frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 06 June 2025
DOI 10.3389/fphy.2025.1599968

OPEN ACCESS

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

REVIEWED BY
Fan Li,
Kunming University of Science and Technology, China
Haicheng Bai,
Yunnan Normal University, China

*CORRESPONDENCE
Zheng Liu,
490956823@qq.com

RECEIVED 25 March 2025
ACCEPTED 19 May 2025
PUBLISHED 06 June 2025

CITATION
Hu D, Wang K, Zhang C, Liu Z, Che Y, Dong S and Kong C (2025) Target-aware unregistered infrared and visible image fusion.
Front. Phys. 13:1599968.
doi: 10.3389/fphy.2025.1599968

Target-aware unregistered infrared and visible image fusion

Dengshu Hu, Ke Wang, Cuijin Zhang, Zheng Liu*, Yukui Che, Shoubing Dong and Chuirui Kong

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

Introduction: Infrared (IR) and visible (VI) image fusion can provide richer texture details for subsequent object detection tasks. Conversely, object detection can offer semantic information about targets, which in turn helps improve the quality of the fused images. As a result, joint learning approaches that integrate infrared-visible image fusion and object detection have attracted increasing attention.

Methods: However, existing methods typically assume that the input source images are perfectly aligned spatially—an assumption that does not hold in real-world applications. To address this issue, we propose a novel method that enables mutual enhancement between infrared-visible image fusion and object detection, specifically designed to handle misaligned source images. The core idea is to use the object detection loss, propagated via backpropagation, to guide the training of the fusion network, while a specially designed loss function mitigates the modality gap between infrared and visible images.

Results: Comprehensive experiments on three public datasets demonstrate the effectiveness of our approach.

Discussion: In addition, our approach can be used with other radiation frequencies where different modalities require image fusion like, for example, radio-frequency, x- and gamma rays used in medical imaging.

KEYWORDS

infrared and visible image fusion, object detection, feature alignment, target-aware, unregistered

1 Introduction

Images captured by a single sensor often fail to provide a comprehensive description of a scene. For example, infrared (IR) sensors can capture thermal radiation emitted by objects and highlight salient targets, but they lack the ability to represent fine texture details and are more susceptible to noise. On the other hand, visible-light (VI) sensors capture visual information with clear texture details but are easily affected by lighting conditions and occlusions. If the information from both infrared and visible images can be integrated into a single, information-rich fused image, the scene representation can be significantly enhanced. As a result, infrared and visible image fusion has been widely applied as a low-level preprocessing task in various high-level vision applications, such as object detection [1], tracking [2], person re-identification [3], and semantic segmentation [4]. An example in Figure 1 visually illustrates the application of fused images in object detection. It can be observed that detection results obtained from individual sensor images are less accurate than those derived from fused images.

Due to its practical value, infrared and visible image fusion has garnered substantial attention in the research community. Over the past decades, numerous image fusion

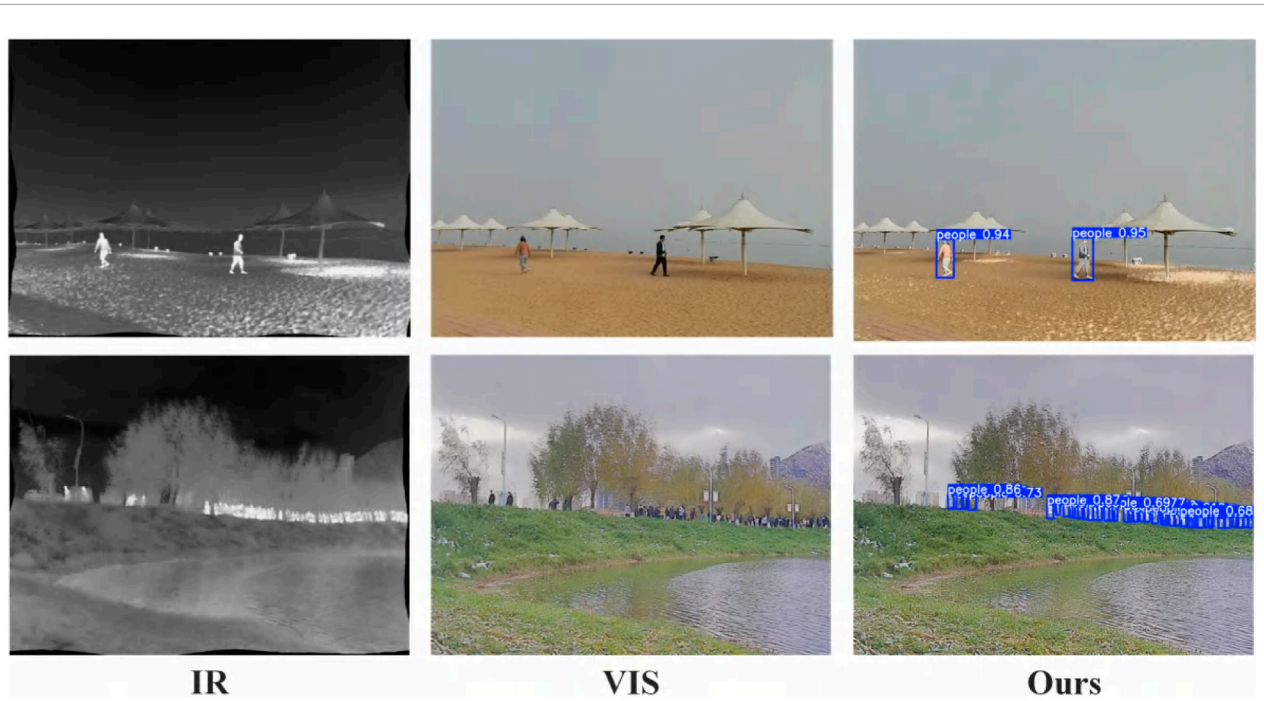

FIGURE 1
Object detection results of the proposed method on the M^3FD dataset.

techniques have been proposed, including both traditional and deep learning-based methods. Traditional methods typically fall into two categories: multi-scale transform-based methods [5–7] and sparse representation-based methods [8–12]. Deep learning-based approaches include methods based on autoencoders (AE) [9, 13, 14], convolutional neural networks (CNNs) [15–18], and generative adversarial networks (GANs) [19, 20].

Although recent deep learning-based fusion algorithms can generate visually pleasing results, several critical challenges remain unsolved. On one hand, most existing fusion algorithms focus on optimizing visual quality and evaluation metrics, but rarely consider whether the fused results benefit downstream task performance. On the other hand, even recent methods that incorporate high-level vision tasks into the fusion process—such as TarDAL [21], which proposes a dual-level optimization model using a task-aware dual adversarial learning network to simultaneously address fusion and object detection; SeAFusion [22], which constrains the fusion process with semantic loss to retain richer semantic information; and DetFusion [23], which guides multimodal fusion using target-related features learned by the object detection network—still assume that the source images are perfectly aligned spatially. This assumption does not hold in real-world applications.

In this study, we propose a framework named Target-Aware Unregistered Infrared and Visible Image Fusion Network, designed to achieve robust performance in both misaligned image fusion and high-level vision tasks. Specifically, we introduce an object detection network to predict detection results on the fused image and construct a detection loss. This loss is then backpropagated to guide the training of the fusion network, encouraging the fused image to retain more information useful for object detection. Additionally, to effectively align unregistered images, we design a modality consistency loss to reduce the domain gap between infrared and visible images.

In summary, our main contributions are as follows:

(1) We are the first to unify unregistered image fusion and object detection within a single framework, breaking the limitations of object detection in real-world applications.
(2) We propose a modality consistency loss that effectively eliminates the domain discrepancy between infrared and visible images, improving image registration accuracy.
(3) Our method demonstrates excellent performance in image alignment, fusion, and object detection across multiple datasets. And our method can be used with other radiation frequencies where different modalities require image fusion like, for example, radio-frequency, x- and gamma rays used in medical imaging.

