Journal of Beijing Institute of Technology, 2026, Vol. 35, No. 2

A Foreground-Guided Fusion Network for Infrared and Visible Images

Enqing Chen, Jinkai Feng, Song Wang✉, Qiang Li

(School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract: Infrared and visible image fusion aims to combine the complementary information from both modalities into a single image that simultaneously retains salient thermal targets and rich texture details. However, current fusion approaches mainly emphasize visual quality of the fused images, overlooking the compatibility with the downstream tasks. To address this issue, this paper proposes a foreground-guided fusion framework that adaptively enhances target regions while preserving global contextual information. Specifically, we design a two-branch network where the fusion branch aims to reconstructs high quality fused images while the foreground extraction branch captures semantic representations of salient objects to guide the fusion process toward target-related regions. To validate the effectiveness of the proposed framework, we build an aircraft keypoint dataset named VIRcraft to assess the performance. The fused images are also applied to semantic segmentation and object detection to verify the generalization of the proposed framework. The experimental results on different tasks demonstrate the superiority and generalization of the proposed fusion framework.

Keywords: infrared and visible image fusion; foreground guidance; generalization; aircraft keypoint detection

1 Introduction

Due to limitations in imaging hardware and imaging principles, a single modality sensor or a single imaging system only capture limited information for specific tasks [1, 2]. For instance, visible cameras rely on reflected light to provide texture and color details. However, under conditions such as nighttime, smoke and occlusion, they may fail to record key target information. In contrast, infrared sensors detect thermal radiation and highlight salient targets with strong robustness, but they usually suffer from low resolution and notable noise, making it difficult to preserve fine structural details [3, 4].

A representative example is provided in Fig. 1. In nighttime, the visible image offers environmental detail but is easily affected by illumination interference from aircraft lights. The infrared image captures the thermal radiation of the aircraft, making its outline and main components clearer. As shown in Fig. 1(c), the fused image combines salient target information from the infrared image with structural and ambient details from the visible image. Such a fusion improves both human observation and machine perception. With these advantages, image fusion is widely applied in nighttime monitoring [5] and intelligent detection [6].

In recent years, image fusion has attracted increasing attention. Early image fusion methods

Manuscript received Oct. 30, 2025; revised Jan. 9, 2026; accepted Mar. 18, 2026. The associate editor coordinating the review of this manuscript was Dr. Lijuan Jia. The work was supported by the National Natural Science Foundation of China (No. 62301497), the Science and Technology Research Program of Henan (No. 252102211024), the Key Research and Development Program of Henan (No.231111212000).

✉ Corresponding author. Email: ieswang@zzu.edu.cn

DOI: 10.15918/j.jbit1004-0579.2025.080

Infrared Visible Fusion

Fig. 1 Comparison of fusion and visible-infrared images

mainly rely on traditional signal processing techniques to combine information from different modalities, such as Laplacian pyramid [7], wavelet transform [8], and discrete cosine transform [9]. With the development of deep learning, researchers have further explored fusion models based on autoencoders (AEs) [10, 11], convolutional neural networks (CNNs) [12, 13] and generative adversarial networks (GANs) [14, 15], to improve the quality of fused images. These methods have achieved noticeable improvements in fusion performance and visual appearance. However, most existing approaches are task-independent and mainly focus on perceptual quality of the fused images, overlooking the compatibility with the various downstream tasks.

To address this issue, Tang et al. [16] and Zhang et al. [17] propose fusion methods guided by semantic perception and semantic segmentation. However, these methods mainly focus on the semantic segmentation task. PSFusion [3] is further introduced to incorporate a semantic perception branch that extracts task-relevant semantic information and gradually injects these cues into the fusion network to guide feature alignment between the infrared and visible modalities. Such a design maintains semantic consistency while preserving scene structure of the fused images. Similar to other task-driven approaches, PSFusion is designed for a specific task, which means that the model has to be retrained for other tasks.

To address the limitation, we propose a generalizable foreground-guided fusion framework which contains two cooperative branches to support downstream task performance and maintain strong task generalization. The fusion branch generates detail-preserving fused images using a dense reconstruction module and a residual fidelity module. The foreground extraction branch identifies the main targets in the scene through a semantic extraction module enhanced with multimodal attention. Both branches share a feature extraction backbone to learn shallow and deep representations jointly. The shallow features provide fine-grained texture information to assist reconstruction in the fusion branch, while the deep features encode semantic information that guides accurate target localization in the foreground extraction branch and delivers semantic cues to the fusion branch through a semantic injection module. Unlike most task-driven fusion methods that depend on specifically annotated datasets for different downstream applications, the proposed framework only requires foreground object masks as auxiliary supervision. This significantly improves its generalization across different tasks. Furthermore, to validate the effectiveness of proposed framework, we build an aircraft key-point dataset named VIRcraft to assess its performance. The fused images are also applied to semantic segmentation and object detection to verify the generalization of the proposed framework.

The main contributions of this paper are summarized as follows:

- We propose a generalizable foreground-guided fusion framework. The framework comprises two cooperative branches where the fusion branch reconstructs high quality fused images while the foreground extraction branch captures semantic representations of salient objects to guide the fusion process toward target-related regions.
- To validate the effectiveness of proposed framework, we build an aircraft key-point dataset, i.e. VIRcraft, with the annotation of foreground masks and keypoints to assess its performance.
- The fused images are applied to semantic segmentation and object detection to verify the

generalization of the proposed framework. The experimental results on different tasks demonstrate the superiority and generalization of the proposed fusion framework.

2 Methods

In this section, we provide a detailed description of the proposed image fusion framework. We first introduce the overall architecture and then describe the loss functions used for image fusion and semantic segmentation.

2.1 Overall Framework

The overall foreground-guided fusion framework is shown in Fig. 2. The proposed framework is composed of three modules: feature extraction, feature fusion, and semantic injection. Producing high quality fused images requires preserving high resolution feature maps, while extracting foreground objects relies on adequate downsampling to obtain deep semantic features. It is hard to meet these two requirements simultaneously within a single feature extraction network. Therefore, the proposed framework adopts a two branch architecture where one branch performs image fusion while the other is used for foreground object extraction. The deep semantic features from the foreground extraction branch are transferred to the fusion branch through the semantic injection module, guiding the fusion process.

