the low resolution of LiDAR and to handle occlusion situations, the IoU threshold α was set to 0.5. After the consistency check, k bounding boxes pairs across multimodal data are obtained.

4.2 | Fusion Model

The fusion model comprises three components: the fusion encoder, the decoder, and the neural network. The fusion encoder takes bounding box pairs as input and encodes them into a suitable format for the neural network. The neural network then processes the encoded data and produces an output. This output is subsequently passed through the decoder, which converts it into the predicted coordinates of the fusion bounding box. In this way, the fusion encoder, decoder, and neural network work collaboratively to handle the fusion of multimodal data, generate refined locations, and provide accurate estimates for the fusion bounding box coordinates.

After the one-to-one matching of bounding box pairs is achieved, a fusion encoder is employed to integrate the multimodal information. To enhance the model's generalisability, we refrain from utilising the spatial distribution of bounding boxes in each modal within a single image. Instead, we propose that the coordinates of the fused bounding box can be derived from the confidence scores of each modality and the coordinates of the individual bounding boxes. We also measure the uncertainty in position between modes by calculating the coordinates of the minimum enclosing rectangle $B^{\text{pair}} \in \mathbb{R}^4$ of bounding box pairs. This encoding approach enables the model to learn the joint probability distribution of multimodal bounding boxes and subsequently determine the coordinates of the fused bounding box with the highest probability.

Following the fusion encoder, the multimodal data is combined into 14-dimensional input vectors v_{in} denoted as follows:

$$v_{\text{in}} = \left(s_{\text{C}}, s_{\text{L}}, B^{\text{img}}, B^{\text{point}}, B^{\text{pair}}\right) \tag{7}$$

where s_{C} and s_{L} are the confidence scores of proposals from YOLOv5 and PointPillars, bounding boxes are represented by the top-left and bottom-right vertexes in coordinates $\left(x_1, y_1, x_2, y_2\right)$.

The input vector is then forwarded to an MLP (multi-layer perception) neural network consisting of three hidden layers, with 64, 256, and 256 neurons, respectively, assigned to each layer. This MLP aims to learn the non-linear mapping from the input fusion vector to the fused bounding box coordinates and confidence scores. The MLP generates output that is passed to the decoder module by leveraging this mapping. The output vectors v_{out} could be represented as follows:

$$v_{\text{out}} = \left(s', p_{\text{x}}, p_{\text{y}}, p_{\text{w}}, p_{\text{h}}\right) \tag{8}$$

where p_{x} and p_{y} are the location factor of the fused bounding box's centre point in the pixel coordinate system, p_{w} and p_{h} are the width and height factors, and s' is the confidence score of the fused bounding box.

The decoder module transforms the output generated by the MLP into the corresponding coordinates of the fused bounding boxes. Let the coordinates of the top-left vertex of B^{pair} be $\left(c_x, c_y\right)$, and c_w, c_h be its width and height in pixels. The location parameters of fused bounding boxes B^{fuse} could be calculated as follows:

$$\begin{cases} b_x = c_x + c_w \sigma\left(p_x\right) \\ b_y = c_y + c_h \sigma\left(p_y\right) \\ b_w = c_w e^{p_w} \\ b_h = c_h e^{p_h} \end{cases} \tag{9}$$

where b_x, b_y are the coordinates of the centre of B^{fuse} and b_w, b_h are its width and height, respectively. $\sigma()$ represents the Sigmoid function, which restricts the value range of p_x and p_y to lie between 0 and 1. By multiplying c_w and c_h, the centre of B^{fuse} is restricted within B^{pair}. The exponential items in b_w and b_h ensure the non-negativity of width and height.

The loss function of the fusion model comprises regression loss and confidence loss. B^{fuse} with an IoU greater than 0.5 aligned with the ground truth are considered positive. For each positive B^{fuse}, we compute the IoU loss between the box and its corresponding ground truth as regression loss. The confidence loss is the binary cross-entropy loss after logistic regression among all the positive and negative boxes.

