frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 22 May 2025
DOI 10.3389/fphy.2025.1599937



OPEN ACCESS

EDITED BY
Yu Liu,
Hefei University of Technology, China

REVIEWED BY
Guohua Lv,
Qilu University of Technology, China
Min Li,
Xinjiang University, China

*CORRESPONDENCE
Yukui Che,
454983185@qq.com

RECEIVED 25 March 2025
ACCEPTED 06 May 2025
PUBLISHED 22 May 2025

CITATION
Wang K, Hu D, Cheng Y, Che Y, Li Y, Jiang Z, Chen F and Li W (2025) Infrared and visible image fusion driven by multimodal large language models.
*Front. Phys.* 13:1599937.
doi: 10.3389/fphy.2025.1599937



# Infrared and visible image fusion driven by multimodal large language models

Ke Wang, Dengshu Hu, Yuan Cheng, Yukui Che*, Yuelin Li, Zhiwei Jiang, Fengxian Chen and Wenjuan Li

Qujing Power Supply Bureau, Yunnan Power Grid Co., Ltd., Kunming, China

**Introduction:** Existing image fusion methods primarily focus on obtaining high-quality features from source images to enhance the quality of the fused image, often overlooking the impact of improved image quality on downstream task performance.

**Methods:** To address this issue, this paper proposes a novel infrared and visible image fusion approach driven by multimodal large language models, aiming to improve the performance of pedestrian detection tasks. The proposed method fully considers how enhancing image quality can benefit pedestrian detection. By leveraging a multimodal large language model, we analyze the fused images based on user-provided questions related to improving pedestrian detection performance and generate suggestions for enhancing image quality. To better incorporate these suggestions, we design a Text-Driven Feature Harmonization (Text-DFH) module. Text-DFH refines the features produced by the fusion network according to the recommendations from the multimodal large language model, enabling the fused image to better meet the needs of pedestrian detection tasks.

**Results:** Compared with existing methods, the key advantage of our approach lies in utilizing the strong semantic understanding and scene analysis capabilities of multimodal large language models to provide precise guidance for improving fused image quality. As a result, our method enhances image quality while maintaining strong performance in pedestrian detection. Extensive qualitative and quantitative experiments on multiple public datasets validate the effectiveness and superiority of the proposed method.

**Discussion:** In addition to its effectiveness in infrared and visible image fusion, the method also demonstrates promising application potential in the field of nuclear medical imaging.

KEYWORDS

**infrared and visible image fusion, pedestrian detection, multimodal large language models, text-guided, model fine-tuning**

## 1 Introduction

Multimodal sensor technology has facilitated the application of multimodal images across various fields. Among them, infrared and visible images have been widely used in diverse tasks due to the complementary nature of the information they contain. Specifically, infrared images provide thermal radiation information of objects and are

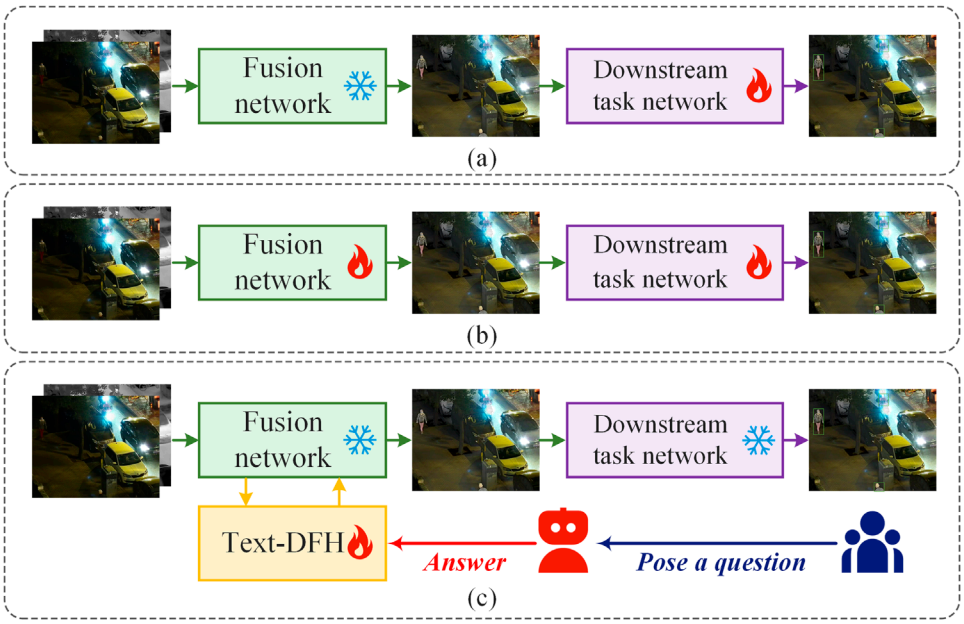


FIGURE 1
Comparison of different joint training strategies for image fusion and downstream tasks.

not affected by lighting conditions, but they lack detailed textures. In contrast, visible images capture rich texture details of the scene but are highly sensitive to lighting variations. Therefore, numerous methods [1–7] have focused on fusing infrared and visible images, aiming to integrate the complementary information from both modalities into a single, more informative fused image. This facilitates better decision-making and judgment in downstream tasks such as object detection [8–10] and semantic segmentation [11–14].

Current approaches that jointly train infrared-visible image fusion with downstream tasks can be broadly categorized into two types: independent optimization and joint optimization. Independent optimization methods first train a fusion network for infrared and visible images and then use the resulting fused images to train a downstream task network, as shown in Figure 1a. Consequently, most independent optimization methods focus on improving fusion quality, for example, by designing new network architectures [15–19] or introducing specific constraints [20–23]. However, such approaches neglect the potential guidance from downstream tasks and fail to establish a deep connection between fusion and task performance, often leading to suboptimal results. Simply chaining the fusion and downstream networks makes it difficult for the fused image to specifically cater to the downstream task's requirements. On the other hand, joint optimization methods use the downstream task network as a constraint to train the image fusion network, thereby forcing it to produce fused images that meet task-specific needs [24–28], as illustrated in Figure 1b. Nevertheless, the effectiveness of directly using high-level vision task supervision to guide fusion remains limited.

Recently, Multimodal Large Language Models (MLLMs) have gained popularity due to their strong capability in modeling data across different modalities, such as images and text. For instance, Text-IF [29] and TeRF [30] leverage large models to encode user instructions and guide various types of fusion tasks. However, these methods do not consider the possibility of using large language models to feed back the specific needs of high-level vision tasks to the image fusion process, which could further improve the quality of fused images.

To address this challenge, we propose a novel infrared and visible image fusion method driven by a Multimodal Large Language Model, aiming to simultaneously enhance fusion quality and pedestrian detection accuracy, as shown in Figure 1c. By leveraging the deep semantic understanding and scene analysis capabilities of MLLMs, we provide precise guidance for improving fused image quality while ensuring better pedestrian detection performance. Specifically, our method analyzes the fused images based on user-provided questions related to pedestrian detection, then generates optimization suggestions using feedback from the language model. To fully utilize these suggestions, we design a Text-Driven Feature Harmonization (Text-DFH) module, which refines the fusion network's output features under the guidance of the MLLM, allowing the fused images to better meet the demands of pedestrian detection.

In summary, the main contributions of this paper are as follows:

(1) We are the first to leverage Multimodal Large Language Models to provide feedback on the quality of fused images based on the specific requirements of downstream tasks, thus further improving infrared and visible image fusion.
(2) We propose an effective Text-Driven Feature Harmonization (Text-DFH) module that enables text-based guidance to assist in enhancing image quality.
(3) Our proposed method achieves excellent performance in infrared and visible image fusion, nuclear medical imaging, and pedestrian detection across multiple datasets.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of related work on multimodal large language models, infrared and visible image fusion, and pedestrian detection. Section 3 presents our proposed method in detail. Section 4 discusses the experimental results and analysis. Section 5 concludes the paper.

# 2 Related work

In this section, we first briefly introduce multimodal large language models, and then review existing infrared and visible image fusion methods.

## 2.1 Multimodal large language models

With the advent of the multimodal data fusion era, the capability of unimodal systems is no longer sufficient to handle complex real-world tasks. As a result, multimodal large language models (MLLMs) have been proposed to integrate information from multiple data sources, enabling more comprehensive and accurate representations. These models have demonstrated significant practical value across various domains, including natural language processing, vision tasks, and audio tasks. In the visual domain, MLLMs enhance the performance of tasks such as image classification, object detection, and image captioning by combining textual descriptions with visual instructions. For example, GPT-4V [31] and Gemini [32] integrate image content with natural language descriptions to produce more vivid and accurate annotations. NExT-GPT [33] and Sora [34] are at the forefront of multimodal video generation, producing rich and realistic content by learning from multimodal data. Moreover, VideoChat [35] and Video-LLaVA [36] demonstrate excellent capabilities in analyzing and understanding video content in intelligent video understanding scenarios.

In the field of image fusion, Text-IF [29] and MGFusion [37] uses CLIP [38] to encode user requirement texts, guiding the model to fuse images. TeRF [30] utilizes LLaMA [39] to encode user instruction texts and generate prompts for guiding image fusion across different tasks. Although these methods employ MLLMs to tackle some challenges in image fusion, they do not consider the specific requirements of high-level downstream visual tasks for image fusion quality, which limits the application of infrared and visible image fusion in such tasks.