We adopt a modified ResNet [18] as the feature extraction backbone, as shown in Fig. 3. To better balance the preservation of textural intricacy and the target saliency, we design a dual-output backbone consisting of a shallow feature extraction block (SFEB) and a deep feature extraction block (DFEB). SFEB utilizes 7×7 convolution kernels with a large receptive field in the initial layer to capture spatial context and removes downsampling layers to preserve the fine-grained resolution of the feature maps. Meanwhile, DFEB extracts semantic representations through a stride-2 convolution layer to facilitate target localization. The feature extraction process is defined as follows

$$\begin{aligned} I_{\mathrm{ir}}^{\mathrm{S}} &= \mathrm{SFEB}(I), \quad I_{\mathrm{vi}}^{\mathrm{S}} = \mathrm{SFEB}(V) \\ I_{\mathrm{ir}}^{\mathrm{D}} &= \mathrm{DFEB}(I), \quad I_{\mathrm{vi}}^{\mathrm{D}} = \mathrm{DFEB}(V) \end{aligned} \tag{1}$$

Here I_{ir} and I_{vi} denote the infrared and visible input images. SFEB generates high-resolution features $I_{\mathrm{ir}}^{\mathrm{S}}$ and $I_{\mathrm{vi}}^{\mathrm{S}}$ to preserve textural intricacy, while DFEB captures semantic features $I_{\mathrm{ir}}^{\mathrm{D}}$ and $I_{\mathrm{vi}}^{\mathrm{D}}$ target saliency preservation. This hierarchical

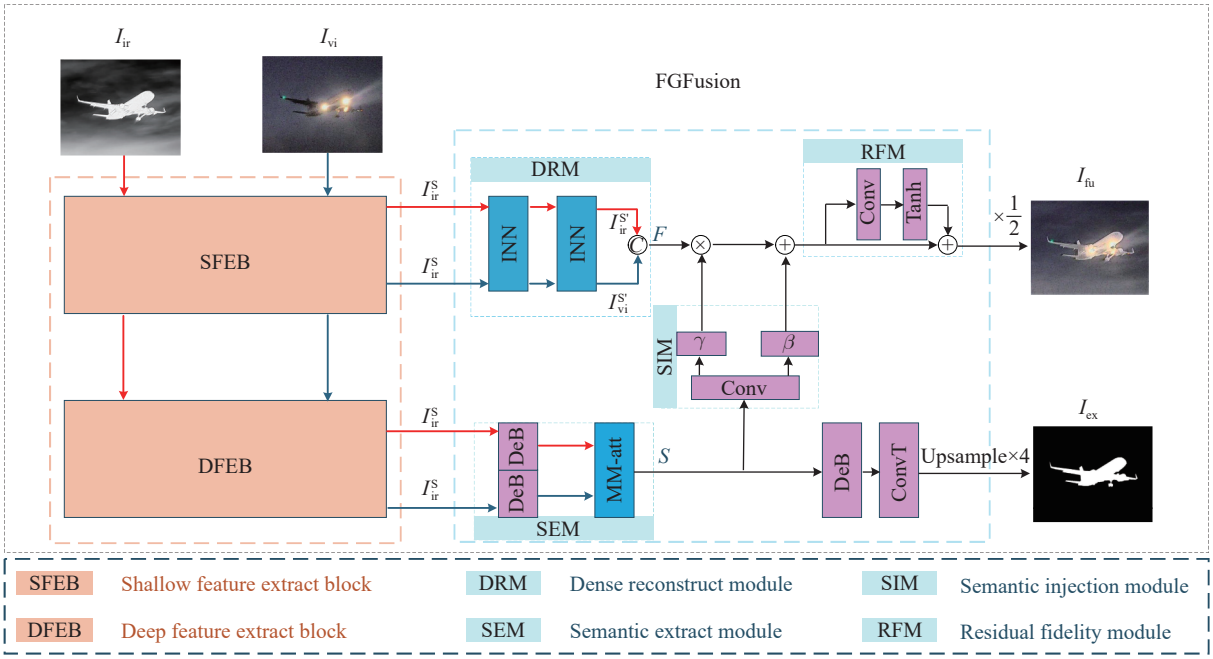

Fig. 2 The overall framework of the proposed Foreground-Guided fusion network

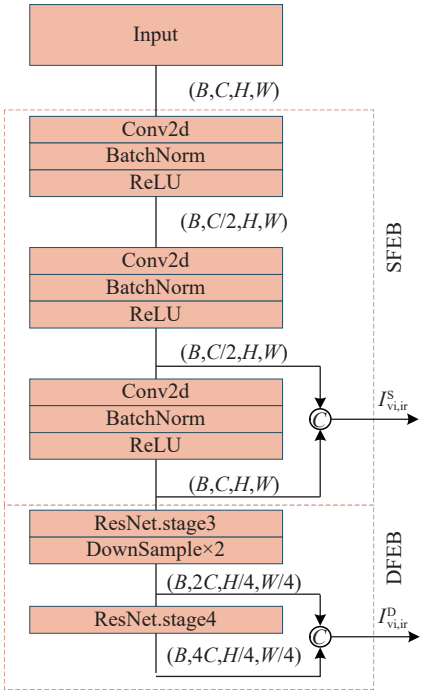

Fig. 3 Framework of feature extract block

design ensures that both high-frequency texture cues and low-frequency semantic information are fully preserved for the subsequent fusion process.

To make full use of the complementary characteristics of visible and infrared images, we design different fusion modules for the fusion branch and the foreground extraction branch.

In the fusion branch, we introduce a dense reconstruction module (DRM) to retain fine texture information. DRM is designed based on an invertible neural network (INN) [19], enabling the preservation of the original information. The proposed architecture utilizes a cross-modal coupling layer to establish a bijective mapping between visible and infrared modalities, ensuring a lossless information flow. As illustrated in Fig. 4, the layer allows the two modalities to interact with each other. Infrared structural information is directly embedded through an addition operation to strengthen object outlines, while visible textures are used to create weights that dynamically adjust the strength of the infrared signals through a multiplication operation. In this case, INN effectively preserves sharp infrared boundaries and fine visible textures, generating a fused image with rich complementary

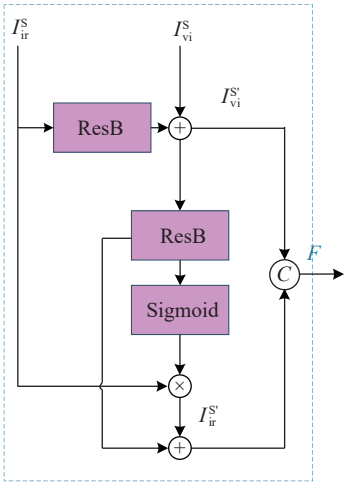

Fig. 4 Framework of invertible neural network

information. For the infrared input I_{ir} and the visible input I_{vi}, the INN transformation is expressed as

$$\begin{cases} I_{vi}^{S'} = I_{vi}^{S} + \mathrm{ReB}(I_{ir}^{S}) \\ I_{ir}^{S'} = I_{ir}^{S} \cdot \mathrm{Sigmoid}[\mathrm{ReB}(I_{vi}^{S})] + \mathrm{ReB}(I_{vi}^{S}) \end{cases} \tag{2}$$

where ReB denotes a Residual Block, and the Sigmoid function normalizes the fusion weights to the range (0,1). By stacking multiple INN blocks, the features of the infrared and visible modalities are iteratively updated, enabling effective interaction and complementary enhancement, as illustrated in Fig. 5.