5 | Experiments and Results

5.1 | Dataset

The dataset used to train the GAN was nirscene1 [32], which contains 477 aligned pairs of RGB and NIR images captured using commercial high-end cameras equipped with 750 nm band-cutoff filters. Image registration was achieved through feature matching between RGB and NIR domains. The dataset contains nine semantic categories: country, field, forest, mountain, old building, street, urban, water, and indoor scenes. Training GANs on this limited dataset presents significant challenges due to insufficient data for learning cross-domain mapping functions. To mitigate this, we employed multiple data augmentation strategies during training including local/global flipping, rotation, stitching, and cropping, thereby enhancing the diversity and robustness of the training set.

The object detection experiments utilised the KITTI dataset [37], collected using a sensor-equipped station waggon containing: two colour and two grayscale PointGrey Flea2 video cameras (10 Hz, resolution: 1392 × 512 pixels, opening: 90° ×35°), a Velodyne HDL-64E 3D laser scanner (10 Hz, 64 laser beams, range: 100 m), a GPS/IMU localisation unit with RTK correction signals (open sky localisation errors < 5 cm) and a powerful computer running a real-time database.

The four cameras were intrinsically and extrinsically calibrated with subsequent image rectification. We established 3D rigid motion parameters to align the coordinate systems of the LiDAR, GPS/IMU, and reference camera.

The KITTI object detection benchmark contains 7481 training frames with 51,867 annotated objects across nine categories. For our evaluation, we specifically used the 'Pedestrian' class labels.

5.2 | Low-Illumination Environment Simulation

To simulate low-light environments for evaluation, we modified the KITTI dataset, originally captured under normal lighting, using a systematic image degradation process. Following the method of Cui et al. [38], the simulation involved three steps: (1) transforming RGB images into RAW format to preserve original sensor data, (2) applying controlled low-light degradation techniques such as brightness reduction and noise addition, and (3) converting the degraded RAW images back to RGB format through forward image signal processing (ISP). This approach accurately replicates real-world low-light scenarios, ensuring consistent evaluation conditions for RMF-ED. The resultant dataset includes variations in brightness, noise levels, and edge clarity, reflecting practical challenges in low-illumination detection.

5.3 | Experimental Setup

Experiments were conducted on an Ubuntu system with an NVIDIA 3090 graphics card. The GAN architecture employed the Unet256 network as the generator and the 70 × 70 PatchGAN network as the discriminator. The GAN was trained for 200 epochs, and the final model was obtained from the last epoch. The MLP network in the post-fusion model was trained for 100 epochs using the Adam optimiser, with an initial learning rate of 1×10^{-4}.

The nirscene dataset was partitioned into three subsets: training, validation, and test sets, following a split ratio of 6:2:2. Following the guidelines in the official KITTI documentation, the KITTI training set was further divided into training and validation sets, with an approximate 1:1 ratio. The final test results must be uploaded to the KITTI official website for online evaluation.

Various loss functions were employed to train our GAN model on the nirscene dataset. The generated NIR images were evaluated using the Fréchet Inception Distance (FID), which measures the dissimilarity between the generated and authentic NIR images. FID is a metric that combines feature representations from the Inception model and computes the Fréchet distance between their multivariate Gaussian distributions. It is commonly used to evaluate the quality of generative models.

5.4 | Results of NIR-GAN

Table 1 presents the FID scores obtained using different loss functions. For models trained with L1 loss alone, SSIM loss alone, and the combined L1 and SSIM loss, the corresponding FID scores are 61.00, 49.72, and 46.95, respectively. Incorporating SSIM in addition to L1 loss results in an approximately 18.5% improvement in the GAN's performance. Furthermore, employing both L1 and SSIM concurrently further enhances the realism of the generated NIR images.

To investigate the influence of different SSIM ratios on model performance further, we trained the model with various proportions of SSIM, ranging from 0 to 1 in steps of 0.1. Subsequently, we calculated the FID scores for these models. Figure 5 reveals that as the SSIM ratio increases, the FID score generally exhibits a decreasing trend, indicating an improvement in the model's performance. The FID score reaches its minimum at ratios around 0.6–0.7, suggesting that the model achieves optimal performance at these ratios.