The rest of this paper is organized as follows. Section 2 briefly reviews related work on high-level vision task-driven image fusion and unregistered infrared-visible image fusion. Section 3 describes the proposed method in detail. Section 4 presents and discusses the experimental results. Section 5 concludes the paper.

2 Related work

In this section, we first provide a brief overview of high-level vision task-driven infrared and visible image fusion methods, and then review existing approaches for unregistered infrared and visible image fusion.

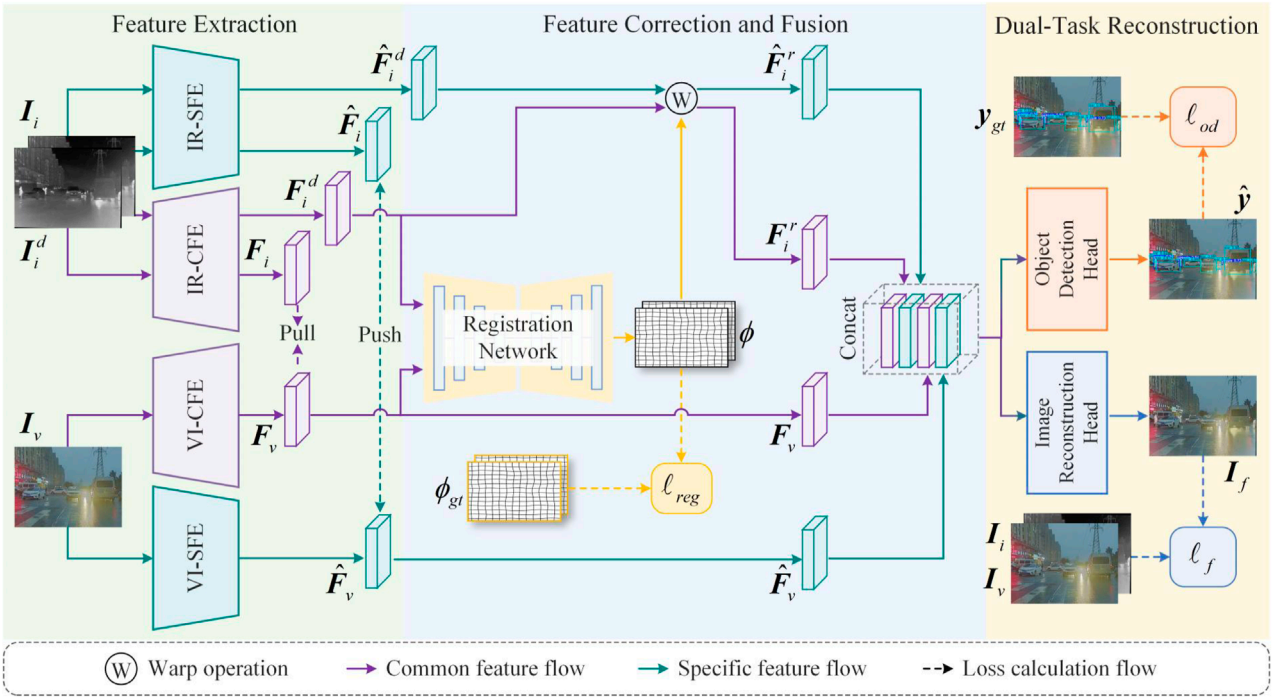

FIGURE 2
Overall framework of the proposed method. We use IR/VI-CFE and IR/VI-SFE to extract common and specific features from the infrared and visible images, respectively. To obtain the deformation field for spatial correction, the infrared/visible common features are fed into the registration module to predict the deformation field. This deformation field is then applied to the infrared common/specific features to correct spatial deformation. The corrected infrared features are concatenated with the visible features and then fed into the image reconstruction head and the object detection head, respectively, to generate the fused image and the object detection result.

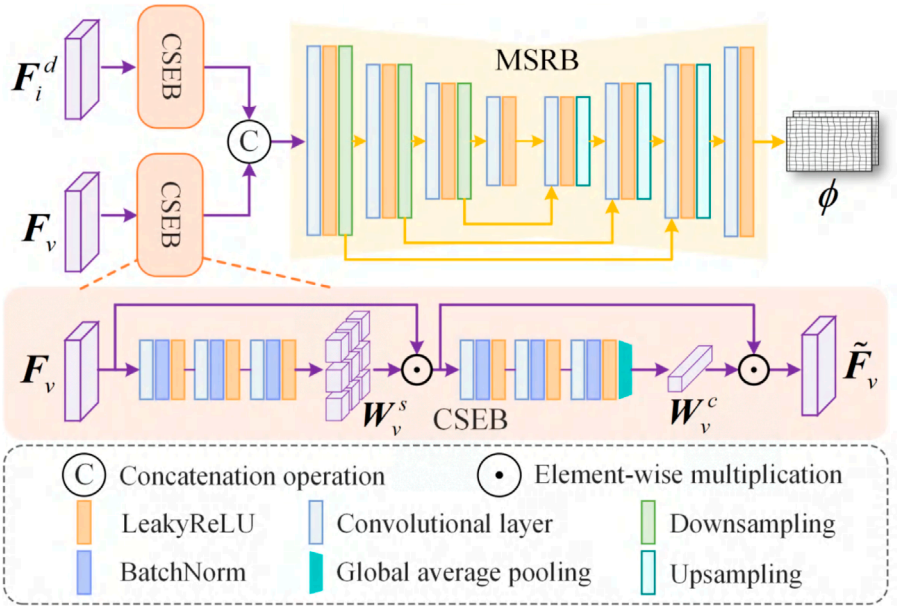

FIGURE 3
Structure of the registration network. The registration network mainly consists of the Channel and Spatial Enhancement Block (CSEB) and the Multi-Scale Registration Block (MSRB).

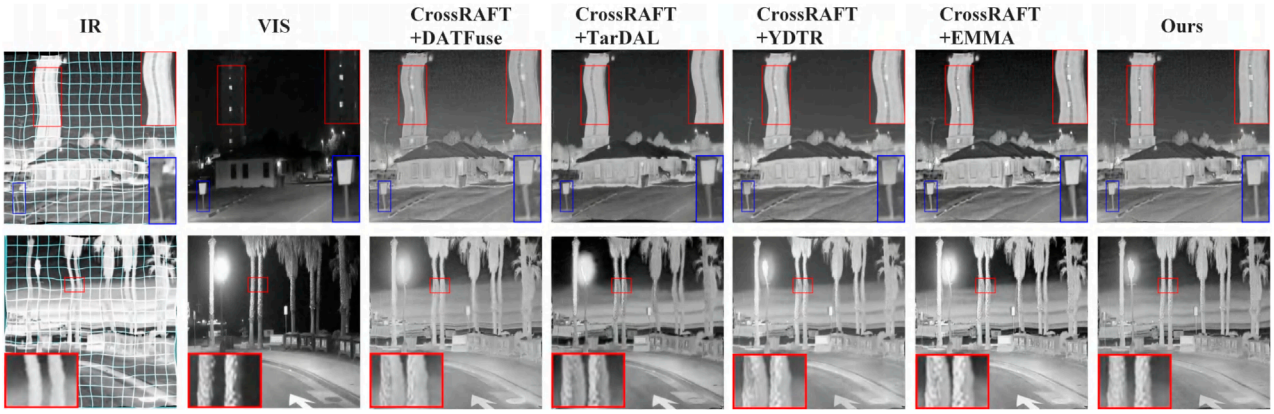

FIGURE 4
Qualitative comparison of fusion results between the Registration + Fusion + Object Detection methods and the proposed method. The first two columns show the unaligned source images as input. The grid in the first column illustrates the deformation present in the image. Columns 3 to 7 present the fusion results obtained by different methods.