## 2.2 Infrared and visible image fusion

Conventional infrared and visible image fusion methods mainly focus on designing sophisticated feature extraction networks and fusion strategies to ensure the quality of the fused results. From the perspective of network design, these methods can be broadly categorized into CNN-based methods, CNN-Transformer hybrid methods, and GAN-based methods. CNN-based methods [40–45] typically apply convolution, activation, and pooling operations to extract features from the input images, then fuse and reconstruct the final result using the extracted features. However, since CNNs can only perceive local features within a limited receptive field, they struggle to capture long-range contextual information, limiting their representational capacity. In contrast, Transformers [46] are better at modeling long-range dependencies and are more suited for capturing global features in images. ViT [47] was the first to introduce Transformer architectures into computer vision, achieving promising results. Subsequently, to combine the respective strengths of CNNs and Transformers, hybrid methods have gained increasing attention in the image fusion domain. For instance, CGTF [48], SwinFusion [16], YDTR [17], and DATFuse [49] insert Transformer layers after CNN layers to jointly leverage local and global feature extraction. CDDFuse [50] and EMMA [51] adopt dual-branch architectures combining CNNs and Transformers to simultaneously extract features from the input images and integrate them for fusion.

GAN-based methods enhance the model's feature extraction capabilities by introducing adversarial learning between generators and discriminators. Depending on the number of discriminators used, these methods can be classified into single-discriminator and dual-discriminator approaches. Single-discriminator methods [2, 52] tend to favor one modality over the other, potentially leading to information loss and reduced visual quality of the fusion results. To address this, dual-discriminator methods [53–56] are proposed to preserve important features from both source images simultaneously.

However, all of these methods primarily focus on designing effective feature extraction networks to produce high-quality fusion features and images. They overlook how fusion quality impacts downstream task performance, and fail to consider the potential feedback from downstream tasks that could help guide fusion more effectively.

## 2.3 Pedestrian detection

Pedestrian detection is a fundamental problem in computer vision with a wide range of applications. Cascade R-CNN [57] extends R-CNN [58] into a multi-stage framework, improving the ability to filter hard negative samples. Faster R-CNN [59] introduces a Region Proposal Network (RPN) that shares convolutional features with the detection network, making region proposals nearly cost-free. YOLO [60] reformulates object detection as a regression problem, allowing real-time inference directly on images through a convolutional neural network. SSD [61] uses multi-scale feature maps and predefined anchors for pedestrian detection, addressing YOLO's limitations in detecting small objects. DETR [62] adopts a Transformer-based encoder-decoder architecture for object detection. BAS Wu et al. [63] learns to represent the whole foreground region by leveraging foreground guidance and domain constraints. CREAM [64] proposes a clustering-based method to enhance activation within target regions. Group R-CNN [65] builds instance groups to perform pedestrian detection from point annotations.

However, most pedestrian detection methods are designed for unimodal images, which often leads to degraded detection performance due to incomplete scene information. In this work, we perform pedestrian detection on fused infrared and visible images, and incorporate task-specific prompts generated by large language

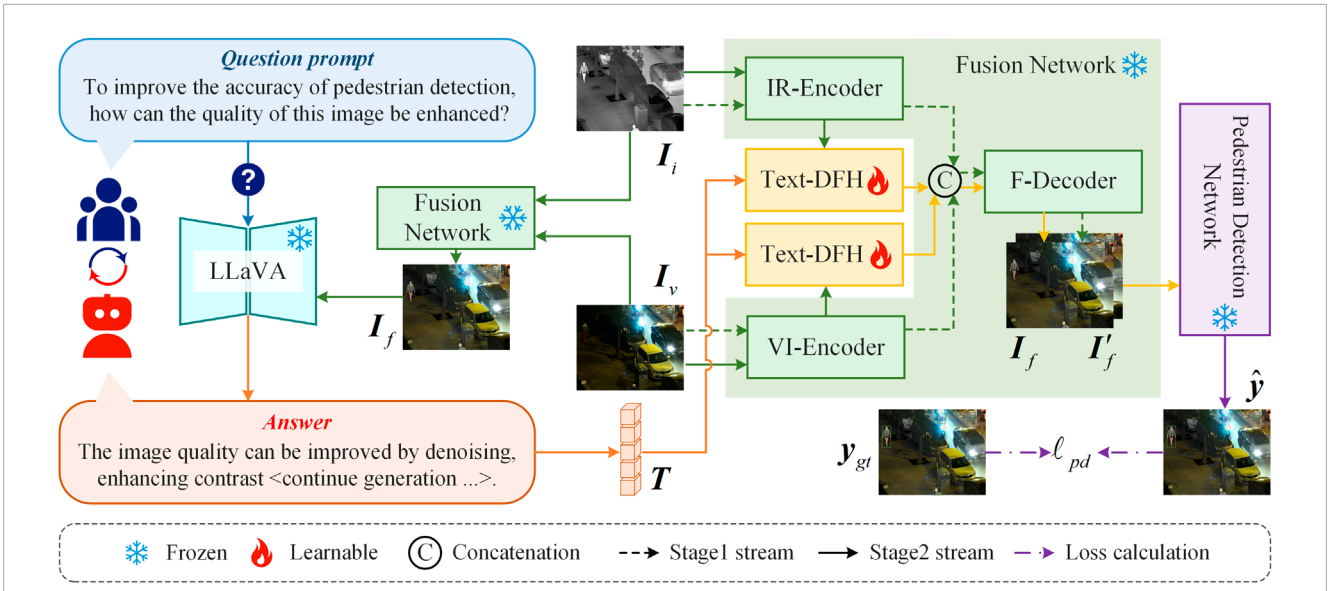


FIGURE 2
Overall framework of the proposed method. We use the IR-Encoder and VI-Encoder to extract features from the infrared and visible images, respectively. To ensure that the fused output meets the requirements of the pedestrian detection task, we input both a question related to pedestrian detection (e.g., To improve the accuracy of pedestrian detection, how can the quality of this image be enhanced?) and the unmodulated fused image into a Multimodal Large Language Model. The model provides suggestions for improving the quality of the fused image. Based on these suggestions, the Text-DFH module refines the output features of the fusion network, so that the final fusion result better aligns with the needs of the pedestrian detection task.

models. This not only improves the quality of the fused images but also enhances pedestrian detection performance.

# 3 Methods

## 3.1 Overview

As shown in Figure 2, the proposed method consists of two training stages. The first stage is dedicated to training the Fusion Network, enabling it to perform basic infrared and visible image fusion. In the second stage, the parameters of the pretrained fusion network are frozen, and a Text-Driven Feature Harmonization (Text-DFH) module is trained to refine the fusion results to better align with the requirements of pedestrian detection.The fusion network is composed of three main components: an Infrared Image Feature Encoder (IR-Encoder), a Visible Image Feature Encoder (VI-Encoder), and a Fusion Feature Decoder (F-Decoder). The IR/VI-Encoders are responsible for extracting features from the input infrared and visible images, respectively, while the F-Decoder reconstructs the fused image based on the combined features.The Text-DFH module adjusts the features extracted by the IR/VI-Encoders based on responses from a Multimodal Large Language Model (MLLM), ensuring that the resulting fused image better satisfies the needs of pedestrian detection. In this work, we adopt LLaVA [66] as the MLLM. LLaVA analyzes the unmodulated fused image and generates suggestions in response to user queries related to pedestrian detection tasks (e.g., To improve the accuracy of pedestrian detection, how can the quality of this image be enhanced?). More text examples of LLaVA answers are shown in Figure 3.

## 3.2 Feature extraction and fusion

In the first training stage, we train the fusion network to perform the basic task of infrared and visible image fusion. The fusion network primarily consists of three components: the IR-Encoder, VI-Encoder, and F-Decoder. Each of the IR-Encoder, VI-Encoder, and F-Decoder is composed of three feature extraction layers. Each layer is constructed by stacking a convolutional layer (kernel size = $3 \times 3$, stride = 1), a Batch Normalization layer, and a LeakyReLU activation function. It is worth noting that the LeakyReLU activation function in the final feature extraction layer of the F-Decoder is replaced with a Tanh activation function to facilitate image reconstruction. We input the infrared image $\mathbf{I}_i$ and the visible image $\boldsymbol{I}_v$ into the IR-Encoder and VI-Encoder, respectively, to extract features $\boldsymbol{F}_i$ and $\boldsymbol{F}_v$. To reconstruct the fused image, we concatenate $\boldsymbol{F}_i$ and $\boldsymbol{F}_v$ along the channel dimension and feed the result into the F-Decoder, which generates the final fused image $\boldsymbol{I}_f$.