Figure 6 compares generated NIR images with the original image under three loss function configurations. Models trained solely with the L1 loss exhibited mosaic-like artefacts in the red and grassy regions. Conversely, using SSIM loss alone resulted in pronounced blurring, particularly affecting local structural details. The combined use of L1 and SSIM loss produced the most accurate results, effectively restoring intricate local information about the building while maintaining overall image

TABLE 1 | FID score of different loss functions chosen for training NIR-GAN.

SSIM	L1	Propotion of SSIM	FID score
✓		1.0	49.72
	✓	0.0	61.00
✓	✓	0.7	46.95

quality. This demonstrates that the combination of SSIM and L1 maintains the colour tone and overall consistency of NIR images and accurately restores dense local structures.

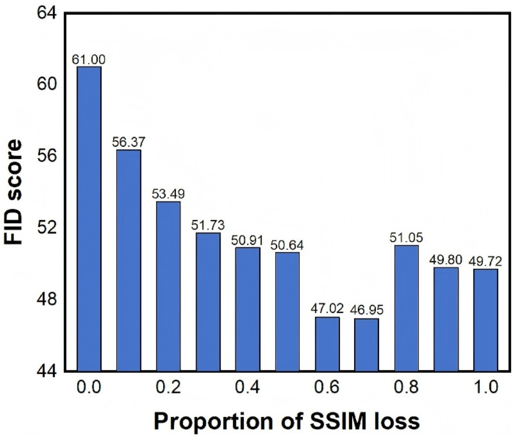

FIGURE 5 | Model performance at different SSIM ratios.

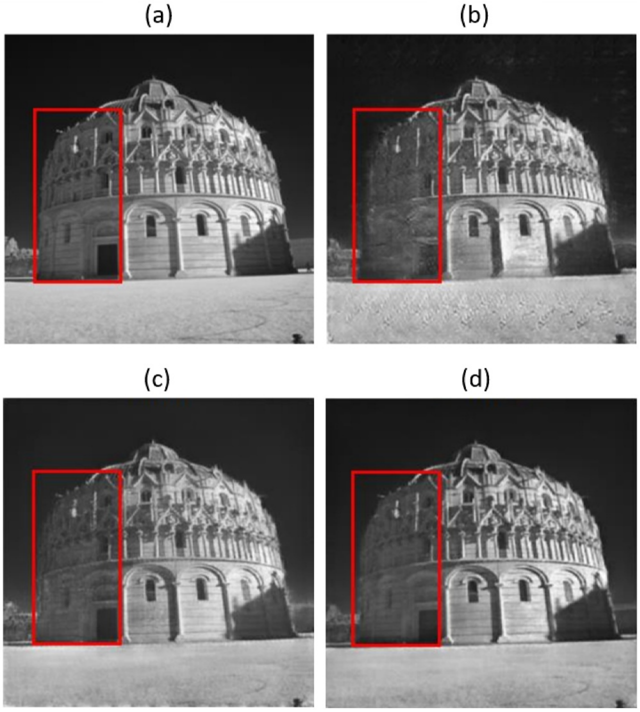

FIGURE 6 | Results comparison with different loss functions. (a) Origin, (b) L1 loss, (c) SSIM loss, (d) L1 + SSIM loss.

5.5 | Results of Fusion System

Our model was submitted for evaluation on the KITTI online testing platform. During the official tests, prediction boxes with IOU values greater than 0.5 relative to the ground truth labels are considered positive results. The test cases are categorised into three levels: easy, moderate, and difficult, based on factors such as bounding box height, occlusion level, and truncation.

Table 2 presents a comparative analysis, highlighting the superior performance of RMF-ED over other notable algorithms in moderate detection scenarios. PointPillars, used as modules in our work, solely utilise LiDAR data, while Faraway-Frustum, PointPainting and Fast-CLOCs incorporate LiDAR and image data. Our method significantly improves on all three difficulty levels compared to algorithms using only LiDAR data. Compared to fusion algorithms leveraging LiDAR and image data, although our approach shows performance gaps in simple and difficult scenarios, we achieve better results in moderate scenarios, especially when compared with PointPainting and Faraway-Frustum which are designed for 3D detection results. In particular, KITTI officially employs the performance score in the moderate scenario as the primary ranking criterion for evaluating algorithms.