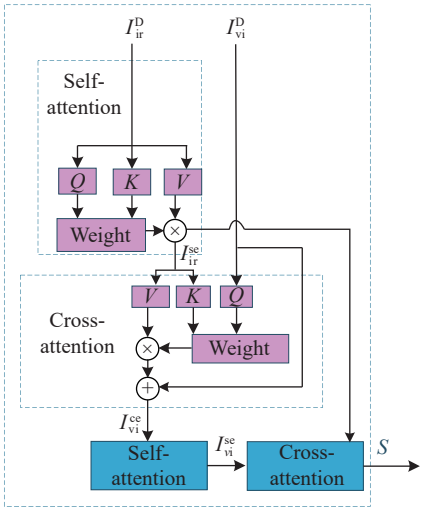

Fig. 5 Framework of multimodal attention

Unlike the fusion branch, the foreground extraction branch focuses on semantic understanding of salient regions. To accomplish this, we introduce a Multimodal Attention (MM-Att)

module that combines self-attention and cross-attention. This structure enhances modality specific representations while making use of complementary information between the infrared and visible inputs. For the input features I_{ir}^{s} and I_{vi}^{s}, the foreground extraction process is formulated as follows

$$\begin{cases} I_{\text{ir}}^{\text{se}} = \text{Self}(I_{\text{ir}}^{\text{D}}) \\ I_{\text{vi}}^{\text{ce}} = \text{Cross}(I_{\text{vi}}^{\text{D}}, I_{\text{ir}}^{\text{se}}) \\ I_{\text{vi}}^{\text{se}} = \text{Self}(I_{\text{vi}}^{\text{ce}}) \\ S = \text{Cross}(I_{\text{ir}}^{\text{se}}, I_{\text{vi}}^{\text{se}}) \end{cases} \tag{3}$$

Here, $\text{Self}(\cdot)$ denotes the self-attention mechanism [20], and $\text{Cross}(\cdot)$ denotes the cross-attention mechanism [21]. Based on these operations, we construct the Semantic Extract Module (SEM) to obtain discriminative semantic representations. Before deep feature fusion, a Dense Block (DeB) is applied to refine the high level features. The resulting deep semantic feature is represented as S.

To make use of the semantic information obtained from the foreground extraction branch, we design a Semantic Injection Module (SIM) which adaptively adjusts the normalization parameters based on the semantic features. The fused features are obtained by

$$F_{\text{u}} = \text{SIM}(F, S) = \gamma \otimes \text{Norm}(F) + \beta \tag{4}$$

Here, F and S denote the outputs of DRM and SEM, respectively. The parameters γ and β are computed as

$$\beta/\gamma = \text{Conv}_{\beta/\gamma}(\text{ReLU}(\text{Conv}(S))) \tag{5}$$

The fused features F_{u} are further processed by the Residual Fidelity Module (RFM) to generate the final fused image. This process preserves semantic information while maintaining natural visual appearance.

$$I_{\text{fu}} = \frac{\text{Tanh}(\text{Conv}(F_{\text{u}})) + 1}{2} \tag{6}$$

The foreground extraction result I_{ex} is obtained by restoring the spatial resolution of the semantic representation S through transposed convolution, which is expressed as

$$I_{\text{ex}} = \uparrow^{4} \text{Conv}(S) \tag{7}$$

2.2 Loss Function

For the fusion network, the objective is not only to produce fused images that are visually coherent, but also to introduce effective semantic representations. To achieve the objective, a fusion loss and a semantic loss are employed to ensure visual fidelity and guide the extraction of semantic features. Thus, the overall loss function is defined as

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{ssim}} + \frac{\sum (\lambda_2 \mathcal{L}_{\text{int}} + \lambda_3 \mathcal{L}_{\text{grad}} + \lambda_4 \mathcal{L}_{\text{ex}}) \odot W_{\text{mask}}}{\sum W_{\text{mask}}}. \tag{8}$$

where $\lambda_{1\sim4}$ are weighting coefficients for each loss term, W_{mask} denotes the mask weights, and $\odot$ indicates element-wise multiplication.

The fusion loss is composed of the gradient loss $\mathcal{L}_{\text{grad}}$, the structural similarity loss $\mathcal{L}_{\text{ssim}}$, and the intensity loss $\mathcal{L}_{\text{int}}$. The gradient loss is expressed as follows

$$\mathcal{L}_{\text{grad}} = \left\| |\nabla I_{f_{\text{u}}}| - \max(|\nabla I_{\text{ir}}|, |\nabla I_{\text{vi}}|) \right\|_1 \tag{9}$$

In this formulation, ∇ denotes the Sobel operator. By computing the L_1 distance between the gradient of the fused image and the maximum gradient of the infrared and visible inputs, the gradient loss encourages the fused result to retain prominent edge structures. The structural similarity loss $\mathcal{L}_{\text{ssim}}$ evaluates the structural consistency between the fused image and the input images using normalized cross correlation, with the objective of enhancing the correlation of structural features. To encourage the fused image to effectively retain foreground information from both input modalities, the intensity loss is introduced as

$$\mathcal{L}_{\text{int}} = \frac{1}{HW} \begin{cases} \sigma(I_{\text{ir}}, I_{\text{fu}}), & \text{if } \sigma(I_{\text{ir}}) > \sigma(I_{\text{vi}}) \\ \sigma(I_{\text{vi}}, I_{\text{fu}}), & \text{otherwise} \end{cases} \tag{10}$$

where $\sigma(\cdot)$ denotes the mean squared error. H and W represent the height and width of the image, respectively.

A cross entropy loss is applied to guide the

model in distinguishing foreground regions from background regions

$$\mathcal{L}_{\text{ex}} = -\frac{1}{N}\sum_{i=1}^{N}\left[M_i \lg(P_i) + (1 - M_i)\lg(1 - P_i)\right] \tag{11}$$

where M is a foreground mask that labels each pixel as either foreground or background. P denotes the predicted probability that each pixel belongs to the foreground.