TABLE 1 Quantitative comparison of fusion results between the Registration + Fusion + Object Detection methods and the proposed method.

Methods	$Q_{CC}\uparrow$	$Q_{AB/F}\uparrow$	$Q_{CV}\downarrow$	$Q_{SSIM}\uparrow$
DATFuse	0.8303	0.3246	1425.2631	1.2189
TarDAL	0.8317	0.3313	1396.1484	1.2205
YDTR	0.8246	0.3179	1383.2556	1.2133
EMMA	0.8255	0.3341	1399.4075	1.2236
Ours	**0.8325**	**0.3420**	**1375.5238**	**1.2271**

Bolded values indicate the best performance.

2.1 High-level vision task-driven infrared and visible image fusion

High-level vision task-driven fusion methods typically incorporate a semantic segmentation [24–27] or object detection network [23, 28] after the fusion network, using the loss functions from these downstream tasks to constrain the fusion results and improve the quality of the fused image. However, introducing high-level vision tasks at the fused image level only provides indirect guidance for the feature extraction network to learn features relevant to the downstream tasks.

To provide direct task-level guidance at the feature level and further enhance fusion performance, PSFusion [29] injects semantic features extracted from a segmentation task directly into the fusion network. SegMiF [25] feeds the fused result into a semantic segmentation network to extract semantic features, which are then interacted with the multimodal image features from the encoder to enhance the fusion result. MRFS [26] interacts and fuses the source image features before feeding them into a semantic segmentation head to enforce semantic supervision, thereby improving the global scene perception of the fusion network. MetaFusion [28] sends the fused result into an object detection network to extract features, which are then combined with the source image features and passed into a meta-feature generator to guide feature extraction in the fusion branch.

Although these methods improve fusion performance to some extent by leveraging downstream high-level tasks, they all assume that the input images are perfectly aligned in spatial position—a condition rarely met in real-world applications. In practice, such methods rely on additional image registration algorithms to achieve accurate alignment before performing fusion. This not only makes the fusion quality highly dependent on the registration accuracy but also significantly increases the complexity of the overall network design.

2.2 Unregistered infrared and visible image fusion

To address the problem of unregistered infrared and visible image fusion, most existing approaches combine registration and fusion algorithms, i.e., first aligning the input misaligned image pairs and then performing fusion. However, due to the large modality gap between infrared and visible images, ignoring the adverse

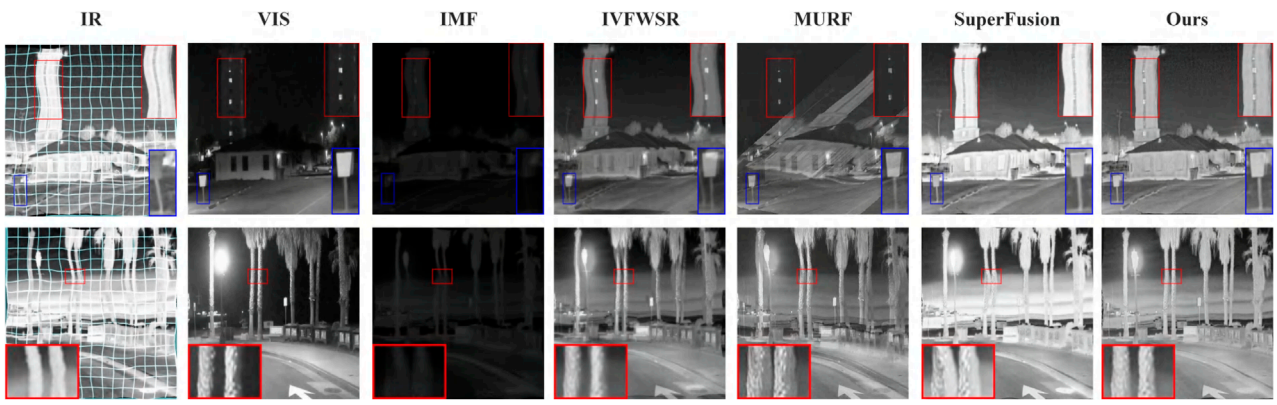

FIGURE 5
Qualitative comparison of fusion results between the Joint Registration and Fusion + Object Detection methods and the proposed method. The first two columns show the unaligned source images as input. The grid in the first column illustrates the deformation in the image. Columns 3 to 7 display the fusion results produced by different methods.

TABLE 2 Quantitative comparison of fusion results between the Joint Registration and Fusion + Object Detection methods and the proposed method.

Methods	$Q_{CC}\uparrow$	$Q_{AB/F}\uparrow$	$Q_{CV}\downarrow$	$Q_{SSIM}\uparrow$
IMF	0.8221	0.3119	1477.6932	1.2058
IVFWSR	0.8269	0.3208	1586.8251	1.2115
MURF	0.8315	0.3254	1456.3259	1.2140
SuperFusion	0.8320	0.3396	1399.4521	1.2207
Ours	**0.8325**	**0.3420**	**1375.5238**	**1.2271**

Bolded values indicate the best performance.

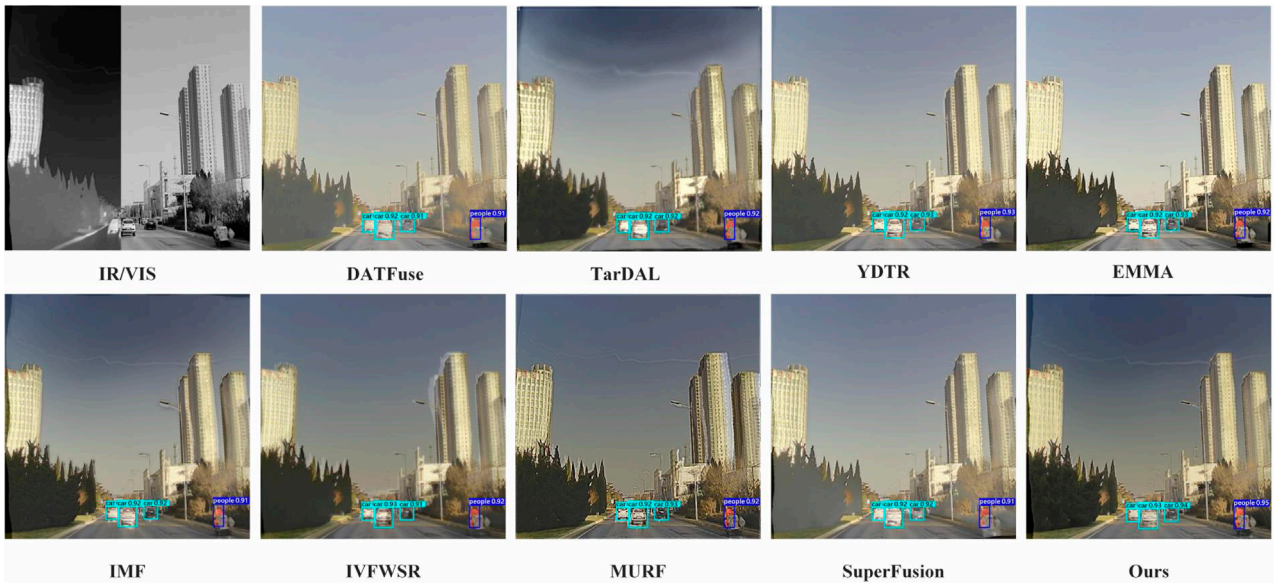

FIGURE 6
Visualization of object detection results using different fusion methods on the M^3FD dataset.

TABLE 3 Quantitative object detection results of different fusion methods on the M^3FD dataset.