Regarding computation time, our post-fusion system demonstrates efficient performance due to the utilisation of the PointPillars and YOLO algorithms as the underlying frameworks. The system's computation time is approximately 20 ms with more than 30 FPS, enabling real-time predictions. Specifically, the PointPillars, YOLO, and post-fusion algorithms require 14, 5, and 2 ms for prediction. Compared to other fusion algorithms, our approach achieves comparable performance while maintaining the lowest computation time. Moreover, the model's computation time can be further reduced by implementing parallel computing techniques. This aspect is highly advantageous for deploying algorithms in various industrial applications.

Figure 7 presents qualitative results comparing the proposed method with backbone models. The first row depicts the original RGB images from the KITTI dataset, while the second row shows the corresponding NIR images generated from them. Due to image size limitations, the original images were halved and reassembled, resulting in slight brightness differences between the left and right halves. Red and blue bounding boxes represent detection results from YOLOv5 and PointPillars, with confidence scores exceeding 0.25 and 0.40, respectively. Green

TABLE 2 | Comparison of our methods with other algorithms on KITTI validation dataset.

	Source				2D AP (%)		
Dimension	**Camera**	**LiDAR**	**Methods**	**Inference time (s)**	**Easy**	**Moderate**	**Hard**
3D	✓	✓	PointPainting [10]	0.400	61.86	**53.76**	50.61
3D	—	✓	PointPillars [19]	0.016	65.29	**55.10**	52.39
3D	✓	✓	Faraway-Frustum [39]	0.100	67.88	**57.35**	54.42
2D	✓	✓	Fast-CLOCs [40]	0.100	76.22	**62.57**	60.13
2D	✓	✓	**RMF-ED (ours)**	0.020	67.62	**63.24**	58.60

Note: Bold entries indicate our proposed method and the moderate AP scores, which are the standard evaluation metrics on the KITTI dataset.