3 Experiments

In this section, we first introduce the VIRcraft dataset and experimental setup. Next, we conduct comparative experiments to evaluate the performance of the proposed method on various downstream tasks. Finally, ablation studies are performed to validate the effectiveness of each component within the proposed framework.

3.1 VIRcraft Dataset

The images of the VIRcraft dataset are collected near the runway of an airport terminal. By recording 53 civil passenger or cargo aircraft during their landing process, we obtain 53 pairs of infrared–visible aircraft videos. The visual and infrared modalities are captured synchronously using a binocular infrared-visible camera system. The imaging parameters of the two sensors are listed in Tab. 1. Several sample frames extracted from the raw videos are shown in Fig. 6. Although the two video streams are read simultaneously, slight temporal offsets occurs when they reach the acquisition terminal. Therefore, the streams are first temporally aligned according to their inherent frame indices. Image frames are sampled at intervals of 10 frames.

Tab. 1 Imaging parameters of the infrared and visible cameras

Modality	Position	Channel	FPS	Width (pixel)	Height (pixel)	Focal length (mm)
Infrared	Lower	Gray	50	640	800	43.0–130.0
Visible	Upper	RGB	50	1080	1980	10.5–317.2

Fig. 6 Sample frames extracted from the raw videos

Due to the differences in imaging parameters between the two sensors, the captured image pairs exhibit spatial misalignment. The visible images have a larger field of view, within which the entire infrared image can be located. To achieve spatial alignment, we adopt a two-stage registration strategy which consists of coarse edge-based alignment followed by ORB-based fine registration. First, structural edge features of both modalities are extracted using an edge detection operator. Multi-scale template matching is then performed on the visible images to search for the corresponding infrared region, where the region with the highest matching response is selected as the coarse alignment result. To further improve geometric accuracy, both images are converted to grayscale. Keypoints along with their descriptors of the corresponding images are extracted. Feature matching is performed using Hamming distance [22]. Finally, the visible image is warped to the infrared image through a perspective transformation based on the estimated homography. The registration results are shown in Fig. 7. After coarse alignment and fine registration, a total of 216 pairs of infrared–visible images are obtained.

After obtaining the infrared–visible pairs,

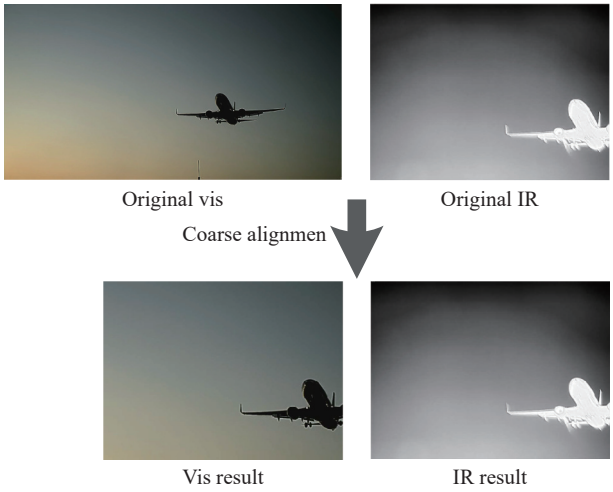

Fig. 7 Coarse edge alignment

the keypoint annotation is conducted using the Labelme software. Following the structural characteristics of civil passenger or cargo aircraft, we chose 9 keypoints based on the rigid body topology of the aircraft. The selection of these keypoints is critical for defining the aircraft's pose during landing. Specifically, the Head, Vertical, and Tail define the central axis of the fuselage, while the wing-based keypoints (Left/Right_front, Left/Right_root, Left/Right_behind) describe the roll angle and span. According to visibility conditions, the keypoints are categorized into three classes: visible, occluded, and invisible. During model training for keypoint detection, these three types are assigned different weights 1.0, 0.5, and 0, respectively.

An overview of the keypoint distribution is shown in Fig. 8. In most cases, the number of visible keypoints substantially exceeds that of occluded keypoints, indicating that the aircraft structures are generally well preserved in the captured frames. In terms of specific distribution, due to the side-view observation angle, the left and right wingtips are often occluded and difficult to precisely localize, resulting in a relatively high proportion of invisible annotations for these keypoints. Although the tail end is also frequently occluded, its position remains easier to infer structurally, leading to a higher proportion of occluded rather than invisible labels.

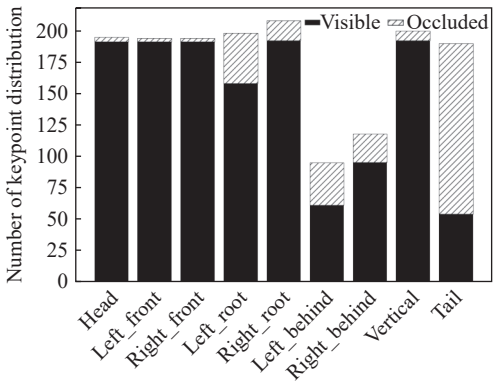

Fig. 8 Annotation of keypoints and foreground mask

The annotations are illustrated in Fig. 9. There are three types of annotations in VIRcraft, e.g. bounding box, keypoint and semantic mask. While the bounding box and the keypoint localize the foreground targets and nine meaningful points on the airframe separately, the mask identifies the pixel-level semantics in the samples, supporting the implementation of the proposed fusion framework.

(a) (b)

Fig. 9 Annotations of VIRcraft: (a) keypoint and box; (b) mask

Prevalent infrared and visible image fusion datasets, such as M3FD [14] and MSRS [16], primarily cater to urban surveillance and autonomous driving. Their targets are largely restricted to common ground objects like pedestrians and vehicles. In terms of annotation, only bounding boxes or masks are provided in these datasets. In this paper, VIRcraft is specifically tailored for the aviation domain. Except for bounding box and mask annotations, VIRcraft also includes keypoint labels.

3.2 Experimental Setup

To evaluate the proposed fusion method, we conducted experiments on the self-constructed VIRcraft dataset and the publicly available MSRS and M3FD datasets. In addition to the infrared and visible image pairs, each dataset contains specific annotations, for example, aircraft keypoint annotations in VIRcraft, pixel-level semantic segmentation labels in MSRS and object bounding boxes in M3FD.

In the experiments, the proposed framework is compared with five state-of-the-art (SOTA) fusion methods: MRFS [17], GIFNet [23], CDDFuse [10], PSFusion [3], and TarDAL [14]. The implementations and experimental settings of these methods are publicly available.