Methods	$mAP_{50\rightarrow90}\uparrow$
DATFuse	53.10
TarDAL	53.20
YDTR	53.80
EMMA	54.20
IMF	52.40
IVFWSR	52.60
MURF	52.20
SuperFusion	53.80
Ours	**54.50**

Bolded values indicate the best performance.

impact of modality discrepancy on registration can greatly degrade fusion quality. For instance, ReCoNet [30] adopts this strategy but produces suboptimal fusion results due to this issue. UMF-CMGR [31] and IMF [32] consider the effect of modality differences on registration results. They propose to convert visible images into pseudo-infrared images via an image generation network and then perform mono-modal registration between the pseudo-infrared and misaligned infrared images. However, the quality of the generated image has a direct impact on the final performance of these methods. Moreover, these methods treat registration and fusion as two independent tasks, failing to establish a unified framework where both tasks can benefit each other.

To address this, RFNet [33] and MURF [34] treat image fusion as a downstream task of registration and improve registration performance by enhancing the sparsity of the gradient in the fused result. However, to tackle the modality discrepancy issue during registration, both methods aim to transform the multimodal registration into a mono-modal one. Specifically, RFNet uses an image generation model to produce a pseudo-image with the same modality as the misaligned one before performing mono-modal registration, while MURF leverages contrastive learning to extract modality-invariant features from the input image pair for registration. Similarly, Super-Fusion [35] extracts modality-invariant features using shared-parameter encoders and consistency constraints on the fused result for registration.

Nevertheless, the information carried by modality-invariant features in infrared-visible pairs is often far less rich than the complementary information present in the image pair. As a result, it is difficult to achieve satisfactory cross-modal registration using only modality-invariant features. In addition, the above methods all follow a two-stage approach (registration + fusion). This two-stage strategy greatly limits deployment in practical applications due to computational constraints. Although RFVIF [36], IVFWSR [37] and MulFS-CAP [38] attempt to achieve registration and fusion within a single-stage framework, the types of deformations they can handle remain limited. Unlike the methods mentioned above, our approach considers multiple challenges simultaneously: the impact of modality discrepancy on cross-modal registration, the deployment limitations of two-stage processing, and the feature requirements of downstream high-level vision tasks for both registration and fusion.

3 Methods

3.1 Overview

As shown in Figure 2, the proposed method consists of three core components: feature extraction, feature alignment and fusion, and dual-task reconstruction. The feature extraction component is designed to obtain both modality-specific and modality-common features from the source images. The feature alignment and fusion component is used to predict a deformation field, which is then used to spatially align the infrared-specific and common features. These aligned features are then fused with the corresponding visible image's specific and common features. In the dual-task reconstruction stage, the fused features are fed into the object detection head and the image reconstruction head, respectively, to generate both the object detection result map and the fused image.

3.2 Feature extraction

The main objective of feature extraction is to extract both the common and specific features of infrared and visible images, in order to facilitate subsequent cross-modal registration and feature fusion. This process consists of four modules: the IR-Specific Feature Extraction (IR-SFE) module, the VI-Specific Feature Extraction (VI-SFE) module, the IR-Common Feature Extraction (IR-CFE) module, and the VI-Common Feature Extraction (VI-CFE) module. Among them, the IR/VI-SFE modules are used to extract modality-specific features from the infrared/visible images, while the IR/VI-CFE modules are used to extract their common features. Assume that each sample in the training dataset contains three images: a pixel-wise strictly aligned infrared image $\boldsymbol{I}_i$, a visible image $\boldsymbol{I}_v$, and a deformed infrared image $\boldsymbol{I}_i^d$. We feed $\boldsymbol{I}_i$ and $\boldsymbol{I}_i^d$ into the IR-CFE and IR-SFE, respectively, to obtain the infrared common feature $\boldsymbol{F}_i$, the deformed infrared common feature $\boldsymbol{F}_i^d$, the infrared specific feature $\widehat{\boldsymbol{F}}_i$, and the deformed infrared specific feature $\widehat{\boldsymbol{F}}_i^d$. At the same time, we feed $\boldsymbol{I}_v$ into the VI-CFE and VI-SFE to obtain the visible common feature $\boldsymbol{F}_v$ and the visible specific feature $\widehat{\boldsymbol{F}}_v$.

In the cross-modal registration process, it is usually necessary to rely on the common information between cross-modal images to establish pixel-wise correspondences. To reduce the modality gap between infrared and visible images and thus establish more accurate pixel-wise correspondences, we introduce a modality consistency loss ℓ_c:

$$\ell_c = \frac{1}{HWC}\left\|\boldsymbol{F}_i - \boldsymbol{F}_v\right\|_1, \tag{1}$$

Here, H, W, and C denote the height, width, and number of channels of the feature maps, respectively, and $\|\cdot\|_1$ represents the l1-norm. In addition, considering that the goal of image fusion is to integrate as much complementary information as possible

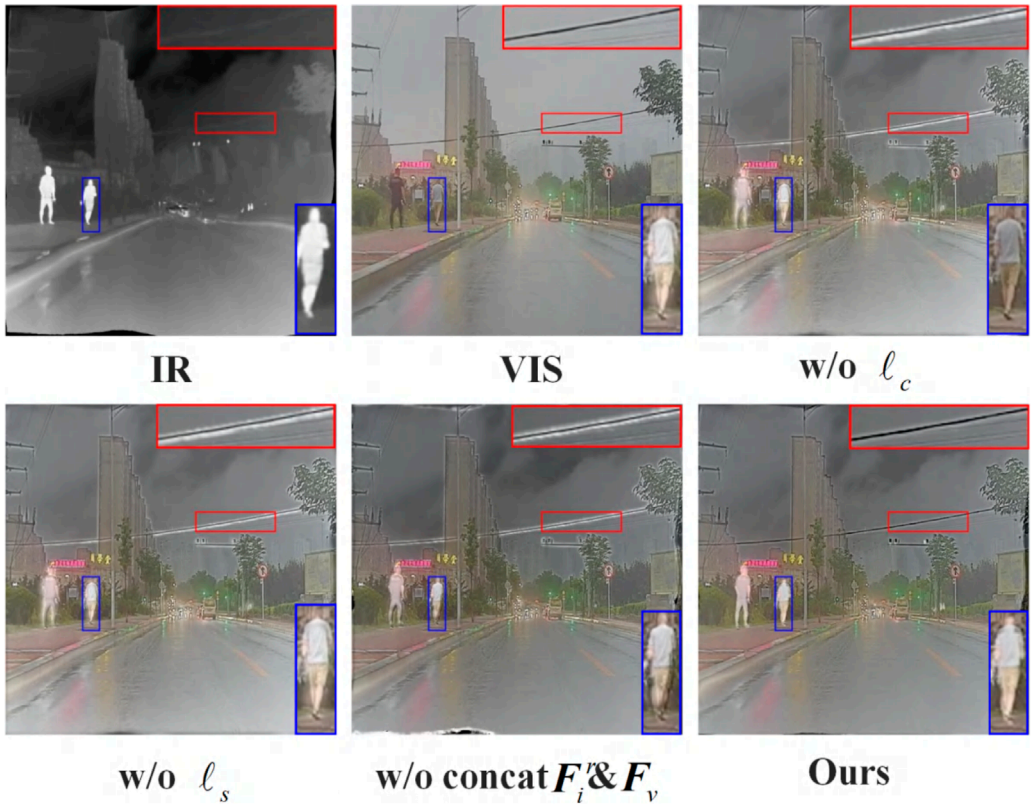

FIGURE 7
Ablation study of the core designs.

TABLE 4 Quantitative results of the ablation study on the core designs.