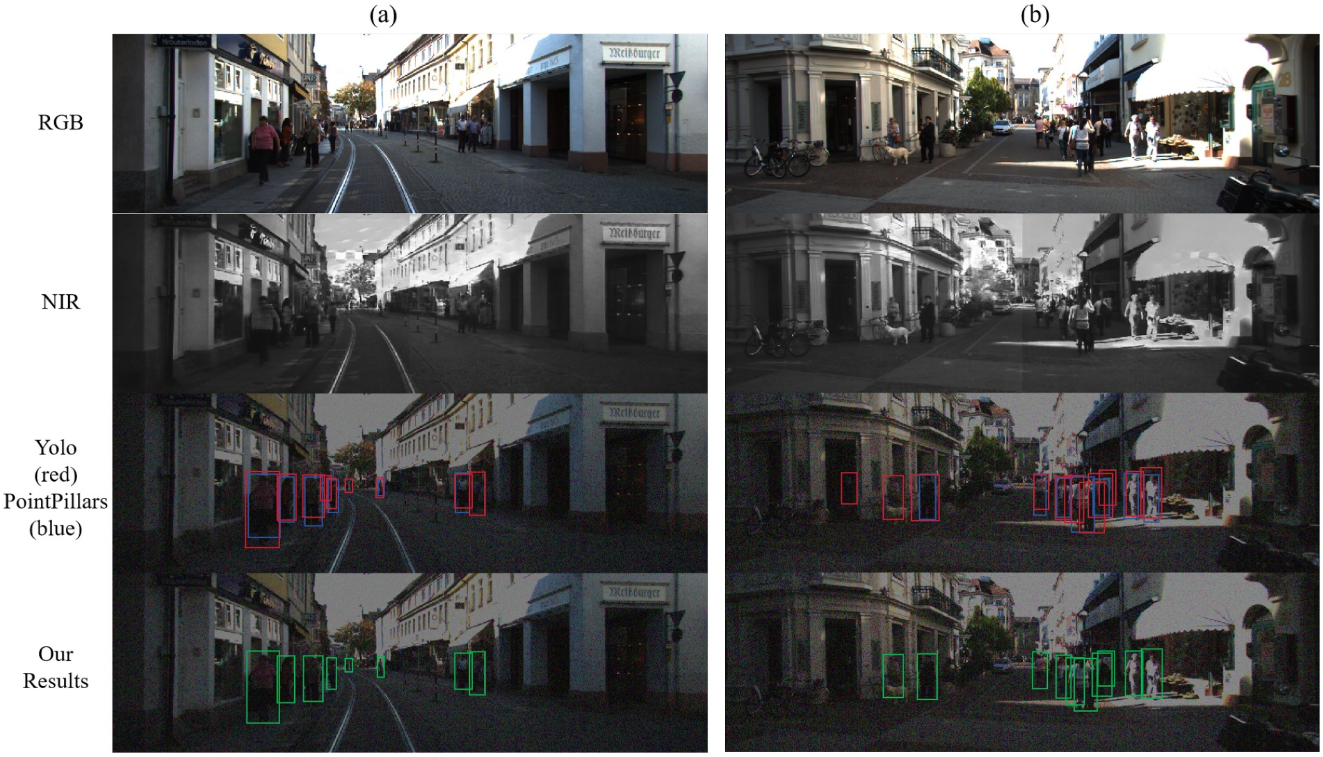

FIGURE 7 | Qualitative comparison of RMF-ED with YOLOv5 and PointPillars on KITTI dataset. (a) The proposed fusion approach corrected missed detections from individual branches, (b) a false positive was effectively filtered. Top: original RGB images; second: generated NIR images; third and fourth: low-illumination simulations. Red, blue, and green bounding boxes denote YOLOv5, PointPillars, and RMF-ED detections, respectively.

bounding boxes illustrate detection results of the proposed method, with confidence scores above 0.4. In Figure 7a, the proposed fusion approach corrected missed detections from individual branches, while in Figure 7b, a false positive was effectively filtered. Additionally, the bounding boxes were refined for improved accuracy.

The results of the ablation experiment shown in Table 3 further investigate the contribution of each data source to our algorithm, including low-illumination RGB images generated by ISP, NIR images generated by NIR-GAN and LiDAR points. For clarity, we denote the experimental setups as follows: (a) RGB, (b) NIR, (c) LiDAR, (d) RGB + NIR, (e) RGB + LiDAR, (f) NIR + LiDAR, and (g) the proposed RMF-ED method integrating RGB, NIR, and LiDAR data.

In summary, LiDAR points are crucial for the algorithm. LiDAR data dramatically enhances performance across single-source experiments (a, b, and c) and dual-source combinations (d, e, f). RGB images show the lowest performance mainly due to weak contrast and excessive noise in low-light conditions. Experiments (a) and (b) reveal that including NIR images improves visual detection performance. Still, results from (e) and (f) indicate that the cooperation of LiDAR data substantially reduces the performance gap between low-illumination RGB and generated NIR images, underscoring the decisive role of LiDAR data.

Finally, experiment (g), which employs the RMF-ED method combining all three modalities (RGB, NIR, and LiDAR), achieves the best performance across all difficulty levels. This highlights the effectiveness of our proposed fusion strategy in leveraging complementary strengths from multiple data sources, especially in challenging low-illumination scenarios.

TABLE 3 | Ablation study of input modalities (Cases: a = RGB, b = NIR, c = LiDAR, d = RGB + NIR, e = RGB + LiDAR, f = NIR + LiDAR, g = RGB + NIR + LiDAR).

	Data sources			Results (%)		
Case	**RGB**	**NIR**	**LiDAR**	**Easy**	**Moderate**	**Hard**
a	✓	—	—	28.52	**24.13**	16.08
b	—	✓	—	48.92	**45.23**	35.87
c	—	—	✓	65.29	**55.10**	52.39
d	✓	✓	—	49.77	**45.46**	35.16
e	✓	—	✓	65.48	**55.37**	52.43
f	—	✓	✓	67.46	**62.93**	58.49
g	✓	✓	✓	67.62	**63.24**	58.60

Note: Bold values in the Moderate column denote the primary evaluation metric according to the KITTI benchmark standard.