To encourage the fused image to retain as much scene information from the source images as possible, we introduce an intensity loss $\ell_{in}$ and an edge loss $\ell_{ed}$, which together form the fusion loss $\ell_f$:

$$\ell_f = \ell_{in} + \varepsilon \ell_{ed}, \tag{1}$$

Here, $\varepsilon$ denotes a hyperparameter used to balance the contribution of each sub-loss term. The intensity loss $\ell_{in}$ is defined as:

$$\ell_{in} = \frac{1}{HW}\left(\left\|\boldsymbol{I}_f - \boldsymbol{I}_i\right\|_1 + \left\|\boldsymbol{I}_f - \boldsymbol{I}_v\right\|_1\right), \tag{2}$$

The edge loss $\ell_{ed}$ is defined as:

$$\ell_{ed} = \frac{1}{HW}\left(\left\|\nabla\boldsymbol{I}_f - \nabla\boldsymbol{I}_i\right\|_1 + \left\|\nabla\boldsymbol{I}_f - \nabla\boldsymbol{I}_v\right\|_1\right), \tag{3}$$

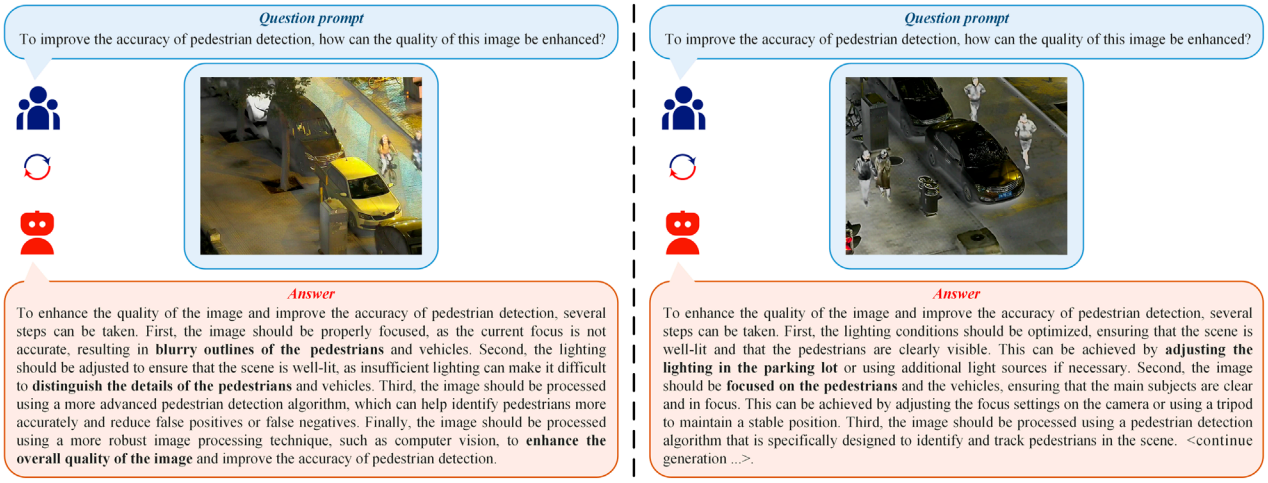


FIGURE 3
Visualized images of text examples of LLaVA answers.

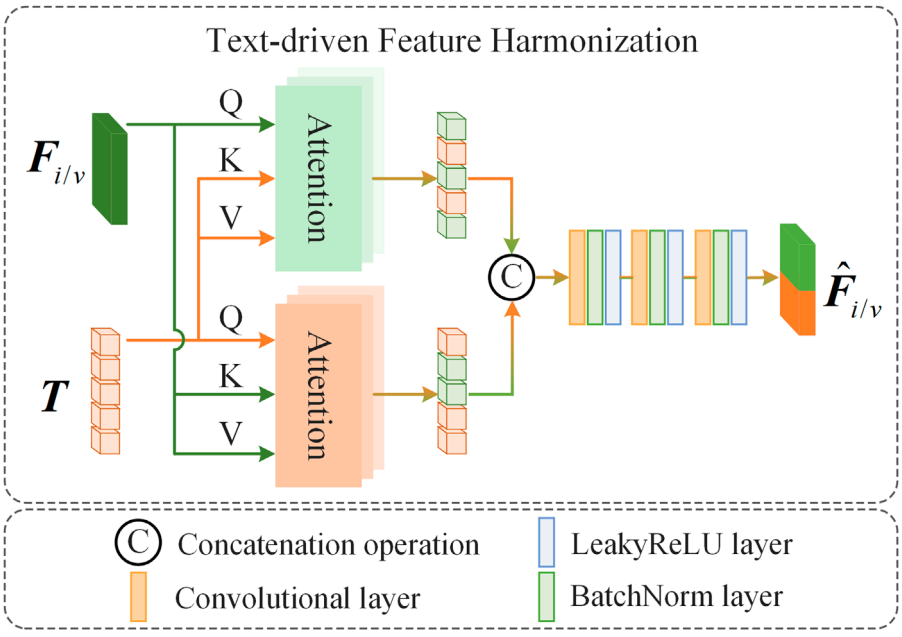


FIGURE 4
Text-driven feature harmonization module.

Here, $H$ and $W$ denote the height and width of the fused image, respectively; $\|\cdot\|_1$ represents the l1-norm, and $\nabla$ denotes the Sobel edge extraction operator.

## 3.3 Text-driven feature harmonization

In the second training stage, we freeze the parameters of the pretrained fusion network and focus on training the Text-DFH module to ensure that the fusion results meet the requirements of the pedestrian detection task. Text-DFH refines the features output by the IR/VI-Encoders in the fusion network based on the responses from the multimodal large language model, enabling the fused image to better align with the needs of pedestrian detection. As shown in Figure 4, Text-DFH mainly consists of a dual-branch Cross Attention (CA) module and three feature extraction layers. The dual-branch cross attention computes the cross-attention between the features extracted by the IR/VI-Encoders and the textual features, allowing the model to extract useful information from the text that can help improve pedestrian detection accuracy. Subsequently, the three feature extraction layers integrate this textual information with the image scene features to generate refined

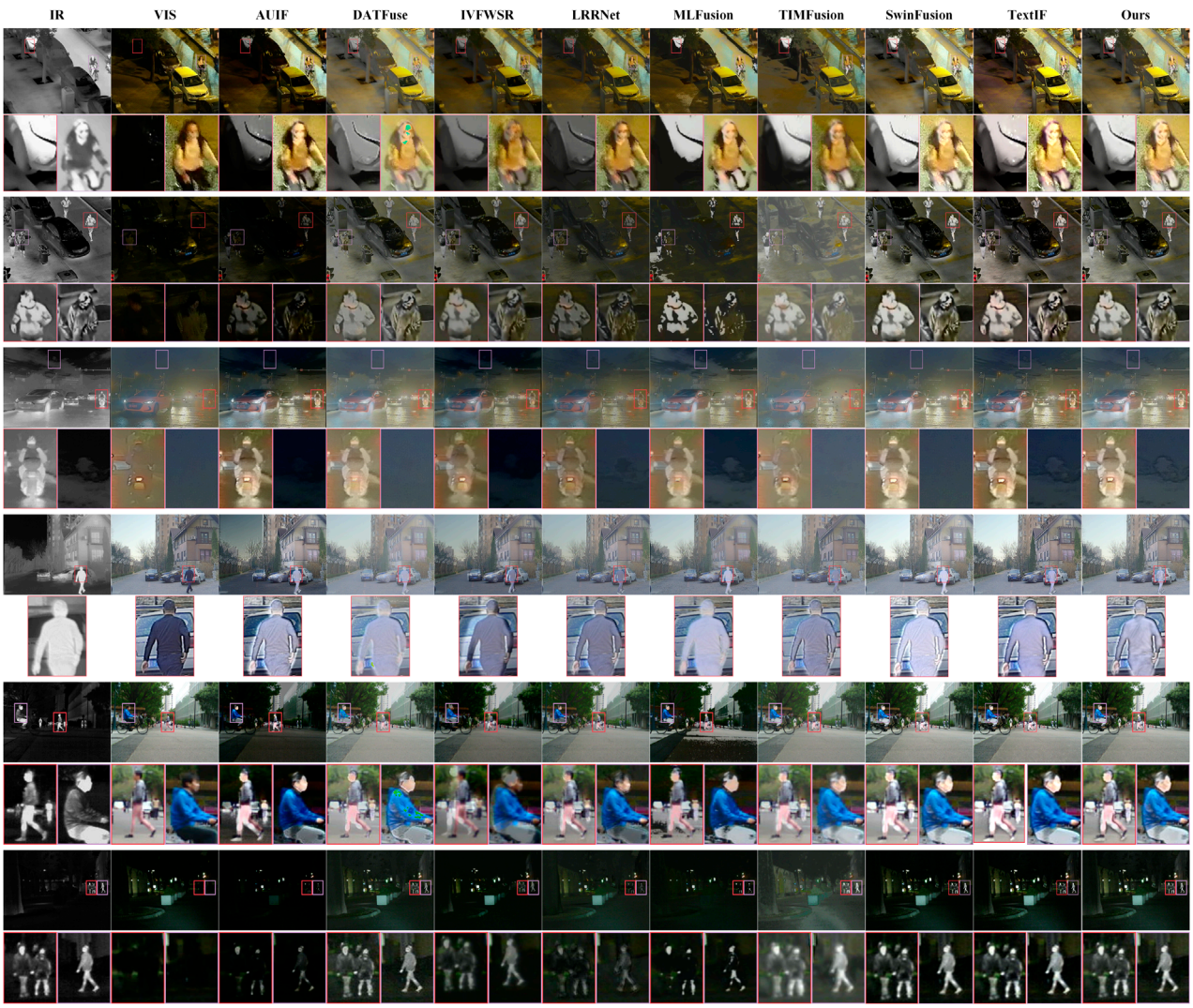


FIGURE 5
Visual comparison with SOTA methods. The top two rows, middle two rows, and bottom two rows of images are from the LLVIP, $M^3FD$, and MSRS datasets, respectively. The first and second columns show the infrared and visible source images, while the third to ninth columns display the fusion results produced by the compared methods.

features. The structure of the CA module is similar to the Multi-Scale Attention (MSA) module used in DATFuse.