Methods	$Q_{CC}\uparrow$	$Q_{AB/F}\uparrow$	$Q_{CV}\downarrow$	$Q_{SSIM}\uparrow$
w/o ℓ_c	0.8304	0.3469	1339.9062	1.2204
w/o ℓ_s	0.8313	0.3439	1336.8814	1.2256
w/o Concat F_i^r and F_v	0.8274	0.3451	1369.4537	1.2114
Ours	**0.8325**	**0.3420**	**1375.5238**	**1.2271**

Bolded values indicate the best performance.

from cross-modal source images into a single image, we introduce the modality complementary information loss ℓ_s to further enrich the complementary information from the source images in the fused image:

$$\ell_s = -\frac{1}{HWC}\left\|\hat{F}_i - \hat{F}_v\right\|_1. \quad (2)$$

3.3 Feature alignment and fusion

Feature alignment corrects the deformation in infrared features by predicting a deformation field, thereby achieving spatial alignment between infrared and visible features. This process is mainly implemented by the registration network. Subsequently, the aligned infrared features are fused with the visible features to obtain the fused features. As shown in Figure 3, the registration network is composed of a Channel and Spatial Enhancement Block (CSEB) and a Multi-Scale Registration Block (MSRB). The CSEB is mainly used to enhance the information beneficial to registration at both the channel and spatial levels, thereby improving the accuracy of the predicted deformation field. The CSEB consists of six feature extraction layers and a Global Average Pooling (GAP) layer. Each feature extraction layer is composed of a convolutional layer with a kernel size of 3×3, stride 1, followed by Batch Normalization (BatchNorm) and a LeakyReLU activation function. The MSRB is used to predict the deformation field to correct the deformed infrared features and ensure spatial alignment between the infrared and visible features. The MSRB adopts a U-Net-like architecture.

TABLE 5 Quantitative analysis results of the hyperparameter study.

γ	λ_1	λ_2	λ_3	$Q_{CC}\uparrow$	$Q_{AB/F}\uparrow$	$Q_{CV}\downarrow$	$Q_{SSIM}\uparrow$
2	10	5	1	0.8235	0.3352	1450.3498	1.2222
2	10	5	10	0.8198	0.3389	1683.8772	1.2195
2	10	1	5	0.8123	0.3321	1502.6641	1.2088
2	10	10	5	0.8260	0.3334	1465.2293	1.2247
2	1	5	5	0.8011	0.3195	1450.3288	1.1954
2	20	5	5	0.8059	0.3248	1529.1245	1.1996
1	10	5	5	0.8144	0.3340	1499.3888	1.2111
5	10	5	5	0.8080	0.3302	1775.1124	1.2020
2	10	5	5	**0.8325**	**0.3420**	**1375.5238**	**1.2271**

Bolded values indicate the best performance.

TABLE 6 Computational efficiency comparison of four SOTA Joint Registration and Fusion methods, the value is tested on GPU.

Methods	FLOPs(G)	Size(M)	Time(s)
IMF	1724.08	13.30	0.82
IVFWSR	859.43	14.09	0.33
MURF	120.72	1.76	1.18
SuperFusion	65.43	**0.14**	**0.27**
Ours	**60.12**	0.97	0.40

Bolded values indicate the best performance.

We input the deformed infrared common feature $\boldsymbol{F}_i^d$ and the visible common feature $\boldsymbol{F}_v$ into two CSEBs with unshared parameters, obtaining the enhanced features $\tilde{\boldsymbol{F}}_i^d$ and $\tilde{\boldsymbol{F}}_v$, respectively. Taking the enhancement process of $\boldsymbol{F}_v$ as an example, $\boldsymbol{F}_v$ is fed into three feature extraction layers to generate the spatial enhancement weights $\boldsymbol{W}_v^s$. To enhance registration-relevant information at the spatial level, we perform element-wise multiplication between $\boldsymbol{W}_v^s$ and $\boldsymbol{F}_v$:

$$\boldsymbol{F}_v^s = \boldsymbol{W}_v^s \odot \boldsymbol{F}_v, \tag{3}$$

Here, $\boldsymbol{F}_v^s$ denotes the feature enhanced at the spatial level, and $\odot$ represents the element-wise multiplication operation. We feed $\boldsymbol{F}_v^s$ into three feature extraction layers and a global average pooling (GAP) layer to obtain feature $\boldsymbol{W}_v^c$ for channel-level enhancement. Then, $\boldsymbol{W}_v^c$ is element-wise multiplied with $\boldsymbol{F}_v^s$ to produce the enhanced feature $\tilde{\boldsymbol{F}}_v$, which has been refined at both the spatial and channel levels:

$$\tilde{\mathbf{F}}_v = \mathbf{W}_v^c \odot \mathbf{F}_v^s, \tag{4}$$

Similarly, we obtain the deformed infrared common feature $\tilde{\boldsymbol{F}}_i^d$ enhanced at both the spatial and channel levels. We concatenate $\tilde{\boldsymbol{F}}_i^d$and $\tilde{\boldsymbol{F}}_v$ along the channel dimension and feed the resulting feature into the MSRB to predict the deformation field ϕ. To ensure the accuracy of the predicted deformation field, we introduce a registration loss ℓ_{reg}:

$$\ell_{reg} = \frac{1}{2HW}\left\|\phi - \phi_{gt}\right\|_1, \tag{5}$$

Here, ϕ_{gt} is the label of ϕ.

We use ϕ to correct $\boldsymbol{F}_i^d$ and $\widehat{\boldsymbol{F}}_i^d$ respectively, resulting in the corrected infrared common feature $\boldsymbol{F}_i^r$ and infrared-specific feature $\widehat{\boldsymbol{F}}_i^r$:

$$\begin{aligned} \boldsymbol{F}_i^r &= \phi \circ \boldsymbol{F}_i^d, \\ \widehat{\boldsymbol{F}}_i^r &= \phi \circ \widehat{\boldsymbol{F}}_i^d, \end{aligned} \tag{6}$$

Here, $\circ$ denotes the Warp operation, which resamples the deformed feature maps based on ϕ to correct the deformations within them. During the fusion process, to minimize information loss, we concatenate $\boldsymbol{F}_i^r$, $\widehat{\boldsymbol{F}}_i^r$, $\boldsymbol{F}_v$, and $\widehat{\boldsymbol{F}}_v$ along the channel dimension to obtain the fused feature $\boldsymbol{F}_f$:

$$\boldsymbol{F}_f = \left[\boldsymbol{F}_i^r, \widehat{\boldsymbol{F}}_i^r, \boldsymbol{F}_v, \widehat{\boldsymbol{F}}_v\right], \tag{7}$$

Here, $[\cdot]$ represents the operation of concatenation along the channel dimension.

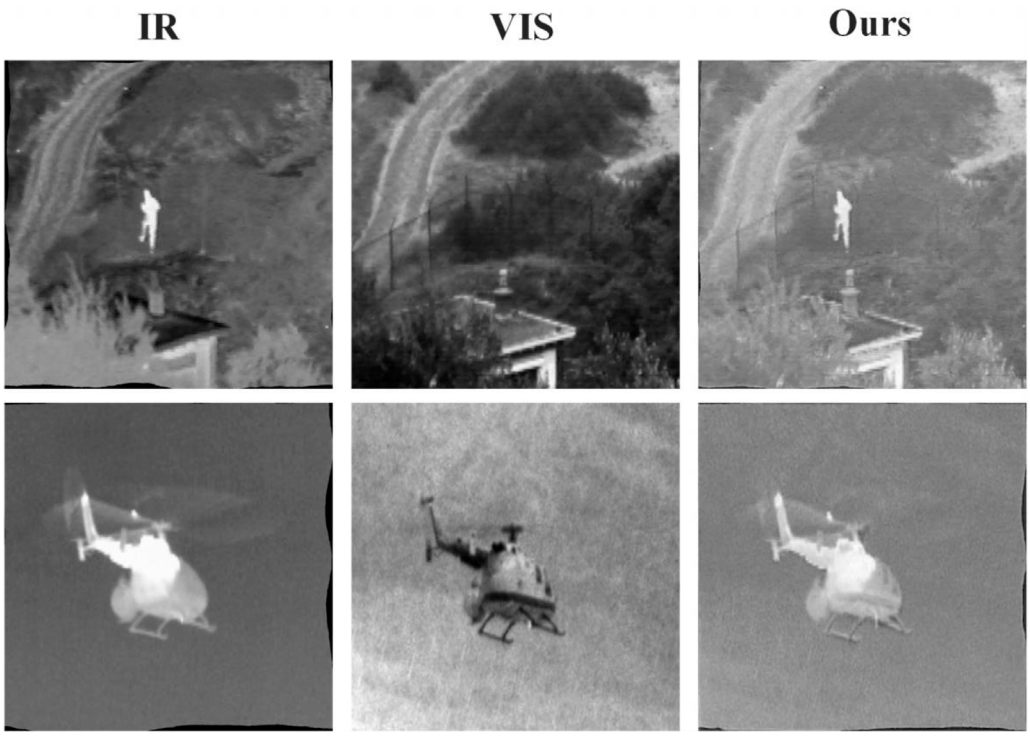

FIGURE 8
Fusion results of our method on different scenarios.