6 | Conclusions

This study introduces RMF-ED, a late fusion framework integrating LiDAR and multi-camera data to improve pedestrian detection in low-light environments. By addressing the challenges of nighttime driving and drone-based surveillance, RMF-ED contributes to the evolving field of multi-sensor fusion, demonstrating its potential for robust target detection under adverse lighting conditions.

The RMF-ED framework comprises two key branches: the LiDAR branch, which utilises the PointPillars algorithm for point cloud processing, and the image branch, which processes RGB and NIR data through the YOLOv5 network. To address the limitations posed by the scarcity of annotated NIR datasets, we developed an NIR-GAN model based on cycleGAN, augmented with SSIM and L1 loss functions. This approach facilitates the generation of high-quality NIR images from RGB data, thereby expanding the dataset available for training and improving detection capabilities.

The late fusion model integrates the outputs from both branches using an MLP, producing refined bounding boxes and confidence scores. By leveraging the complementary strengths of LiDAR and camera data, RMF-ED demonstrates promising improvements in detection accuracy and robustness under low-light conditions.

Experimental evaluations on the KITTI dataset suggest that the optimised NIR-GAN model generates NIR images that closely approximate real-world equivalents, outperforming the original cycleGAN model in this context. Additionally, the post-fusion algorithm achieves real-time performance with a processing time of only 21 ms, which is suitable for deployment in autonomous systems. While comparisons with other state-of-the-art algorithms reveal some trade-offs, RMF-ED shows competitive or superior performance in moderate and challenging detection scenarios, indicating its potential applicability in real-world settings.

In conclusion, RMF-ED represents a step forward in integrating GAN-generated NIR images with late fusion strategies to address the limitations of single-modality detection systems. The results suggest that multimodal fusion techniques, as demonstrated in this study, hold promise for improving detection reliability in low-light conditions. Future work could extend our framework to multi-object tracking scenarios by integrating appearance features and high-performance detection modules [41]. We hope this work will provide valuable insights for further developments in intelligent sensing and measurement technologies, particularly in challenging environments.

Due to time constraints and limited experimental resources, this study has not yet incorporated comprehensive experimental validation. However, we have explicitly prioritised this verification as critical future work to rigorously evaluate our method's effectiveness. A multi-sensor UAV system integrating RGB/NIR cameras and LiDAR has been deployed for ongoing data collection focused on pedestrian detection under low-illumination scenarios, supported by a semi-automatic annotation pipeline with manual verification. Upon completion of dataset collection, we plan to conduct real-environment evaluations and incorporate the results into a camera-ready version for formal publication.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The codes of RMF-ED and NIR-GAN are publicly available at: https://github.com/DrErwin/RMF-ED.

References

1. S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection With Region Proposal Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, no. 6 (June 2017): 1137–1149, https://doi.org/10.1109/TPAMI.2016.2577031.

2. R. Varghese and M. Sambath, “YOLOv8: A Novel Object Detection Algorithm With Enhanced Performance and Robustness,” in *Proceedings of the 2024 International Conference Advances in Data Engineering and Intelligent Computing Systems (ADICS)* (April 2024), 1–6, https://doi.org/10.1109/adics58448.2024.10533619.

3. W. Liu, D. Anguelov, D. Erhan, et al., “SSD: Single Shot Multibox Detector,” in *Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands* (October 2016), 21–37, https://doi.org/10.1007/978-3-319-46448-0_2.

4. S. Liu and D. Huang, “Receptive Field Block Net for Accurate and Fast Object Detection,” in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), 385–400, https://doi.org/10.1007/978-3-030-01252-6_24.

5. Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, “Object Detection in 20 Years: A Survey,” *Proceedings of the IEEE* 111, no. 3 (2023): 257–276, https://doi.org/10.1109/jproc.2023.3238524.