We input the infrared image $\boldsymbol{I}_i$ and visible image $\boldsymbol{I}_v$ into the pretrained fusion network with frozen parameters to obtain the fused image $\boldsymbol{I}_f$. To obtain effective textual feedback that helps ensure the fused image meets the requirements of the pedestrian detection task, we input both $\boldsymbol{I}_f$ and the text prompt "To improve the accuracy of pedestrian detection, how can the quality of this image be enhanced?" into LLaVA, resulting in the textual feature $\boldsymbol{T}$. We then input the outputs $\boldsymbol{F}_{i/v}$ from the IR/VI-Encoders and the textual feature $\boldsymbol{T}$ into Text-DFH to harmonize the information in $\boldsymbol{F}_{i/v}$. To comprehensively extract the task-relevant information from the textual features, we design a dual-branch processing strategy. In the first branch, we take $\boldsymbol{F}_{i/v}$ as the Query (Q) and $\boldsymbol{T}$ as the Key (K) and Value (V) for cross-attention computation:

$$\boldsymbol{F}^1_{i/v} = \text{softmax}\left(\frac{\boldsymbol{Q}^1_{i/v}\left(\boldsymbol{K}^1_{i/v}\right)^T}{\sqrt{d_1}}\right)\boldsymbol{V}^1_{i/v}, \tag{4}$$

Here, $\boldsymbol{F}^1_{i/v}$ represents the features injected with textual information in the first branch, $d_1$ denotes the dimensionality of $\boldsymbol{Q}^1_{i/v}$, $\boldsymbol{Q}^1_{i/v} = \boldsymbol{W}^{Q,1}_{i/v}\boldsymbol{F}_{i/v}$, $\boldsymbol{K}^1_{i/v} = \boldsymbol{W}^{K,1}_{i/v}\boldsymbol{T}$, $\boldsymbol{V}^1_{i/v} = \boldsymbol{W}^{V,1}_{i/v}\boldsymbol{T}$. In the second branch, we use $\boldsymbol{T}$ as the Query (Q) and $\boldsymbol{F}_{i/v}$ as the Key (K) and Value (V) for cross-attention computation:

$$\boldsymbol{F}^2_{i/v} = \text{softmax}\left(\frac{\boldsymbol{Q}^2_{i/v}\left(\boldsymbol{K}^2_{i/v}\right)^T}{\sqrt{d_2}}\right)\boldsymbol{V}^2_{i/v}, \tag{5}$$

Here, $\boldsymbol{F}^2_{i/v}$ represents the features injected with textual information in the second branch, $d_2$ denotes the dimensionality of $\boldsymbol{Q}^2_{i/v}$, and $\boldsymbol{Q}^2_{i/v} = \boldsymbol{W}^{Q,2}_{i/v}\boldsymbol{T}, \boldsymbol{K}^2_{i/v} = \boldsymbol{W}^{K,2}_{i/v}\boldsymbol{F}_{i/v}, \boldsymbol{V}^2_{i/v} = \boldsymbol{W}^{V,2}_{i/v}\boldsymbol{F}_{i/v}$. To comprehensively

TABLE 1 Quantitative results on the LLVIP dataset. The best and second-best values for each evaluation metric are highlighted in red and blue, respectively.

| Methods | $Q_{AB/F}\uparrow$ | $Q_{CV}\downarrow$ | $Q_{SSIM}\uparrow$ | $Q_{AG}\uparrow$ | $Q_{SCD}\uparrow$ |
|---|---|---|---|---|---|
| AUIF | 0.3869 | 610.74 | 1.2016 | 3.5256 | 1.3413 |
| DATFuse | 0.4548 | 453.42 | 1.3130 | 3.1243 | 1.3351 |
| IVFWSR | 0.2925 | 512.77 | 1.2348 | 2.5252 | 1.1235 |
| LRRNet | 0.4426 | 534.89 | 1.3022 | 2.4625 | 0.9999 |
| MLFusion | 0.3239 | 523.41 | 1.2624 | 2.1613 | 0.9966 |
| TIMFusion | 0.2325 | 845.75 | 1.1742 | 2.1761 | 0.5368 |
| SwinFusion | 0.4266 | 598.53 | 1.2743 | 2.6346 | 1.3527 |
| TextIF | 0.5235 | 356.35 | 1.3056 | 3.4856 | 1.4527 |
| Ours | 0.5845 | 287.43 | 1.3441 | 3.9867 | 1.5462 |

TABLE 2 Quantitative results on the M$^3$FD dataset. The best and second-best values for each evaluation metric are highlighted in red and blue, respectively.

| Methods | $Q_{AB/F}\uparrow$ | $Q_{CV}\downarrow$ | $Q_{SSIM}\uparrow$ | $Q_{AG}\uparrow$ | $Q_{SCD}\uparrow$ |
|---|---|---|---|---|---|
| AUIF | 0.5425 | 852.56 | 1.3003 | 6.6735 | 1.5353 |
| DATFuse | 0.4854 | 563.57 | 1.3067 | 4.8326 | 1.3461 |
| IVFWSR | 0.4532 | 722.22 | 1.2735 | 3.5628 | 1.2452 |
| LRRNet | 0.5164 | 579.55 | 1.3735 | 4.5624 | 1.3461 |
| MLFusion | 0.4253 | 689.44 | 1.2835 | 4.4527 | 1.2687 |
| TIMFusion | 0.5352 | 616.16 | 1.2872 | 4.3336 | 1.2004 |
| SwinFusion | 0.5537 | 588.24 | 1.3086 | 6.0463 | 1.3456 |
| TextIF | 0.5423 | 534.21 | 1.2986 | 6.4026 | 1.5035 |
| Ours | 0.5856 | 454.45 | 1.3095 | 6.4561 | 1.6187 |

aggregate the textual information, we concatenate $\boldsymbol{F}^1_{i/v}$ and $\boldsymbol{F}^2_{i/v}$ along the channel dimension and feed the result into three feature extraction layers to obtain the harmonized features $\widehat{\boldsymbol{F}}_{i/v}$. We then concatenate $\widehat{\boldsymbol{F}}_i$ and $\widehat{\boldsymbol{F}}_v$ along the channel dimension and input the result into the F-Decoder to reconstruct the refined fused image $\boldsymbol{I'}_f$.

To ensure that the refined fused image $\boldsymbol{I'}_f$ meets the requirements of the pedestrian detection task, we introduce a pretrained pedestrian detection network with frozen parameters to supervise the fused image. We input $\boldsymbol{I'}_f$ into the detection network and obtain the pedestrian detection result $\widehat{\boldsymbol{y}}$. To make $\widehat{\boldsymbol{y}}$ as close as possible to its ground truth $\boldsymbol{y}_{gt}$, we constrain the Text-DFH module using the loss function $\ell_{pd}$, which is the same as the one used during the training of YOLOv5.

TABLE 3 Quantitative results on the MSRS dataset. The best and second-best values for each evaluation metric are highlighted in red and blue, respectively.

| Methods | $Q_{AB/F}\uparrow$ | $Q_{CV}\downarrow$ | $Q_{SSIM}\uparrow$ | $Q_{AG}\uparrow$ | $Q_{SCD}\uparrow$ |
|---|---|---|---|---|---|
| AUIF | 0.1736 | 799.97 | 0.9853 | 1.8844 | 1.1963 |
| DATFuse | 0.6326 | 416.67 | 1.2421 | 3.5481 | 1.5641 |
| IVFWSR | 0.3464 | 734.46 | 1.3462 | 2.1129 | 1.3581 |
| LRRNet | 0.4263 | 666.35 | 1.2952 | 2.5632 | 1.0854 |
| MLFusion | 0.2656 | 745.57 | 1.3457 | 2.6531 | 1.2053 |
| TIMFusion | 0.3346 | 1032.24 | 1.1003 | 2.6422 | 1.1783 |
| SwinFusion | 0.4527 | 439.46 | 1.3163 | 3.0042 | 1.4828 |
| TextIF | 0.6125 | 400.34 | 1.3357 | 3.6426 | 1.5457 |
| Ours | 0.6365 | 334.23 | 1.3537 | 3.5474 | 1.6854 |

# 4 Experiments

## 4.1 Datasets

The proposed method consists of two training stages. In both the first and second training stages, we train the fusion network and the text-driven feature harmonization module on the publicly available LLVIP dataset [67], respectively, in accordance with standard practices in the field [68–70]. Specifically, we randomly select 2,000 pairs of infrared and visible images from the LLVIP dataset as the training set. To enhance the diversity of training samples, we apply random flipping, random rotation, and random cropping as data augmentation techniques. For evaluation, we randomly select 200 pairs of infrared and visible images from each of the LLVIP, M$^3$FD [71], and MSRS [3] datasets to form the test set, in order to assess both the fusion performance and pedestrian detection performance of the proposed method. Among them, LLVIP, M$^3$FD, and MSRS are used to evaluate fusion performance, while LLVIP is specifically used to evaluate pedestrian detection performance.

## 4.2 Implementation details

The proposed method involves two training stages. In the first stage, the fusion network is trained. In the second stage, the parameters of the fusion network are frozen, and the text-driven feature harmonization module is trained. Both training stages use the Adam optimizer to update the network parameters, with a batch size of 16 and a learning rate of $1\times10^{-3}$. The total number of training epochs is set to 100 for the first stage and 200 for the second stage. In addition, the hyperparameter $\varepsilon$ is set to 0.2. The proposed method is implemented based on the PyTorch framework and is trained on a single NVIDIA RTX A6000 GPU.