FIGURE 9
Failure cases of our method on the real-world dataset CVC-14.

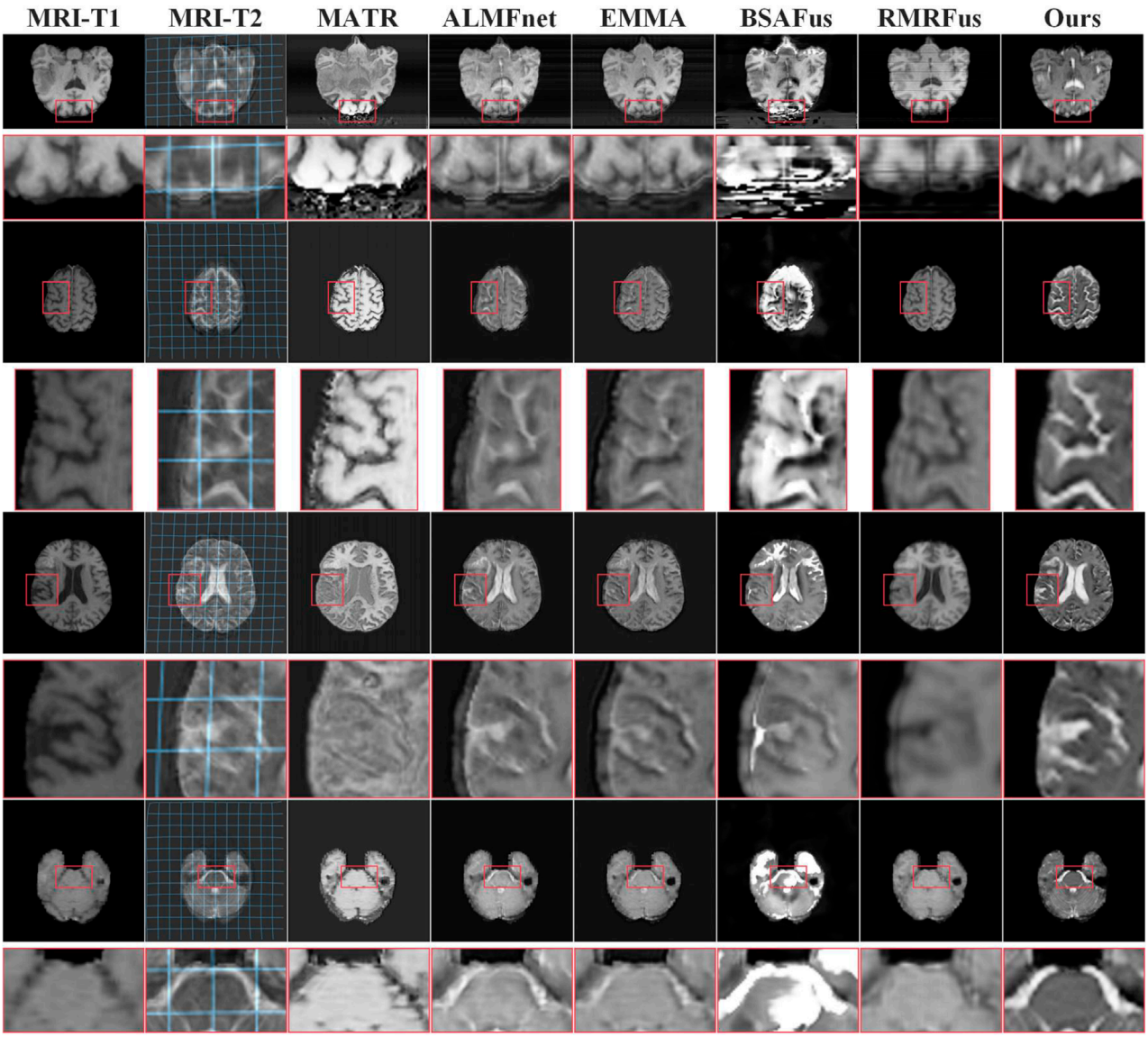

FIGURE 10
Visual comparison on the BraTS2020 dataset.

TABLE 7 Quantitative analysis results on the BraTS2020 dataset.

Methods	$Q_{CC}\uparrow$	$Q_{AB/F}\uparrow$	$Q_{CV}\downarrow$	$Q_{SSIM}\uparrow$
MATR	0.7889	0.2901	1345.4510	1.2299
ALMFnet	0.7749	0.2888	1606.5911	1.2155
EMMA	0.7906	0.2853	1568.7139	1.2220
BSAFus	0.7812	0.3001	1436.1287	1.2318
RMRFus	0.7784	0.2992	1409.9831	1.2007
Ours	**0.7934**	**0.3063**	**1399.5234**	**1.2454**

Bolded values indicate the best performance.

3.4 Dual-task reconstruction

In the dual-task reconstruction, the fused feature is fed into both the object detection head and the image reconstruction head to respectively generate the object detection result map and the fused image. The dual-task reconstruction primarily consists of the object detection head and the image reconstruction head. We adopt YOLOv5 [39] as the object detection head. The image reconstruction head is composed of three feature extraction layers, where the LeakyReLU activation function in the final layer is replaced with a Tanh activation function. The fused feature $\boldsymbol{F}_f$ is input into both the object detection head and the image reconstruction head to obtain the object detection result map $\hat{\boldsymbol{y}}$ and the fused image $\boldsymbol{I}_f$, respectively. To ensure high-quality object detection results, we introduce the object detection loss ℓ_{ob} to constrain the network:

$$\ell_{ob} = c_{yolov5}\left(\boldsymbol{y}, \boldsymbol{y}_{gt}\right), \tag{8}$$

Here, $c_{yolov5}(\cdot)$ refers to the loss function used during the training of YOLOv5. In addition, to encourage the fused image to retain as much shared and complementary information from both infrared and visible images as possible, we introduce luminance loss ℓ_b and gradient loss ℓ_g, and construct the fusion loss ℓ_f accordingly:

$$\ell_f = \ell_b + \gamma \ell_g, \tag{9}$$

Here, γ denotes the balancing hyperparameter. The gradient loss ℓ_g is defined as:

$$\ell_g = \frac{1}{HW}\left\|\nabla \boldsymbol{I}_f - \max\left(\nabla \boldsymbol{I}_i, \nabla \boldsymbol{I}_v\right)\right\|_1, \tag{10}$$

Here, ∇ denotes the Sobel operator. The luminance loss ℓ_b is defined as:

$$\ell_b = \frac{1}{HW}\left\|\boldsymbol{I}_f - \max\left(\boldsymbol{I}_i, \boldsymbol{I}_v\right)\right\|_1. \tag{11}$$

Finally, we define the total loss ℓ_t as follows:

$$\ell_t = \ell_c + \ell_s + \lambda_1 \ell_{reg} + \lambda_2 \ell_f + \lambda_3 \ell_{ob}, \tag{12}$$

Here, $\lambda_n (n = 1, 2, 3)$ denotes the balancing hyperparameter.