6. R. Zhang, J. Zhang, and H. Yu, “Review of Modeling and Control in UAV Autonomous Maneuvering Flight,” in *Proceedings of the 2018 IEEE International Conference Mechatronics and Automation (ICMA)* (August 2018), 1920–1925, https://doi.org/10.1109/icma.2018.8484542.

7. M. L. Fung, M. Z. Chen, and Y. H. Chen, “Sensor Fusion: A Review of Methods and Applications,” in *Proceedings of the 29th Chinese Control and Decision Conf. (CCDC)* (May 2017), 3853–3860, https://doi.org/10.1109/ccdc.2017.7979175.

8. S. M. Pizer, E. P. Amburn, J. D. Austin, et al., “Adaptive Histogram Equalization and Its Variations,” *Computer Vision, Graphics, and Image Processing* 39, no. 3 (September 1987): 355–368, https://doi.org/10.1016/S0734-189X(87)80186-X.

9. L. Tao, C. Zhu, J. Song, T. Lu, H. Jia, and X. Xie, “Low-Light Image Enhancement Using CNN and Bright Channel Prior,” in *Proceedings of the 2017 IEEE International Conference Image Processing (ICIP), Beijing, China* (2017), 3215–3219, https://doi.org/10.1109/ICIP.2017.8296876.

10. Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, “Retinexformer: One-Stage Retinex-Based Transformer for Low-Light Image Enhancement,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), 12504–12513, https://doi.org/10.1109/iccv51070.2023.01149.

11. E. Land and J. J. McCann, “Lightness and Retinex Theory,” *Journal of the Optical Society of America* 58, no. 1 (1967): 1–11, https://doi.org/10.1364/josa.61.000001.

12. Y. Jiang, X. Gong, D. Liu, et al., “Enlightengan: Deep Light Enhancement Without Paired Supervision,” *IEEE Transactions on Image Processing* 30 (2021): 2340–2349, https://doi.org/10.1109/tip.2021.3051462.

13. L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, and J. Ma, “MSR-Net: Low-Light Image Enhancement Using Deep Convolutional Network,” *arXiv preprint* (2017), arXiv:1711.02488, http://arxiv.org/abs/1711.02488.

14. C. H. Hu, J. Yu, F. Wu, et al., “Face Illumination Recovery for the Deep Learning Feature Under Severe Illumination Variations,” *Pattern Recognition* 111 (2021): 107724, https://doi.org/10.1016/j.patcog.2020.107724.

15. L. Zhou, G. Wu, Y. Zuo, X. Chen, and H. Hu, “A Comprehensive Review of Vision-Based 3D Reconstruction Methods,” *Sensors* 24, no. 7 (2024): 2314, https://doi.org/10.3390/s24072314.

16. G. Choe, S. H. Kim, S. Im, J. Y. Lee, S. G. Narasimhan, and I. S. Kweon, "RANUS: RGB and NIR Urban Scene Dataset for Deep Scene Parsing," *IEEE Robotics and Automation Letters* 3, no. 3 (2018): 1808–1815, https://doi.org/10.1109/lra.2018.2801390.

17. G. Zamanakos, L. Tsochatzidis, A. Amanatiadis, and I. Pratikakis, "A Comprehensive Survey of LIDAR-Based 3D Object Detection Methods With Deep Learning for Autonomous Driving," *Computers & Graphics* 99 (2021): 153–181, https://doi.org/10.1016/j.cag.2021.07.007.

18. C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," in *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)* (2017), 5099–5108.

19. S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential Fusion for 3D Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 4604–4612, https://doi.org/10.1109/cvpr42600.2020.00466.

20. F. Bartoccioni, É. Zablocki, A. Bursuc, P. Pérez, M. Cord, and K. Alahari, "Lara: Latents and Rays for Multi-Camera Bird's-Eye-View Semantic Segmentation," in *Proceedings of the Conference on Robot Learning* (March 2023), 1663–1672, https://doi.org/10.48550/arXiv.2206.13294.

21. C. Chang, J. Zhang, K. Zhang, et al., "BEV-V2X: Cooperative Birds-Eye-View Fusion and Grid Occupancy Prediction via V2X-Based Data Sharing," *IEEE Trans. Intell. Veh.* 8, no. 11 (2023): 4498–4514, https://doi.org/10.1109/tiv.2023.3293954.