We first performed image fusion experi-

ments on the VIRcraft, MSRS, and M3FD datasets to evaluate the visual quality of the fused results. Subsequently, the fused images were applied to aircraft keypoint detection, semantic segmentation, and object detection tasks. Both qualitative and quantitative assessments are performed to examine the influence of the fusion results on downstream high-level vision applications.

All experiments are carried out on a workstation configured with an Intel Core i7-13700K processor (13th Generation, 3.4 GHz, 16 cores and 24 threads), 64 GB RAM and an NVIDIA Tesla V100 GPU with 32 GB memory. In the preprocessing stage, the training samples are randomly cropped and rotated into 256×256 image patches. The total number of training epochs is set to 500 with the batch_size = 24. The optimizer is SGD with an initial learning rate of 10^{-3} and a weight decay of 5×10^{-4}. For the loss function, the hyperparameters λ_1 to λ_4 are assigned values of 1, 10, 10, and 5, respectively. Mask weight W_{mask} is set to 3.

3.3 Results Analysis

We first conduct image fusion experiments on the VIRcraft, MSRS, and M3FD datasets to evaluate the visual quality of the fused results. Subsequently, the fused images are applied to aircraft keypoint detection, semantic segmentation, and object detection tasks. Both qualitative and quantitative assessments are performed to examine the influence of the fusion results on downstream high-level vision applications.

3.3.1 Image Fusion

To assess the fusion performance of the proposed framework under different conditions, fused images are produced on three datasets and compared with existing fusion approaches. The comparison emphasizes preservation of texture detail, clarity of structural contours and contrast in target regions. Two representative scenes from each dataset are selected to illustrate the advantages of the proposed framework in visual analysis.

Compared with other fusion approaches, the proposed framework highlights foreground targets more effectively, yielding fusion results with sharper, more natural structural boundaries and richer texture details. For instance, in scene "218" from the VIRcraft dataset (Fig. 10(a)), both the proposed framework and TarDAL successfully enhance the aircraft's silhouette. Other methods suffer from lighting variations that blur the distinction between foreground and background. Similarly, in scene "70", MRFS and CDDFuse exhibit poorly defined boundaries, while TarDAL loses critical texture information. In scene "00710N" from the MSRS dataset (Fig. 10(c)), both PSFusion and the proposed framework achieve reasonable separation between the pedestrians and the background. Nevertheless, only the proposed framework preserves the intricate textures of ground cracks, producing more continuous and high-contrast edges. In scene "00298D", our method is the only one to clearly distinguish the buildings from the background, as highlighted within the red boxes. In scene "00922" from the M3FD dataset (Fig. 10(e)), the proposed framework distinctly separates the foliage from the sky, creating a pronounced sense of depth. Furthermore, the proposed framework successfully restores the contours of the shipping containers obscured by smoke in scene "00968".

In general, the proposed framework effectively highlights structural patterns and key texture information in regions of interest, resulting in fused images with enhanced contrast and improved detail representation. This yields a more reliable input for high-level vision tasks.

3.3.2 Keypoint Detection

HRNet-w32 [24] is adopted as the keypoint detection network for experiments on the VIRcraft dataset. Detection accuracy is evaluated using average precision (AP). The detection model is trained using the Adam optimizer, with a batch size of 6 and a total of 200 training epochs. The initial learning rate is set to 10^{-3}. The weight decay coefficient is 5×10^{-4}. A weighted cross-

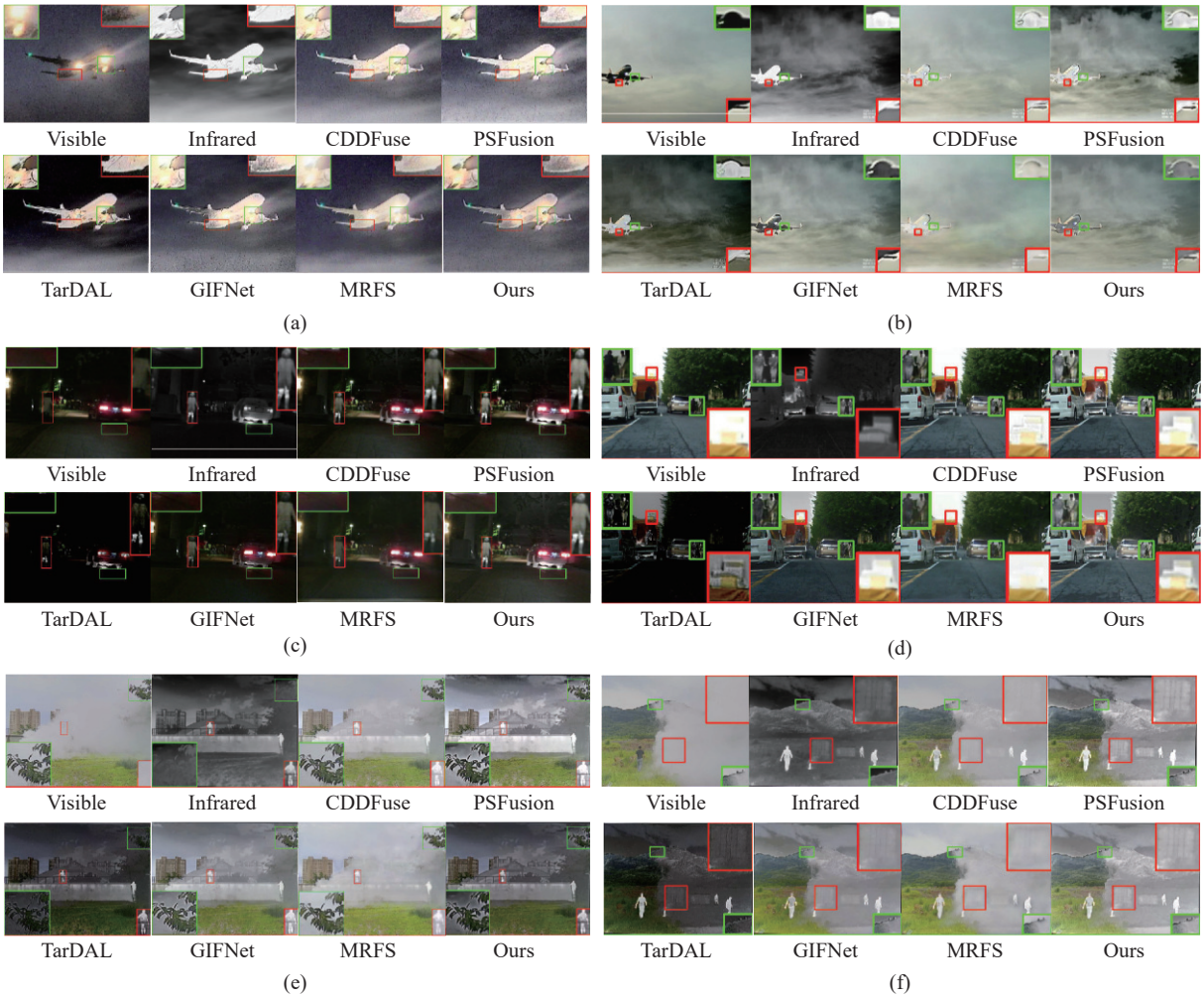

Fig. 10 Fusion comparison of the proposed framework with 5 SOTA methods: (a) scene "218" from VIRcraft; (b) scene "70" from VIRcraft; (c) scene "00710N" from MSRS; (d) scene "00298D" from MSRS; (e) scene "00922" from M3FD; (f) scene "00968" from M3FD