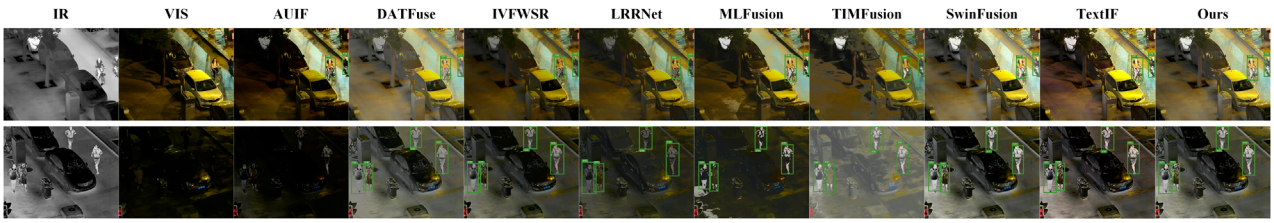


FIGURE 6
Qualitative comparison of pedestrian detection performance with "retraining methods." The first and second columns show the infrared and visible source images, while the third to ninth columns display the pedestrian detection results of the compared methods.

TABLE 4 Quantitative comparison of pedestrian detection performance with "retraining methods." The best and second-best values for each evaluation metric are highlighted in red and blue, respectively.

| Methods | $mAP_{50}\uparrow$ | $mAP_{75}\uparrow$ | $mAP_{50\rightarrow95}\uparrow$ |
|---|---|---|---|
| AUIF | 98.2 | 91.8 | 74.4 |
| DATFuse | 99.0 | 91.5 | 74.3 |
| IVFWSR | 97.2 | 89.6 | 72.9 |
| LRRNet | 98.0 | 90.8 | 73.8 |
| MLFusion | 97.8 | 89.9 | 73.6 |
| TIMFusion | 97.9 | 88.4 | 74.0 |
| SwinFusion | 98.5 | 90.4 | 74.3 |
| TextIF | 98.9 | 91.7 | 74.6 |
| Ours | 99.1 | 92.8 | 75.0 |

## 4.3 Evaluation metrics

We adopt five commonly used objective evaluation metrics to quantitatively assess the fusion performance of the proposed method. These metrics include Edge Preservation Index ($Q_{AB/F}$) [72, 73], Chen-Varshney Index ($Q_{CV}$) [74], Structural Similarity Index ($Q_{SSIM}$) [75], Average Gradient ($Q_{AG}$) [76], and Sum of Correlations of Differences ($Q_{SCD}$) [77]. $Q_{AB/F}$ measures how well edge information from the source images is preserved in the fused image. $Q_{AB/F}$ higher value indicates less loss of texture details in the fused image. $Q_{CV}$ evaluates fusion quality from the perspective of human visual perception; a lower value means the fused image aligns better with human visual preferences. $Q_{SSIM}$ quantifies the similarity between the fused image and the source images in terms of luminance, contrast, and structure. A higher value indicates less information difference between the fused and source images. $Q_{AG}$ measures the richness of gradient information in the fused image. A higher value means the fused image contains more detailed gradient content. $Q_{SCD}$ assesses information loss during the fusion process by computing difference maps between the fused image and source images. A higher value indicates less distortion in the fused image. Among these, $Q_{AB/F}$, $Q_{SSIM}$, $Q_{AG}$ and $Q_{SCD}$ are positive indicators, meaning a higher value indicates better fusion performance. $Q_{CV}$ is a negative indicator, meaning a lower value represents better fusion performance. In addition, to objectively evaluate the effectiveness of the fused images in the pedestrian detection task, we adopt three widely used metrics in the pedestrian detection domain for quantitative analysis: Mean Average Precision (mAP) at IoU threshold of 0.5 ($mAP_{50}$), mAP at IoU threshold of 0.75 ($mAP_{75}$), and the averaged mAP at IoU threshold from 0.5 to 0.95 ($mAP_{50\rightarrow95}$).

## 4.4 Comparison with state-of-the-art methods

In this study, we conduct a series of qualitative and quantitative comparisons between the proposed method and eight state-of-the-art (SOTA) methods to verify its superiority in both fusion performance and pedestrian detection performance. These methods include AUIF [78], DATFuse [49], IVFWSR [79], LRRNet [80], MLFusion [81], TIMFusion [82], SwinFusion [16], and TextIF [29]. The comparative experiments are divided into two distinct groups: In the first group, we compare the fusion performance of our method with that of the SOTA methods. In the second group, we freeze the fusion networks of the compared methods and retrain their pedestrian detection networks using the corresponding fused results. The retrained detection networks are then used to perform pedestrian detection on the fused images. This setup is designed to demonstrate that our proposed method can achieve strong pedestrian detection performance without requiring retraining of the detection network.

### 4.4.1 Fusion performance comparison

We conduct both quantitative and qualitative comparisons of the proposed method against AUIF, DATFuse, IVFWSR, LRRNet, MLFusion, TIMFusion, SwinFusion, and TextIF on the LLVIP, $M^3FD$, and MSRS datasets to validate the superiority of our method in terms of fusion performance. As shown in the enlarged regions of Figure 5, our method effectively highlights the thermal radiation information from the infrared image while preserving fine texture details from the visible image. Compared to existing SOTA methods, the fused images produced by our method exhibit clearer local details as well as higher overall brightness and contrast at the global level. This not only improves visual quality but also facilitates better object recognition in downstream tasks. This

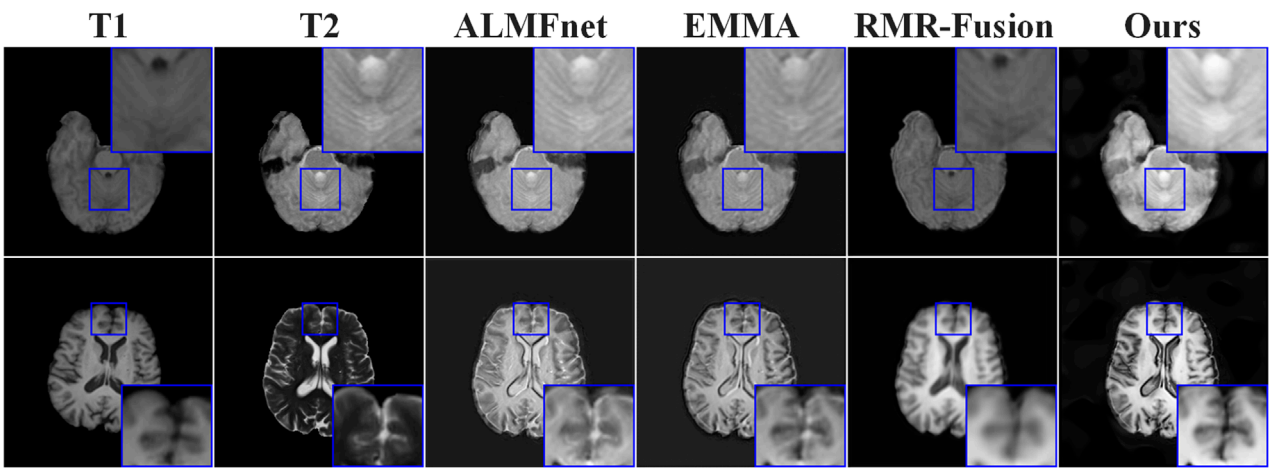


FIGURE 7
Qualitative analysis results on the medical image fusion task.

TABLE 5 Quantitative Analysis Results on the Medical Image Fusion Task. The best and second-best values for each evaluation metric are highlighted in red and blue, respectively.

| Methods | $Q_{AB/F}\uparrow$ | $Q_{CV}\downarrow$ | $Q_{SSIM}\uparrow$ | $Q_{AG}\uparrow$ | $Q_{SCD}\uparrow$ |
|---|---|---|---|---|---|
| ALMFnet | 0.4700 | 1330.59 | 1.3432 | 3.5826 | 1.2991 |
| EMMA | 0.4682 | 1288.99 | 1.3232 | 3.1826 | 1.2999 |
| RMR-Fusion | 0.4419 | 1344.12 | 1.2967 | 3.2621 | 1.3781 |
| Ours | 0.4792 | 1203.12 | 1.3631 | 3.7521 | 1.3629 |

advantage is also reflected in the quantitative evaluation results, as shown in Tables 1–3. Specifically, our method achieves the lowest values in metric $Q_{CV}$, and ranks first in both metrics $Q_{AB/F}$ and $Q_{AG}$, indicating that the fused images contain richer edge information and are more consistent with human visual perception. In summary, both qualitative and quantitative results demonstrate that our proposed method offers significant improvements in fusion performance over the compared methods.

### 4.4.2 Pedestrian detection performance comparison

A common practice to improve the performance of fusion networks in downstream tasks is to freeze the parameters of the fusion network and retrain the downstream task network based on the generated fused results. Such approaches are referred to as "retraining methods." To evaluate the effectiveness of our proposed method in pedestrian detection, we perform both quantitative and qualitative comparisons against these retraining methods. As shown in Figure 6, the pedestrian detection results of other methods often suffer from issues such as bounding boxes that fail to fully cover the pedestrians' bodies, or boxes that include large amounts of irrelevant background, indicating insufficient detection accuracy. In contrast, the detection results produced by our method show significantly fewer irrelevant regions within the bounding boxes and more accurate box placement. This advantage is also clearly reflected in the quantitative results, as shown in Table 4. Our method achieves the highest scores in metrics $mAP_{50}$, $mAP_{75}$, and $mAP_{50\rightarrow95}$, indicating superior performance in the pedestrian detection task compared to the other methods. In conclusion, our method demonstrates better performance than approaches that require retraining the pedestrian detection network, even without retraining. This highlights the effectiveness and advantage of our method in pedestrian detection tasks.