4 Experiments

4.1 Experimental setup

4.1.1 Datasets and implementation details

4.1.1.1 Datasets

Following standard experimental practices in the image fusion field [40–43], we trained our model on 152 pairs of infrared and visible images with a resolution of 512×512 from the RoadScene Xu et al. [44, 45] dataset. For testing, we used 18 pairs of images from RoadScene and 17 pairs from M^3FD [21]. The misaligned infrared images were generated by randomly applying a combination of rigid and non-rigid deformations to the originally well-aligned infrared images. This type of mixed deformation is applied randomly to the original aligned images in each epoch to augment the training data.

4.1.1.2 Implementation details

The proposed method was implemented using the PyTorch framework and trained on a single NVIDIA GeForce RTX 3090 GPU. The model was trained for 150 epochs with a batch size of 8, a learning rate of 1e-3, and the Adam optimizer was used to update the model parameters. The four hyperparameters in the loss function were set to $\gamma = 2$, $\lambda_1 = 10$, $\lambda_2 = 5$,and $\lambda_3 = 5$.

4.1.2 Evaluation metrics

We selected four commonly used image quality evaluation metrics to objectively assess the quality of the fusion results, including correlation coefficient (Q_{CC}) [46], gradient-based fusion performance ($Q_{AB/F}$) [47], Chen-Varshney metric (Q_{CV}) [48], and structural similarity (Q_{SSIM}) [49]. Metric Q_{CC} evaluates the linear correlation between the fused image and the source images, reflecting their similarity. Metric $Q_{AB/F}$ assesses the amount of edge information transferred from the source images to the fused image. Metric Q_{CV} takes into account both edge information and human visual perception. Metric Q_{SSIM} quantifies information loss and distortion in the fused image by comparing it with the source images. Among these metrics, a lower value of indicates better fusion quality, while higher values of the other metrics indicate better performance. In addition, we adopted metric $mAP_{50\to90}$ [50] as the evaluation metric for the object detection task, where a higher $mAP_{50\to90}$ value indicates better detection performance.

4.2 Comparison with state-of-the-art methods

In our experiments, we first compare the proposed method with two categories of fusion approaches for unaligned infrared and visible images based on their fusion results. We then compare the subsequent object detection results obtained using these two categories of methods. The first category involves registering the images to be fused, followed by image fusion and then object detection. We refer to this category as Registration + Fusion + Object Detection. The second category performs joint training of registration and fusion to directly handle unaligned images, followed by object detection. We refer to this as Joint Registration and Fusion + Object Detection.

4.2.1 Comparison with registration + fusion + object detection methods

For the Registration + Fusion + Object Detection methods, we follow the standard processing pipeline used in prior work. We first adopt the high-performing registration method CrossRAFT [51] to align the images to be fused. Then, we apply four advanced infrared and visible image fusion methods to the aligned results, including DATFuse [52], TarDAL [21], YDTR [53], and EMMA [54]. Figure 4 shows the visual results of different methods. As seen from the fusion results, our proposed method not only demonstrates stronger capability in preserving structures and textures but also effectively avoids distortions and artifacts caused by feature misalignment. In addition, we performed objective evaluations of the results from different methods. As shown in Table 1, our method achieves the best performance across all four evaluation metrics.

4.2.2 Comparison with joint registration and fusion + object detection methods

In recent years, joint registration and fusion methods have attracted significant attention. To demonstrate the superiority of our approach over these methods, we compared its performance with four joint registration and fusion methods: IMF, IVFWSR, MURF, and SuperFusion. Figure 5 presents a qualitative comparison of the fusion results produced by different methods. It can be observed that our method exhibits clear advantages in terms of feature alignment, contrast preservation, and detail retention. In addition, we conducted quantitative experiments to visually compare the performance differences. As shown in Table 2, our method achieves the best performance across all four evaluation metrics.

4.2.3 Performance evaluation on infrared and visible image object detection

We evaluated the object detection performance of the two aforementioned categories of methods, as well as the proposed method, on the M^3FD dataset. Figure 6 shows the visualized results of object detection. In comparison, our proposed method achieves superior performance. Table 3 presents the quantitative results. The fused outputs generated by our method help the detection network achieve the highest object detection accuracy. This further demonstrates the superior fusion capability of our approach for object detection tasks.

4.3 Ablation study

The core of the proposed method lies in the losses designed to eliminate modality differences, namely, losses ℓ_c and ℓ_s. In this section, we conduct ablation studies on these key components to verify their effectiveness. All experiments are conducted on the M^3FD dataset. From the ablation results, it can be observed that removing losses ℓ_c and ℓ_s leads to a decline in the model's ability to correct local deformations, as shown in Figure 7. In addition, when the shared information is excluded during fusion and only complementary information is used for concatenation, the visual quality of the fused image does not deteriorate significantly, but the objective evaluation results in Table 4 show a noticeable drop in performance.

4.4 Analysis of hyperparameters

In our proposed method, four main hyperparameters are defined: $\lambda_1, \lambda_2, \lambda_3$, which balances different losses, i.e., ℓ_{reg}, ℓ_f, and ℓ_{ob}, and γ, which balances luminance loss ℓ_b and gradient loss ℓ_g. During model training, λ_1, λ_2, λ_3, γ are set to 10, 5, 5, two respectively.

Next, we analyze the impact of variations in these hyperparameters on model performance. To analyze the impact of $\lambda_1, \lambda_2, \lambda_3$ on fusion performance, we perform a search over λ_1, λ_2, λ_3 values in the ranges of 1–20, 1 to 10, and 1 to 10. The quantitative evaluation results for both fusion and downstream object detection are presented in Table 5. As shown in Table 5, the model achieves optimal performance on fusion when $\lambda_1 = 10$, $\lambda_2 = 5$, and $\lambda_3 = 5$.

To verify the effectiveness of the hyperparameter γ, we fix λ_1, λ_2 and λ_3 to 10, 5, 5 and analyze the model performance as γ varies from 1 to 5. As shown in Table 5, the model achieves the best fusion performance when γ is set to 2. Therefore, we set the hyperparameter γ to 2.

4.5 Analysis of computational complexity

As shown in Table 6, a complexity evaluation is introduced to evaluate the efficiency of our method from three aspects, i.e., FLOPs, training parameters and runtime. Wherein, for FLOPs calculation, the size of the input images is standardized to 512×512 pixels. The inference time is calculated as the average time taken to process 18 scene images from RoadScene's test dataset. From Table 6, our model performs the best in FLOPs, implying that our method has fast calculation speed and is application-friendly. The average inference time for our model to fuse two source images is 0.40 s, only a bit longer than the SOTA method, demonstrating that our model's inference speed is relatively fast and acceptable. Besides, the parameter size of our model is only 0.97M, which can be easily deployed in practical applications. This indicates the efficiency of our method, which can serve practical vision tasks well with better visual performance.

4.6 Analysis of generalization ability

To validate the generalization ability of our method, we conduct experiments under other scenarios. Fusion results are shown in Figure 8. From the qualitative results we can see that our proposed model performs perfectly under other scenarios.

4.7 Analysis of limitation

The proposed method enables mutual enhancement between infrared-visible image fusion and object detection, specifically designed to handle misaligned source images, achieving better experimental results compared to other methods. However, our approach still has certain limitations. Specifically, since our model is trained on the generated unaligned dataset, where the deformations in real-world images cannot be fully included, failure cases appear under real-world scenarios. As shown in Figure 9, our method fails to handle deformations under real-world scenarios. Improving the robustness of our method is vital for future research.