entropy loss is employed for supervision, where weights of 1, 0.5, and 0 are assigned to visible, occluded, and invisible keypoints, respectively, to enhance the model's ability to learn keypoint representations under varying visibility conditions.

The aircraft keypoint detection results are presented in Tab. 2, where bold values and underlined values indicate the best and the second best performance. From Tab. 2, it is observed that the proposed framework achieves the highest detection accuracy.

Tab. 2 Precision of aircraft keypoint detection

Models	AP75	mAP[50,95]
Visible	60.29	64.85
Infrared	61.76	63.38
CDDFuse	60.29	64.26
PSFusion	<u>63.24</u>	<u>65.44</u>
TarDAL	61.76	63.97
MRFS	61.76	64.85
GIFNet	60.29	64.85
Ours	**64.71**	**66.18**

To intuitively investigate the impact of different fusion strategies on feature discriminability, we visualize the confidence heatmaps generated by the fusion strategies in Fig. 11. In the heatmaps, regions with higher response values are shown in redder areas, indicating a greater likelihood of corresponding to the target keypoint locations. This phenomenon refers to the displacement of the peak activation in the confidence heatmap away from the ground-truth physical location. For occluded keypoints, such as

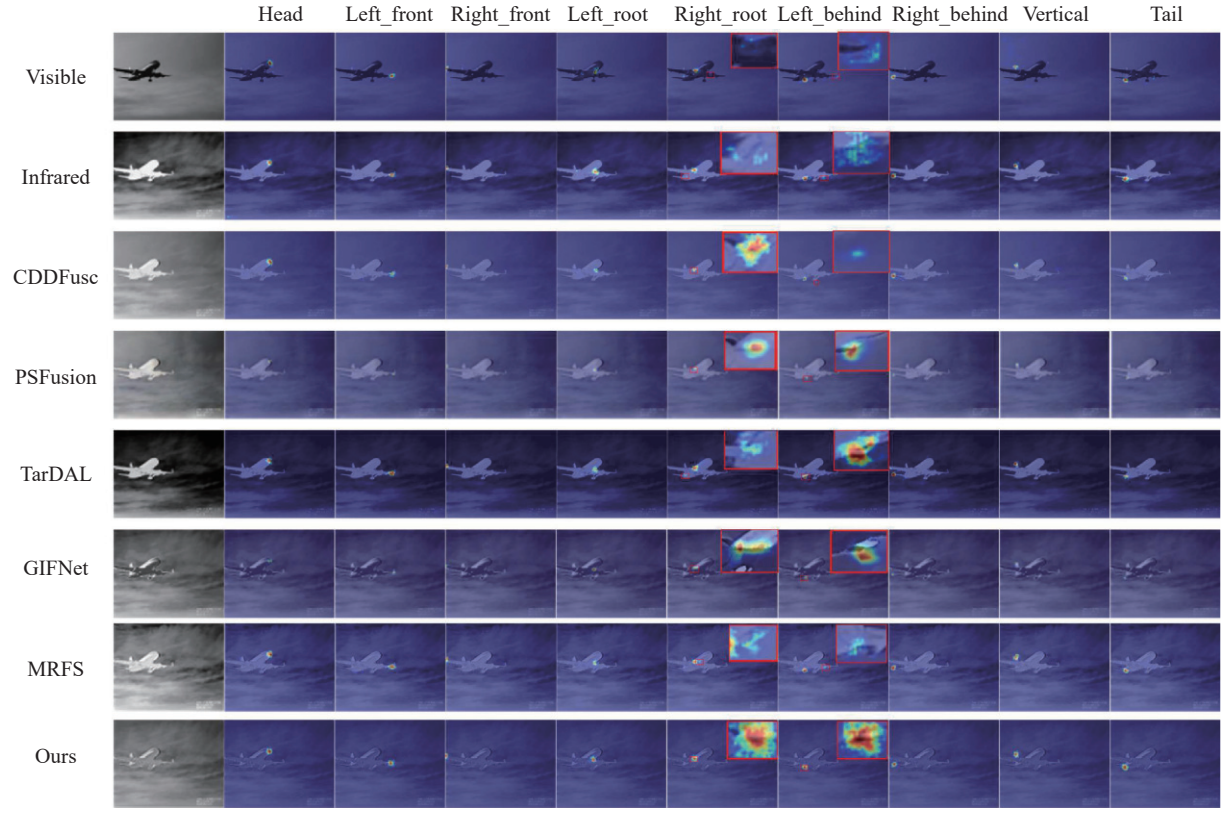

Fig. 11 Heatmap prediction comparison of the proposed framework with 5 SOTA methods on the scene "216" from VIRcraft dataset

the Left_behind in the example, there is a response drift when only the visible or infrared modality is used. Likewise, fusion methods without explicit task guidance, such as CDDFuse and TarDAL, fail to accurately localize these keypoints. GIFNet yields two response peaks near the right-wing tip, while PSFusion produces comparatively accurate predictions. By incorporating foreground guidance to align semantic features, the proposed framework achieves stable localization without response drift as shown in Fig. 11.

3.3.3 Semantic Segmentation

DeepLabV3+ [25] is adopted as the semantic segmentation model to conduct experiments on the MSRS dataset. MSRS focuses on urban road scenes and contains 1 444 pairs of infrared–visible images across nine object categories, covering both daytime and nighttime conditions. The segmentation performance is evaluated using the intersection over union (IoU) metric. All models are trained with a cross-entropy loss function using the Adam optimizer with a batchsize of 12 and a total of 300 training epochs. The initial learning rate is set to 10^{-3}.