### 4.4.3 Analysis of application potential in medical image fusion

Furthermore, to validate the effectiveness and application potential of the proposed method in the field of nuclear medical imaging, we further deployed it in a medical image fusion task. Specifically, we conducted experiments on the BraTS2020 [83] dataset and performed both qualitative and quantitative analyses of the fusion results. As shown in Figure 7, compared with state-of-the-art methods such as ALMFnet [84, 85], and RMR-Fusion [86], the proposed method preserves more texture details and salient information in the fused medical images. As reported in Table 5, our method ranks first or second across most evaluation metrics. These results demonstrate the promising potential of the proposed method for applications in nuclear medical imaging.

## 4.5 Ablation study

The proposed method mainly consists of two core components: the Multimodal Large Language Model (MLLM) and the Text-Driven Feature Harmonization (Text-DFH) module. Within Text-DFH, both the text-guided cross-attention and the image-guided cross-attention play key roles. To validate the effectiveness of these components, we conduct a series of ablation experiments on the LLVIP dataset.

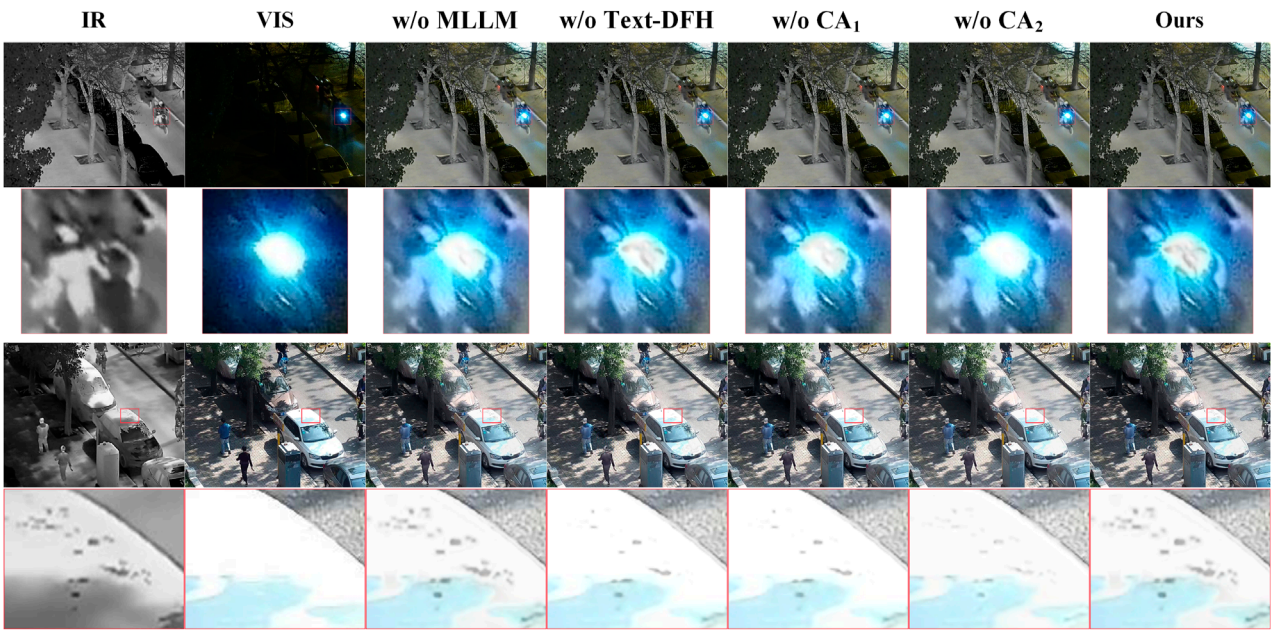


FIGURE 8
Qualitative comparison of fusion performance across different ablation models. The first and second columns show the infrared and visible source images, while the third to seventh columns display the fusion results obtained under different ablation settings.

TABLE 6 Quantitative comparison of fusion performance across different ablation models. The best and second-best values for each evaluation metric are highlighted in red and blue, respectively.

| Methods | $Q_{AB/F}\uparrow$ | $Q_{CV}\downarrow$ | $Q_{SSIM}\uparrow$ | $Q_{AG}\uparrow$ | $Q_{SCD}\uparrow$ |
|---|---|---|---|---|---|
| w/o MLLM | 0.5472 | 298.75 | 1.3244 | 3.6433 | 1.5367 |
| w/o Text-DFH | 0.5763 | 299.46 | 1.3321 | 3.4131 | 1.4992 |
| w/o CA1 | 0.5834 | 305.92 | 1.3234 | 3.6362 | 1.5213 |
| w/o CA2 | 0.5798 | 301.68 | 1.3401 | 3.6524 | 1.5123 |
| Ours | 0.5845 | 287.43 | 1.3441 | 3.9867 | 1.5462 |

TABLE 7 Quantitative comparison of pedestrian detection performance across different ablation models. The best and second-best values for each evaluation metric are highlighted in red and blue, respectively.

| Methods | $mAP_{50}\uparrow$ | $mAP_{75}\uparrow$ | $mAP_{50\rightarrow95}\uparrow$ |
|---|---|---|---|
| w/o MLLM | 98.5 | 91.6 | 73.9 |
| w/o Text-DFH | 98.8 | 92.1 | 74.0 |
| w/o CA1 | 99.0 | 92.4 | 74.5 |
| w/o CA2 | 98.9 | 91.8 | 74.4 |
| Ours | 99.1 | 92.8 | 75.0 |

### 4.5.1 Effectiveness of the multimodal large language model

We utilize the MLLM to analyze the fused images based on user-provided questions related to pedestrian detection performance and generate suggestions for improving image quality. To assess the contribution of the MLLM, we remove it and replace its feedback with a fixed text prompt: "Brighter brightness, higher contrast, and clearer texture details." As shown in Figure 8, the fusion results from the ablation model without the MLLM are noticeably inferior in visual quality compared to the full model. To further validate this, we perform quantitative analysis as presented in Table 6. The results show that the full model outperforms the ablation model on all evaluation metrics. Additionally, we analyze the performance of pedestrian detection, as shown in Table 7 and Figure 9. Both the quantitative and qualitative results indicate that the fused images produced by the ablation model without the MLLM lead to poorer detection performance. In contrast, the full model achieves better pedestrian detection results. In summary, both qualitative and quantitative analyses confirm the effectiveness of the Multimodal Large Language Model in our method.

### 4.5.2 Effectiveness of Text-DFH

Text-DFH refines the output features of the fusion network based on suggestions from the multimodal large language model, enabling the fused image to better meet the requirements of the pedestrian detection task. To verify the effectiveness of Text-DFH, we remove it from the architecture and instead concatenate the text features with the image features to be refined along the channel dimension. The combined features are then processed by CNNs to obtain the refined output. We conduct both quantitative and qualitative analyses of the fusion


FIGURE 9
Qualitative comparison of pedestrian detection performance across different ablation models. The first and second columns show the infrared and visible source images, while the third to seventh columns display the pedestrian detection results under different ablation settings.

performance of the model without Text-DFH, as shown in Table 6 and Figure 8. As observed, the ablation model without Text-DFH performs worse than the full model across multiple evaluation metrics, and the visual quality of the fused images is also inferior. In addition, we evaluate pedestrian detection performance both quantitatively and qualitatively, as presented in Table 7 and Figure 9. The full model achieves higher scores compared to the ablation model without Text-DFH. In summary, a series of experiments clearly demonstrate the effectiveness of the Text-DFH module.

### 4.5.3 Effectiveness of dual-branch cross attention

In the Text-DFH module, we refine image features using text features through a dual-branch cross attention mechanism. To verify its effectiveness, we remove the cross attention from each branch individually, leaving only a single branch to refine the image features. These variants are referred to as CA1 and CA2, respectively. From the quantitative and qualitative results on fusion performance, it is evident that removing either branch of the cross attention leads to a significant drop in performance, as shown in Table 6 and Figure 8. Furthermore, to assess the impact of dual-branch cross attention on pedestrian detection performance, we conduct both quantitative and qualitative analyses. The results demonstrate that pedestrian detection performance is optimal only when both branches of the cross attention are used to refine the image features, as shown in Table 7 and Figure 9. In conclusion, the above experiments confirm the effectiveness of the dual-branch cross attention mechanism.

## 5 Conclusion

To address the limitation of existing methods that primarily focus on improving fused image quality through network design—while overlooking the potential benefits of enhanced image quality for pedestrian detection—we propose a multimodal large language model (MLLM)-driven infrared and visible image fusion method. This method not only aims to improve the quality of the fused images but also emphasizes enhancing their performance in pedestrian detection tasks. By leveraging a multimodal large language model, we analyze the fused images based on user-provided questions related to improving pedestrian detection performance and generate suggestions for enhancing image quality. To fully utilize the guidance provided by the MLLM, we design a Text-Driven Feature Harmonization (Text-DFH) module, which refines the features output by the fusion network according to the textual suggestions. This ensures improved fusion quality while maintaining strong performance in pedestrian detection. In addition, the proposed method also demonstrates significant application potential in the field of nuclear medical imaging. However, under extreme weather conditions such as rain, fog, and snow, the fusion performance of the current method may degrade. Moreover, when such methods are applied to other types of source images [87–90], their performance may degrade. In future work, we plan to extend this research to develop an infrared and visible image fusion framework tailored for extreme weather scenarios, striving to maintain robust downstream task performance even in challenging environments.