22. J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3D Proposal Generation and Object Detection From View Aggregation," in *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2018), 1–8, https://doi.org/10.1109/iros.2018.8594049.

23. I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative Adversarial Networks," *Communications of the ACM* 63, no. 11 (2020): 139–144, https://doi.org/10.1145/3422622.

24. A. Gunasekaran, "Generative Adversarial Networks: A Brief History and Overview," preprint, (December 2022), https://doi.org/10.20944/preprints202212.0191.v1.

25. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation With Conditional Adversarial Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 1125–1134, https://doi.org/10.1109/cvpr.2017.632.

26. J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017), 2223–2232, https://doi.org/10.1109/iccv.2017.244.

27. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing* 13, no. 4 (2004): 600–612, https://doi.org/10.1109/tip.2003.819861.

28. H. Wu, C. Wen, S. Shi, X. Li, and C. Wang, "Virtual Sparse Convolution for Multimodal 3D Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 21653–21662, https://doi.org/10.1109/cvpr52729.2023.02074.

29. S. Pang, D. Morris, and H. Radha, "CLOCs: Camera-LiDAR Object Candidates Fusion for 3D Object Detection," in *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2020), 10386–10393, https://doi.org/10.1109/IROS45743.2020.9341791.

30. S. Pang, D. Morris, and H. Radha, "Fast-CLOCs: Fast Camera-LiDAR Object Candidates Fusion for 3D Object Detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2022), 3736–3745, https://doi.org/10.1109/wacv51458.2022.00380.

31. P. Huang, M. Cheng, Y. Chen, H. Zuniga-Noël, K. A. Hossain, and C. Zhou, "Traffic Sign Occlusion Detection Using Mobile Laser Scanning Point Clouds," *IEEE Transactions on Intelligent Transportation Systems* 18, no. 9 (2017): 2364–2376, https://doi.org/10.1109/tits.2016.2639582.

32. M. Brown and S. Süsstrunk, "Multi-Spectral SIFT for Scene Category Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2011), 177–184, https://doi.org/10.1109/cvpr.2011.5995637.

33. I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, and C. McCool, "deepNIR: Datasets for Generating Synthetic NIR Images and Improved Fruit Detection System Using Deep Learning Techniques," *Sensors* 22, no. 13 (2022): 4721, https://doi.org/10.3390/s22134721.

34. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing* 13, no. 4 (2004): 600–612, https://doi.org/10.1109/tip.2003.819861.

35. K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, no. 9 (2015): 1904–1916, https://doi.org/10.1109/tpami.2015.2389824.

36. Ultralytics, "GitHub - Ultralytics/yolov5," *GitHub* (n.d.), https://github.com/ultralytics/yolov5.

37. A. Geiger, P. Lenz, and R. Urtasun, "Are We Ready for Autonomous Driving? The Kitti Vision Benchmark Suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2012), 3354–3361, https://doi.org/10.1109/cvpr.2012.6248074.

38. Z. Cui, Y. Qi, L. Gu, D. You, X. Zhang, and J. Zhang, "Multitask AET With Orthogonal Tangent Regularity for Dark Object Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 2553–2562, https://doi.org/10.1109/iccv48922.2021.00255.

39. A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast Encoders for Object Detection From Point Clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 12697–12705, https://doi.org/10.1109/cvpr.2019.01298.

40. H. Zhang, M. Wang, Y. Liu, and Y. Wang, "Faraway-Frustum: Dealing With Lidar Sparsity for 3D Object Detection Using Fusion," in *Proceedings of the 2021 International Intelligent Transportation Systems Conference (ITSC)* (2021), 1124–1131, https://doi.org/10.1109/itsc48978.2021.9564990.

41. F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and J. Yan, "Poi: Multiple Object Tracking With High-Performance Detection and Appearance Feature," in *Proceedings of the European Conference on Computer Vision Workshops (ECCV Workshops), Amsterdam, The Netherlands* (October 2016), 36–42, https://doi.org/10.1007/978-3-319-48881-3_3.