The semantic segmentation results are reported in Tab. 3. Across the nine object categories in the MSRS dataset, the proposed framework achieves either the best or second performance, and attains the highest overall mIoU. This improvement is attributed to the introduction of a foreground extraction module that explicitly distinguishes foreground from background, providing clearer and more informative input for segmentation. Furthermore, assigning higher weights to foreground regions during training encourages the model to focus on salient objects, thereby enhancing segmentation accuracy.

Fig. 12 and Fig. 13 presents the visualization of segmentation results, including a daytime scene (00298D) and a nighttime scene (00750N) from the MSRS dataset. In both cases, only the proposed framework avoids misclassifying background areas as guardrails, demonstrating its strong semantic discrimination capability. In the nighttime example, both the proposed frame-

Tab. 3 Accuracy of semantic segmentation

Method	mIoU	BG	Car	Per	Bik	Cur	CS	GR	CC	Bum
Visible	97.43	84.01	53.32	64.81	40.01	51.00	54.59	53.97	63.68	62.54
Infrared	96.87	76.37	67.41	56.76	33.10	28.73	27.84	28.77	58.56	52.71
CDDFuse	97.91	**85.46**	67.13	66.90	46.67	59.73	59.31	59.17	**70.18**	68.05
PSFusion	97.88	84.63	**68.82**	67.01	48.74	57.71	58.82	57.54	65.80	67.44
TarDAL	96.63	73.42	63.25	56.70	21.91	36.90	43.19	26.41	42.32	51.19
GIFNet	97.64	82.70	64.56	66.30	46.48	51.99	56.78	41.78	64.64	63.65
MRFS	97.69	83.50	64.14	65.02	47.69	58.37	**64.98**	52.01	61.30	66.08
Ours	**98.11**	84.95	68.55	**69.71**	**52.82**	**64.74**	59.90	**60.34**	69.87	**69.89**

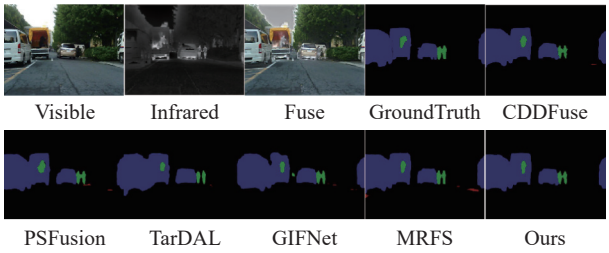

Fig. 12 Semantic segmentation comparison of the proposed framework with 5 SOTA methods on the "00298D" scene from MSRS

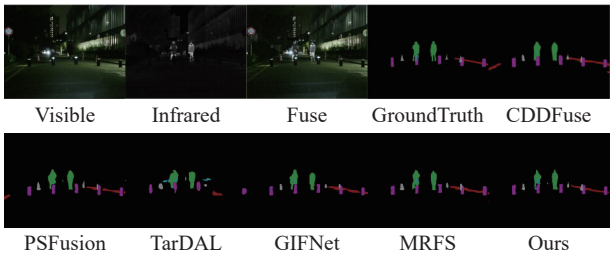

Fig. 13 Semantic segmentation comparison of the proposed framework with 5 SOTA methods on the "00750N" scene from MSRS

work and TarDAL correctly identify the bicycle category, but the performance of TarDAL in other regions drops. In contrast, the proposed framework maintains consistently reliable segmentation across different scenes.

There are also some failure cases in semantic segmentation. As shown in Fig. 14, the "Car" object is misclassified as the "Person" by the proposed framework. One possible explanation is that FGFusion overemphasizes the distinction between the foreground targets and the background during the fusion phase, overlooking the differences within the foreground targets. Such an explanation is also validated in Tab. 3, where the segmentation performance of the proposed

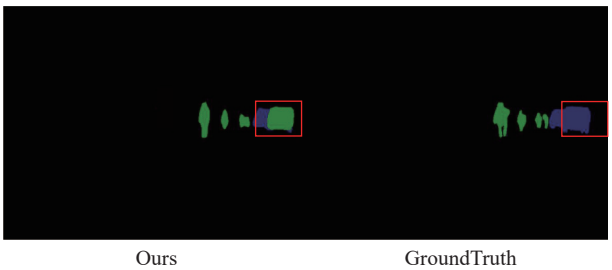

Fig. 14 The failure segmentation case caused by FGFusion

framework for some targets is slightly worse than the existing methods. Nevertheless, the proposed framework maintains a substantial advantage in overall segmentation accuracy, illustrating its effectiveness and robustness on semantic segmentation.

3.3.4 Object Detection

Object detection experiments are conducted on the M3FD dataset. M3FD contains 4 200 pairs of visible and infrared images. The performance is evaluated using mean average precision (mAP). The experiments adopt the YOLOv5 framework as the base network. The models are trained for 300 epochs with a batch size of 8. The other experimental settings follow the default configuration of YOLOv5.

The quantitative performance of the object detection task is reported in Tab. 4. On the M3FD dataset, the proposed method achieves the highest detection accuracy, which proves that incorporating foreground guidance into the fusion process provides more discriminative features for downstream tasks. Unlike methods that only focus on visual quality, our framework effectively bridges the gap between image fusion and high-level scene understanding.

Tab. 4 Precision of object detection.

Models	AP50	mAP[50,95]
Visible	80.9	53.9
Infrared	77.9	51.9
CDDFuse	80.9	54.5
PSFusion	81.2	55.1
TarDAL	81.0	54.7
GIFNet	80.3	54.3
MRFS	81.1	55.0
Ours	**81.4**	**55.4**

Visualization results in Fig. 15 further demonstrate the superiority of the proposed framework. For the third person from the right in the image, GIFNet and TarDAL suffer from "over-detection", misidentifying one individual as multiple objects due to blurred edges. Meantime, the proposed framework achieves a leading confidence score of 0.84 for the person object, surpassing the other method MRFS with 0.73 confidence score. This is because the proposed foreground-guided framework generates better boundaries and suppresses background clutter, ensuring the semantic integrity of objects even in complex environments.

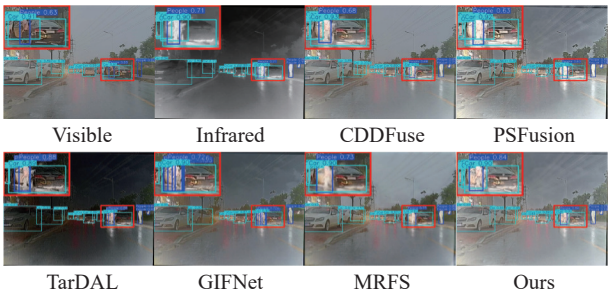

Fig. 15 Object detection comparison of the proposed framework with 5 SOTA methods on the "00052" scene from M3FD

3.4 Ablation Experiments

Ablation experiments are conducted to evaluate the contributions of specific components in the proposed method. Three variants are tested: (i) removing the foreground extraction branch in the dual-branch architecture (w/o Ex), leaving only the fusion branch; (ii) removing the feature interaction design (w/o FE), where INN and MM-Att are replaced with dense connections; and (iii) removing mask weighting (w/o MASK), resulting in uniform loss weights for foreground and background regions.