## 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

KW: Project administration, Writing – original draft, Writing – review and editing, Investigation, Conceptualization, Methodology. DH: Formal Analysis, Writing – review and editing, Data curation, Validation. YaC: Visualization, Supervision, Writing – review and editing, Resources. YkC: Funding acquisition, Project administration, Supervision, Writing – review and editing, Writing – original draft. YL: Validation, Writing – review and editing, Visualization, Formal Analysis. ZJ: Writing – review and editing, Investigation, Data curation, Resources. FC: Formal Analysis,

Writing – review and editing, Data curation. WL: Writing – review and editing, Resources, Visualization.

## Funding

The author(s) declare that financial support was received for the research and/or publication of this article. Science and Technology Project of China Southern Power Grid Co., Ltd. (No. YNKJXM20240052).

## Conflict of interest

Authors KW, DH, YaC, YkC, YL, ZJ, FC, and WL were employed by Yunnan Power Grid Co., Ltd.

## Generative AI statement

The author(s) declare that Generative AI was used in the creation of this manuscript. AI was only used to polish the paper.

## Publisher's note



## References

1. Li H, Wu X-J, Kittler J. Rfn-nest: an end-to-end residual fusion network for infrared and visible images. *Inf Fusion* (2021) 73:72–86. doi:10.1016/j.inffus.2021.02.023

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

3. Tang L, Yuan J, Zhang H, Jiang X, Ma J. Piafusion: a progressive infrared and visible image fusion network based on illumination aware. *Inf Fusion* (2022) 83-84:79–92. doi:10.1016/j.inffus.2022.03.007

4. Xu M, Tang L, Zhang H, Ma J. Infrared and visible image fusion via parallel scene and texture learning. *Pattern Recognition* (2022) 132:108929. doi:10.1016/j.patcog.2022.108929

5. Du K, Li H, Zhang Y, Yu Z. Chitnet: a complementary to harmonious information transfer network for infrared and visible image fusion. *IEEE Trans Instrumentation Meas* (2025) 74:1–17. doi:10.1109/TIM.2025.3527523

6. Shi Y, Liu Y, Cheng J, Wang ZJ, Chen X. Vdmufusion: a versatile diffusion model-based unsupervised framework for image fusion. *IEEE Trans Image Process* (2025) 34:441–54. doi:10.1109/tip.2024.3512365

7. Lv G, Sima C, Gao Y, Dong A, Ma G, Cheng J. Sigfusion: semantic information-guided infrared and visible image fusion. *IEEE Trans Instrumentation Meas* (2024) 73:1–18. doi:10.1109/tim.2024.3457951

8. Fu H, Wang S, Duan P, Xiao C, Dian R, Li S, et al. Lraf-net: long-range attention fusion network for visible–infrared object detection. *IEEE Trans Neural Networks Learn Syst* (2024) 35:13232–45. doi:10.1109/tnnls.2023.3266452

9. Li Y, Pang Y, Cao J, Shen J, Shao L. Improving single shot object detection with feature scale unmixing. *IEEE Trans Image Process* (2021) 30:2708–21. doi:10.1109/tip.2020.3048630

10. Zhao Z-Q, Zheng P, Xu S-T, Wu X. Object detection with deep learning: a review. *IEEE Trans Neural Networks Learn Syst* (2019) 30:3212–32. doi:10.1109/tnnls.2018.2876865

11. Liu Y, Zeng J, Tao X, Fang G. Rethinking self-supervised semantic segmentation: achieving end-to-end segmentation. *IEEE Trans Pattern Anal Machine Intelligence* (2024) 46:10036–46. doi:10.1109/tpami.2024.3432326

12. Wu L, Fang L, He X, He M, Ma J, Zhong Z. Querying labeled for unlabeled: cross-image semantic consistency guided semi-supervised semantic segmentation. *IEEE Trans Pattern Anal Machine Intelligence* (2023) 45:8827–44. doi:10.1109/TPAMI.2022.3233584

13. Zhao S, Zhang Q. A feature divide-and-conquer network for rgb-t semantic segmentation. *IEEE Trans Circuits Syst Video Technology* (2023) 33:2892–905. doi:10.1109/tcsvt.2022.3229359

14. Zhao S, Liu Y, Jiao Q, Zhang Q, Han J. Mitigating modality discrepancies for rgb-t semantic segmentation. *IEEE Trans Neural Networks Learn Syst* (2024) 35:9380–94. doi:10.1109/tnnls.2022.3233089

15. Li H, Wu X. Densefuse: a fusion approach to infrared and visible images. *IEEE Trans Image Process* (2019) 28:2614–23. doi:10.1109/tip.2018.2887342

16. Ma J, Tang L, Fan F, Huang J, Mei X, Ma Y. Swinfusion: cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA J Automatica Sinica* (2022) 9:1200–17. doi:10.1109/jas.2022.105686

17. Tang W, He F, Liu Y. Ydtr: infrared and visible image fusion via y-shape dynamic transformer. *IEEE Trans Multimedia* (2023) 25:5413–28. doi:10.1109/tmm.2022.3192661

18. Li H, Qiu H, Yu Z, Zhang Y. Infrared and visible image fusion scheme based on nsct and low-level visual features. *Infrared Phys and Technology* (2016) 76:174–84. doi:10.1016/j.infrared.2016.02.005

19. Li H, Wang Y, Yang Z, Wang R, Li X, Tao D. Discriminative dictionary learning-based multiple component decomposition for detail-preserving noisy image fusion. *IEEE Trans Instrumentation Meas* (2020) 69:1082–102. doi:10.1109/tim.2019.2912239

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

21. Ma J, Zhang H, Shao Z, Liang P, Xu H. Ganmcc: a generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Trans Instrumentation Meas* (2021) 70:1–14. doi:10.1109/tim.2020.3038013

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

23. Zhang Y, Yang M, Li N, Yu Z. Analysis-synthesis dictionary pair learning and patch saliency measure for image fusion. *Signal Process.* (2020) 167:107327. doi:10.1016/j.sigpro.2019.107327

24. Tang L, Zhang H, Xu H, Ma J. Rethinking the necessity of image fusion in high-level vision tasks: a practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. *Inf Fusion* (2023) 99:101870. doi:10.1016/j.inffus.2023.101870

25. Liu J, Liu Z, Wu G, Ma L, Liu R, Zhong W, et al. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In: *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (2023). p. 8115–24.

26. Zhang H, Zuo X, Jiang J, Guo C, Ma J. Mrfs: mutually reinforcing image fusion and segmentation. In: *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (2024). p. 26964–73.

27. Wang D, Liu J, Liu R, Fan X. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection. *Inf Fusion* (2023) 98:101828. doi:10.1016/j.inffus.2023.101828

28. Yang Z, Zhang Y, Li H, Liu Y. Instruction-driven fusion of infrared–visible images: tailoring for diverse downstream tasks. *Inf Fusion* (2025) 121:103148. doi:10.1016/j.inffus.2025.103148

29. Yi X, Xu H, Zhang H, Tang L, Ma J. Text-if: leveraging semantic text guidance for degradation-aware and interactive image fusion. In: *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (2024). p. 27016–25.

30. Wang H, Zhang H, Yi X, Xiang X, Fang L, Ma J. Terf: text-driven and region-aware flexible visible and infrared image fusion. In: *Proceedings of the 32nd ACM international conference on multimedia* (2024). p. 935–44.

31. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

32. Team G, Anil R, Borgeaud S, Alayrac J-B, Yu J, Soricut R, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

33. Wu S, Fei H, Qu L, Ji W, Chua T-S. Next-gpt: any-to-any multimodal llm. In: *Forty-first international conference on machine learning* (2024).

34. Liu Y, Zhang K, Li Y, Yan Z, Gao C, Chen R, et al. Sora: a review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402* (2024).

35. Li K, He Y, Wang Y, Li Y, Wang W, Luo P, et al. Videochat: chat-centric video understanding. *arXiv preprint arXiv:2305.06355* (2023).

36. Lin B, Ye Y, Zhu B, Cui J, Ning M, Jin P, et al. Video-llava: learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311* (2023).

37. Yang Z, Li Y, Tang X, Xie M. Mgfusion: a multimodal large language model-guided information perception for infrared and visible image fusion. *Front Neurorobotics* (2024) 18:1521603. doi:10.3389/fnbot.2024.1521603

38. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: *International conference on machine learning (PmLR)* (2021). p. 8748–63.

39. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, et al. Llama: open and efficient foundation language models. *arXiv preprint arXiv:2302* (2023).