4.8 Further discussion

To validate the effectiveness of the proposed method in the field of medical imaging, we conduct a comparative study on the publicly available BraTS2020 Menze et al. [55] dataset. Specifically, we first employ the state-of-the-art medical image registration method CorrMLP Meng et al. [56] to align the deformed MRI-T2 images to the reference MRI-T1 images, and subsequently apply several advanced fusion methods (including MATR Tang et al. [57], ALMFnet Mu et al. [58], EMMA Zhao et al. [54], BSAFus Li et al. [47], and RMRFus Zhang et al. [59]) for image fusion. As shown in Figure 10, the fusion images generated by the

proposed method exhibit superior image quality and effectively correct artifacts and spatial deformations. In contrast, existing "registration + fusion" methods often introduce noticeable artifacts when handling unregistered medical images, significantly degrading the visual quality of the fused images. Furthermore, as reported in Table 7, the quantitative analysis results further demonstrate the significant advantages of the proposed method in terms of fusion performance.

5 Conclusion

This paper proposes a mutual promotion algorithm for infrared and visible image fusion and object detection, tailored for unaligned image scenarios. Considering the significant modality differences between infrared and visible images, we design specific loss functions to reduce such differences, thereby easing the difficulty of cross-modality image registration and improving its accuracy. In addition, we adopt a mutually beneficial learning strategy that enables the fusion task and the downstream object detection task to enhance each other, leading to improved quality in both the fused images and detection results. Extensive qualitative and quantitative experiments demonstrate the superiority of our method over existing state-of-the-art approaches. In addition, our approach can be used with other radiation frequencies where different modalities require image fusion like, for example, radio-frequency, x- and gamma rays used in medical imaging.

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

DH: Conceptualization, Methodology, Writing – review and editing, Writing – original draft, Investigation. KW: Writing – review and editing, Project administration, Data curation. CZ: Validation, Writing – review and editing, Formal Analysis. ZL: Methodology, Supervision, Writing – original draft, Funding acquisition, Writing – review and editing. YC: Formal Analysis, Visualization, Project administration, Writing – review and editing. SD: Resources, Data curation, Validation, Writing – review and editing. CK: Resources, Writing – review and editing, Formal Analysis.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article.

Conflict of interest

Authors DH, KW, CZ, ZL, YC, SD, and CK were employed by Yunnan Power Grid Co., Ltd.

The authors declare that this study received funding from the Science and Technology Project of China Southern Power Grid Co., Ltd. (No. YNKJXM20240052). The funder had the following involvement in the study: study design, collection, analysis, interpretation of data, the writing of this article, and the decision to submit it for publication.

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. Cao Y, Guan D, Huang W, Yang J, Cao Y, Qiao Y. Pedestrian detection with unsupervised multispectral feature learning using deep neural networks. *information fusion* (2019) 46:206–17. doi:10.1016/j.inffus.2018.06.005

2. Li C, Zhu C, Huang Y, Tang J, Wang L. Cross-modal ranking with soft consistency and noisy labels for robust rgb-t tracking. In: *Proceedings of the European conference on computer vision (ECCV)* (2018). p. 808–23.

3. Lin X, Li J, Ma Z, Li H, Li S, Xu K, et al. Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (2022). p. 20973–82.

4. Ha Q, Watanabe K, Karasawa T, Ushiku Y, Harada T. Mfnet: towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In: *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE (2017). p. 5108–15.

5. Li S, Kang X, Fang L, Hu J, Yin H. Pixel-level image fusion: a survey of the state of the art. *information Fusion* (2017) 33:100–12. doi:10.1016/j.inffus.2016.05.004

6. Li H, Qi X, Xie W. Fast infrared and visible image fusion with structural decomposition. *Knowledge-Based Syst* (2020) 204:106182. doi:10.1016/j.knosys.2020.106182

7. 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

8. Zhang Q, Liu Y, Blum RS, Han J, Tao D. Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: a review. *Inf fusion* (2018) 40:57–75. doi:10.1016/j.inffus.2017.05.006

9. 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) 98:116400. doi:10.1016/j.image.2021.116400

10. 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

11. 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

12. 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

13. 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

14. 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

15. 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

16. Ma J, Tang L, Xu M, Zhang H, Xiao G. Stdfusionnet: an infrared and visible image fusion network based on salient target detection. *IEEE Trans Instrumentation Meas* (2021) 70:1–13. doi:10.1109/TIM.2021.3075747

17. 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

18. 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

19. 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

20. 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

21. 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.

22. Tang L, Yuan J, Ma J. Image fusion in the loop of high-level vision tasks: a semantic-aware real-time infrared and visible image fusion network. *Inf Fusion* (2022) 82:28–42. doi:10.1016/j.inffus.2021.12.004

23. Sun Y, Cao B, Zhu P, Hu Q. Detfusion: a detection-driven infrared and visible image fusion network. In: *Proceedings of the 30th ACM international conference on multimedia* (2022). p. 4003–11.

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.

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: *2023 IEEE/CVF international conference on computer vision (ICCV)* (2023). p. 8081–90.

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. Yang Z, Zhang Y, Li H, Liu Y. Instruction-driven fusion of infrared-visible images: tailoring for diverse downstream tasks. *arXiv preprint arXiv:2411.09387* (2024).

28. Zhao W, Xie S, Zhao F, He Y, Lu H. Metafusion: infrared and visible image fusion via meta-feature embedding from object detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (2023). p. 13955–65.

29. 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

30. Huang Z, Liu J, Fan X, Liu R, Zhong W, Luo Z. Reconet: recurrent correction network for fast and efficient multi-modality image fusion. In: *European conference on computer vision (ECCV2022)* (2022). p. 539–55.

31. Wang D, Liu J, Fan X, Liu R. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In: *International joint conference on artificial intelligence (IJCAI)* (2022).

32. Wang D, Liu J, Ma L, Liu R, Fan X. Improving misaligned multi-modality image fusion with one-stage progressive dense registration. *IEEE Trans Circuits Syst Video Technology* (2024) 34:10944–58. doi:10.1109/tcsvt.2024.3412743

33. Xu H, Ma J, Yuan J, Le Z, Liu W. Rfnet: unsupervised network for mutually reinforcing multi-modal image registration and fusion. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (2022). p. 19679–88.

34. Xu H, Yuan J, Ma J. Murf: mutually reinforcing multi-modal image registration and fusion. *IEEE Trans Pattern Anal Machine Intelligence* (2023) 45:12148–66. doi:10.1109/tpami.2023.3283682

35. Tang L, Deng Y, Ma Y, Huang J, Ma J. Superfusion: a versatile image registration and fusion network with semantic awareness. *IEEE/CAA J Automatica Sinica* (2022) 9:2121–37. doi:10.1109/jas.2022.106082

36. 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

37. 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* (2023) 132:1625–44. doi:10.1007/s11263-023-01948-x

38. 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

39. 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.

40. 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

41. 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

42. 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

43. 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

44. 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

45. Xu H, Ma J, Le Z, Jiang J, Guo X. Fusiondn: a unified densely connected network for image fusion. In: *In proceedings of the thirty-fourth AAAI Conference on artificial intelligence* (2020).

46. Ma J, Ma Y, Li C. Infrared and visible image fusion methods and applications: a survey. *Inf Fusion* (2019) 45:153–78. doi:10.1016/j.inffus.2018.02.004

47. 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

48. 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

49. 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

50. He L, Todorovic S. Destr: object detection with split transformer. In: *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (2022). p. 9367–76.

51. Zhou S, Tan W, Yan B. Promoting single-modal optical flow network for diverse cross-modal flow estimation. *Proc AAAI Conf Artif Intelligence (Aaai)* (2022) 36:3562–70. doi:10.1609/aaai.v36i3.20268

52. 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

53. 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

54. 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 (CVPR)* (2024).

55. 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

56. Meng M, Feng D, Bi L, Kim J. Correlation-aware coarse-to-fine mlps for deformable medical image registration. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2024). p. 9645–54.

57. Tang W, He F, Liu Y, Duan Y. Matr: multimodal medical image fusion via multiscale adaptive transformer. *IEEE Trans Image Process* (2022) 31:5134–49. doi:10.1109/tip.2022.3193288

58. 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

59. 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