Tab. 5 summarizes the effects of these variants on aircraft keypoint detection, semantic segmentation, and object detection. It is observed from Tab. 5 that removing any of the components leads to a noticeable decrease in performance. For example, eliminating the semantic injection and the foreground extraction branch reduces the amount of semantic information introduced into the fused images. Mask weighting is designed to focus on target regions. Removing it reduce the model's emphasis on salient objects. INN and MM-Att play a key role in capturing complementary information between infrared and visible modalities. Without them, the model becomes unable to effectively model inter-modal dependencies, leading to a simple overlay of modalities, blurred target contours, and loss of structural detail.

Tab. 5 Performance of each variant in the ablation study

Models	KDetection (mAP[50,95])	SSegmentation (mIoU)	ODetection (mAP[50,95])
Baseline	58.37	61.83	54.4
w/o Ex	61.63	62.93	54.9
w/o FE	62.31	63.66	55.1
w/o MASK	63.94	58.84	54.4
Ours	**66.18**	**69.89**	**55.4**

4 Conclusion

This paper presents a foreground-guided infrared–visible image fusion framework that incorporates explicit semantic guidance and modality interaction to enhance downstream high-level vision tasks. By introducing a foreground extraction branch and a mask-weighted loss strategy, the proposed framework effectively emphasizes target regions and suppresses irrelevant background information. Extensive experiments on image fusion, aircraft keypoint detection, semantic segmentation and object detection demonstrate that the proposed framework achieves superior visual quality and performance on downstream tasks. In

future work, we plan to extend the proposed framework to broader modalities and apply it real-world applications.

References:

[1] Y. Liu, X. Chen, H. Peng, and Z. Wang,"Deep learning for pixel-level image fusion: Recent advances and future prospects," *Information Fusion*, vol. 42, pp. 158-173, 2018.

[2] H. Zhang, H. Xu, T. Xie, and J. Ma,"Image fusion meets deep learning: A survey and perspective," *Information Fusion*, vol. 76, pp. 323-336, 2021.

[3] L. Tang, H. Zhang, H. Xu, and J. Ma,"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," *Information Fusion*, vol. 99, pp. 101870, 2023.

[4] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma,"Piafusion: A progressive infrared and visible image fusion network based on illumination aware," *Information Fusion*, vol. 83, pp. 79-92, 2022.

[5] K. Yang, W. Xiang, Z. Chen, J. Zhang, and Y Liu.,"A review on infrared and visible image fusion algorithms based on neural networks," *Information Fusion*, vol. 99, pp. 101870, 2024.

[6] X. Z. Wang, C. L. Zhang, J. M. Hu, Q. Wen, G. F. Zhang, and M. Huang.,"AEFusion:Adaptive enhanced fusion of visible and infrared images for intelligent driving," *Remote Sensing*, vol. 17, no. 18, pp. 3129, 2025.

[7] P. J. Burt and E. H. Adelson, "The laplacian pyramid as a compact image code," in *Readings in Computer Vision*. Morgan Kaufmann, pp. 671-679,1987.

[8] H. Li, B. S. Manjunath, and S. K. Mitra,"Multisensor image fusion using the wavelet transform," *Graphical Models and Image Processing*, vol. 57, no. 3, pp. 235-245, 1995.

[9] N. Ahmed, T. Natarajan, and K. R. Rao,"Discrete cosine transform," *IEEE Transactions on Computers*, vol. C-23, no. 1, pp. 90-93, 1974.

[10] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool, "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)*, pp. 5906-5916, 2023.

[11] H. Li and X. J. Wu,"Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614-2623, 2019.

[12] D. Han, W. Zhang, Y. Liu, L. Li, and J. Ma,"Multi-exposure image fusion via deep perceptual enhancement," *Information Fusion*, vol. 79, pp. 248-262, 2022.

[13] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, and J. Jiang,"Infrared and visible image fusion via detail preserving adversarial learning," *Information Fusion*, vol. 54, pp. 85-98, 2020.

[14] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "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)*, pp. 5802-5811, 2022.

[15] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502-518, 2020.

[16] L. Tang, H. Zhang, and J. Ma,"Image fusion in the loop of high-level vision tasks: A semantic-aware real-time image fusion network," *Information Fusion*, vol. 82, pp. 28-42, 2022.

[17] H. Zhang, X. Zuo, J. Jiang, C. Guo, and J. Ma, "Mrfs: Mutually reinforcing image fusion and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26 974-26 983, 2024.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016.

[19] L. Ardizzone, J. Kruse, K. Wickstrøm, C. Rother, and U. Köthe, "Analyzing inverse problems with invertible neural networks," *arXiv preprint arXiv: 1808.04730*, 2018.

[20] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[21] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint,* arXiv: 1908.07490, 2019.

[22] R. W. Hamming,"Error detecting and error correcting codes," *The Bell System Technical Journal*, vol. 29, no. 2, pp. 147-160, 1950.

[23] C. Cheng, T. Xu, and Z. Feng, "One model for all:

Low-level task interaction is a key to task-agnostic image fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28 102-28 112, 2025.

[24] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5693-5703, 2019.

[25] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801-818, 2018.

Enqing Chen received the Ph.D. degree in communication and information system from the Beijing Institute of Technology, China, in 2007. He is currently a Professor with the School of Electrical and Information Engineering, Zhengzhou University. His research interests are in the areas of machine learning and signal processing, including computer vision, human action recognition, and target detection.

Jinkai Feng received his B.E. degree from the School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou, China. He is currently pursuing a Master's degree at the School of Electrical and Information Engineering, Zhengzhou University. His research interests include image processing, computer vision, and object detection.

Song Wang received the Ph.D. degree in information and communication engineering from Zhengzhou University, Zhengzhou, China. He is currently a Lecturer with the School of Electrical and Information Engineering, Zhengzhou University. His research interests include image processing, image matching and multimedia signal processing.

Qiang Li is currently a Ph.D. candidate and a Senior Engineer at the 27th Research Institute of China Electronics Technology Group Corporation. His research interests include visual measurement and object detection.