40. Xu H, Ma J, Jiang J, Guo X, Ling H. U2fusion: a unified unsupervised image fusion network. *IEEE Trans Pattern Anal Machine Intelligence* (2022) 44:502–18. doi:10.1109/tpami.2020.3012548

41. Zhang Y, Liu Y, Sun P, Yan H, Zhao X, Zhang L. Ifcnn: a general image fusion framework based on convolutional neural network. *Inf Fusion* (2020) 54:99–118. doi:10.1016/j.inffus.2019.07.011

42. Zhang H, Ma J. Sdnet: a versatile squeeze-and-decomposition network for real-time image fusion. *Int J Computer Vis* (2021) 129:2761–85. doi:10.1007/s11263-021-01501-8

43. Li H, Yang Z, Zhang Y, Jia W, Yu Z, Liu Y. Mulfs-cap: multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Trans Pattern Anal Machine Intelligence* (2025) 47:3673–90. doi:10.1109/TPAMI.2025.3535617

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

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

46. 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.

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

48. Li J, Zhu J, Li C, Chen X, Yang B. Cgtf: convolution-guided transformer for infrared and visible image fusion. *IEEE Trans Instrumentation Meas* (2022) 71:1–14. doi:10.1109/tim.2022.3175055

49. Tang W, He F, Liu Y, Duan Y, Si T. Datfuse: infrared and visible image fusion via dual attention transformer. *IEEE Trans Circuits Syst Video Technology* (2023) 33:3159–72. doi:10.1109/tcsvt.2023.3234340

50. Zhao Z, Bai H, Zhang J, Zhang Y, Xu S, Lin Z, et al. Cddfuse: correlation-driven dual-branch feature decomposition for multi-modality image fusion. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (2023). p. 5906–16.

51. Zhao Z, Bai H, Zhang J, Zhang Y, Zhang K, Xu S, et al. Equivariant multi-modality image fusion. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2024). p. 25912–21.

52. Ma J, Liang P, Yu W, Chen C, Guo X, Wu J, et al. Infrared and visible image fusion via detail preserving adversarial learning. *Inf Fusion* (2020) 54:85–98. doi:10.1016/j.inffus.2019.07.005

53. Ma J, Xu H, Jiang J, Mei X, Zhang X. Ddcgan: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans Image Process* (2020) 29:4980–95. doi:10.1109/tip.2020.2977573

54. Li J, Huo H, Li C, Wang R, Feng Q. Attentionfgan: infrared and visible image fusion using attention-based generative adversarial networks. *IEEE Trans Multimedia* (2021) 23:1383–96. doi:10.1109/tmm.2020.2997127

55. Zhou H, Wu W, Zhang Y, Ma J, Ling H. Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network. *IEEE Trans Multimedia* (2021) 25:635–48. doi:10.1109/tmm.2021.3129609

56. Zhang H, Yuan J, Tian X, Ma J. Gan-fm: infrared and visible image fusion using gan with full-scale skip connection and dual markovian discriminators. *IEEE Trans Comput Imaging* (2021) 7:1134–47. doi:10.1109/tci.2021.3119954

57. Cai Z, Vasconcelos N. Cascade r-cnn: high quality object detection and instance segmentation. *IEEE Trans pattern Anal machine intelligence* (2019) 43:1483–98. doi:10.1109/tpami.2019.2956516

58. 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.

59. 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.

60. 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.

61. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, et al. Ssd: single shot multibox detector. In: *Computer vision–ECCV 2016: 14th European conference, Amsterdam, The Netherlands, october 11–14, 2016, proceedings, Part I 14*. Springer (2016). p. 21–37.

62. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: *European conference on computer vision*. Springer (2020). p. 213–29.

63. Wu P, Zhai W, Cao Y. Background activation suppression for weakly supervised object localization. In: *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE (2022). p. 14228–37.

64. Xu J, Hou J, Zhang Y, Feng R, Zhao R-W, Zhang T, et al. Cream: weakly supervised object localization via class re-activation mapping. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022). p. 9437–46.

65. Zhang S, Yu Z, Liu L, Wang X, Zhou A, Chen K. Group r-cnn for weakly semi-supervised object detection with points. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022). p. 9417–26.

66. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. *Adv Neural Inf Process Syst* (2023) 36:34892–916.

67. Jia X, Zhu C, Li M, Tang W, Zhou W. Llvip: a visible-infrared paired dataset for low-light vision. In: *Proceedings of the IEEE/CVF international conference on computer vision workshops (ICCVW)* (2021). p. 3496–504.

68. Tang L, Huang H, Zhang Y, Qi G, Yu Z. Structure-embedded ghosting artifact suppression network for high dynamic range image reconstruction. *Knowledge-Based Syst* (2023) 263:110278. doi:10.1016/j.knosys.2023.110278

69. Liu Y, Shi Y, Mu F, Cheng J, Chen X. Glioma segmentation-oriented multi-modal mr image fusion with adversarial learning. *IEEE/CAA J Automatica Sinica* (2022) 9:1528–31. doi:10.1109/jas.2022.105770

70. Xie M, Wang J, Zhang Y. A unified framework for damaged image fusion and completion based on low-rank and sparse decomposition. *Signal Processing: Image Commun* (2021) 29:116400. doi:10.1016/j.image.2021.116400

71. Liu J, Fan X, Huang Z, Wu G, Liu R, Zhong W, et al. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (2022). p. 5802–11.

72. Liu Y, Qi Z, Cheng J, Chen X. Rethinking the effectiveness of objective evaluation metrics in multi-focus image fusion: a statistic-based approach. *IEEE Trans Pattern Anal Machine Intelligence* (2024) 46:5806–19. doi:10.1109/tpami.2024.3367905

73. Xydeas CS, Petrovic V, et al. Objective image fusion performance measure. *Electronics Lett* (2000) 36:308–9.

74. Chen H, Varshney PK. A human perception inspired quality metric for image fusion based on regional information. *Inf Fusion* (2007) 8:193–207. doi:10.1016/j.inffus.2005.10.001

75. Wang Z, Bovik A, Sheikh H, Simoncelli E. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* (2004) 13:600–12. doi:10.1109/tip.2003.819861

76. Liu J, Wu G, Liu Z, Wang D, Jiang Z, Ma L, et al. Infrared and visible image fusion: from data compatibility to task adaption. *IEEE Trans Pattern Anal Machine Intelligence* (2024) 1–20. doi:10.1109/TPAMI.2024.3521416

77. Zhang X, Ye P, Xiao G. Vifb: a visible and infrared image fusion benchmark. In: *2020 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)* (2020). p. 468–78.

78. Zhao Z, Xu S, Zhang J, Liang C, Zhang C, Liu J. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Trans Circuits Syst Video Technology* (2021) 32:1186–96. doi:10.1109/TCSVT.2021.3075745

79. Li H, Liu J, Zhang Y, Liu Y. A deep learning framework for infrared and visible image fusion without strict registration. *Int J Computer Vis* (2024) 132:1625–44. doi:10.1007/s11263-023-01948-x

80. Li H, Xu T, Wu X-J, Lu J, Kittler J. Lrrnet: a novel representation learning guided fusion network for infrared and visible images. *IEEE Trans Pattern Anal Machine Intelligence* (2023) 45:11040–52. doi:10.1109/tpami.2023.3268209

81. Li H, Cen Y, Liu Y, Chen X, Yu Z. Different input resolutions and arbitrary output resolution: a meta learning-based deep framework for infrared and visible image fusion. *IEEE Trans Image Process* (2021) 30:4070–83. doi:10.1109/tip.2021.3069339

82. Liu R, Liu Z, Liu J, Fan X, Luo Z. A task-guided, implicitly-searched and meta-initialized deep model for image fusion. *IEEE Trans Pattern Anal Machine Intelligence* (2024) 46:6594–609. doi:10.1109/tpami.2024.3382308

83. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans Med Imaging* (2015) 34:1993–2024. doi:10.1109/tmi.2014.2377694

84. Mu P, Wu G, Liu J, Zhang Y, Fan X, Liu R. Learning to search a lightweight generalized network for medical image fusion. *IEEE Trans Circuits Syst Video Technology* (2024) 34:5921–34. doi:10.1109/tcsvt.2023.3342808

85. Zhao Z, Bai H, Zhang J, Zhang Y, Zhang K, Xu S, et al. Equivariant multi-modality image fusion. In: *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (2024). p. 25912–21.

86. Zhang H, Zuo X, Zhou H, Lu T, Ma J. A robust mutual-reinforcing framework for 3d multi-modal medical image fusion based on visual-semantic consistency. *Proc AAAI Conf Artif Intelligence* (2024) 38:7087–95. doi:10.1609/aaai.v38i7.28536

87. Zhang Y, Yang X, Li H, Xie M, Yu Z. Dcpnet: a dual-task collaborative promotion network for pansharpening. *IEEE Trans Geosci Remote Sensing* (2024) 62:1–16. doi:10.1109/tgrs.2024.3377635

88. Li H, Yang Z, Zhang Y, Tao D, Yu Z. Single-image hdr reconstruction assisted ghost suppression and detail preservation network for multi-exposure hdr imaging. *IEEE Trans Comput Imaging* (2024) 10:429–45. doi:10.1109/tci.2024.3369396

89. Li H, Wang D, Huang Y, Zhang Y, Yu Z. Generation and recombination for multifocus image fusion with free number of inputs. *IEEE Trans Circuits Syst Video Technology* (2024) 34:6009–23. doi:10.1109/TCSVT.2023.3344222

90. Liu Y, Yu C, Cheng J, Wang ZJ, Chen X. Mm-net: a mixformer-based multi-scale network for anatomical and functional image fusion. *IEEE Trans Image Process* (2024) 33:2197–212. doi:10.1109/tip.2024.3374072