E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP18;TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)04-1184-17

论文引用格式:Yang T Y, Huo H T, Guo B F, Zheng B W and Liu X W. 2026. Multilevel Mamba network for infrared and visible image fusion. Journal of Image and Graphics, 31(4):1184-1200(杨天宇, 霍宏涛, 郭宝峰, 郑博文, 刘晓文. 2026. 用于红外与可见光图像融合的多层级Mamba网络. 中国图象图形学报, 31(4):1184-1200)[DOI:10.11834/jig.250243]

用于红外与可见光图像融合的多层级Mamba网络

杨天宇,霍宏涛*,郭宝峰,郑博文,刘晓文

中国人民公安大学信息网络安全学院,北京　100038

摘　要:**目的**　现有融合方法普遍存在多层级语义信息的表征退化问题,且缺乏有效的跨层级特征交互机制,导致浅层细节与深层语义信息在融合过程中难以完全耦合。此外,基于Transformer的融合方法在全局特征建模的过程中需要消耗大量计算资源。针对上述问题,提出了一种用于红外与可见光图像融合的多层级Mamba网络。**方法**　融合网络通过构建多层级特征框架,对多分辨率源图像进行全局特征建模与跨层级特征交互,实现了跨模态图像细粒度语义信息的有效保留。同时,特征编码阶段设计F-Mamba模块,在维持线性复杂度的同时,实现了全局特征提取。此外,模型通过设计跨层级特征聚合模块,实现了不同层级间视觉特征与语义信息的深度对齐。**结果**　实验在MSRS(multispectral road scenarios)、LLVIP(visible-infrared paired dataset for low-light vision)和RoadScene数据集上与13种传统以及深度学习融合方法进行比较。主观评价方面,融合结果在目标细节特征恢复以及视觉质量方面具有显著优势;客观指标方面,在MSRS数据集上本文算法在信息熵、空间频率、视觉保真度、峰值信噪比、平均梯度和边缘强度6项指标上取得最优值,相比于对比方法最优值分别提升了3.03%、1.56%、15.89%、7.26%、2.61%和1.62%。在LLVIP数据集上本文算法在空间频率、峰值信噪比、平均梯度和边缘强度4项指标上取得最优值,相比于对比方法最优值分别提升了6.42%、0.45%、6.47%和7.23%。在RoadScene数据集上本文算法在平均梯度和边缘强度2项指标上仍取得最优值。消融实验验证了本文融合网络各组件的有效性。此外,运行效率对比实验和语义分割实验,进一步验证了本文算法在计算效率和深层语义信息保留方面的优势。**结论**　提出了基于Mamba的多层级红外与可见光图像融合网络,在源图像多层级语义特征保留、目标细节特征恢复以及计算效率等方面均具有优越性。

关键词:图像融合;多层级;Mamba;特征聚合;深度学习

Multilevel Mamba network for infrared and visible image fusion

Yang Tianyu, Huo Hongtao*, Guo Baofeng, Zheng Bowen, Liu Xiaowen

Department of Information and Cyber Security, People's Public Security University of China, Beijing　100038, China

Abstract: **Objective** Some network parameters remain in an unstable optimization state during training because of the lack of ground truth images for supervision in infrared and visible image fusion tasks. As a result, deep semantic information from the source images is difficult to preserve effectively during the fusion process. Existing fusion methods commonly suffer from degradation in multilevel semantic representation and lack effective cross-level feature interaction mechanisms; thus, full integration of shallow details with deep semantic information during fusion becomes difficult. In addition, transformer-based fusion methods incur substantial computational overhead when modeling global features. This study addresses these challenges by proposing a multilevel infrared and visible image fusion network based on Mamba, thereby

收稿日期:2025-06-11;**修回日期**:2025-11-13;**预印本日期**:2025-11-20

*通信作者:霍宏涛　huohongtao@ppsuc.edu.cn

leveraging recent advances in state space modeling. **Method** In this study, we propose a novel infrared and visible image fusion algorithm that constructs a multilevel feature extraction architecture. The model effectively alleviates the loss of semantic information during deep network training by employing a progressive downsampling strategy and dense connections during the feature encoding phase, thereby preserving fine-grained textures and high-level semantic features from the source images. During the feature extraction stage, we introduce an innovative F-Mamba module that leverages the selective memory mechanism of state space models and hardware-aware algorithms. This design mitigates the limitations of traditional convolutional receptive fields while maintaining linear computational complexity, thereby enabling the network to capture cross-level feature representations ranging from local textures to global semantic information efficiently. A cross-level feature aggregation module is proposed to enhance the extraction of complementary features further across levels. This module employs multiscale dilated convolutions to align and fuse shallow visual features with deep semantic features, thereby achieving effective preservation of fine-grained semantic information in cross-modal images. **Result** Comparative experiments were conducted on the multispectral road scenarios (MSRSs), visible-infrared paired dataset for low-light vision (LLVIP), and a new dataset of aligned infrared and visible images (RoadScene) to compare 13 traditional and deep learning-based fusion methods. In the subjective evaluation, the proposed method demonstrated a clear advantage in the restoration of target detail features and visual quality. With regard to the objective evaluation, our method achieved optimal values in six objective metrics on the MSRS dataset: entropy, spatial frequency (SF), visual information fidelity, peak signal-to-noise ratio (PSNR), average gradient (AG), and edge intensity (EI). Compared with the best values in the six metrics of 13 existing fusion algorithms, those of our method reached an average increase of 3.03%, 1.56%, 15.89%, 7.26%, 2.61%, and 1.62%. The values achieved by our method in SF, PSNR, AG, and EI on the LLVIP dataset are the best among all the values. Compared with the best values in the four aforementioned metrics of 13 existing fusion algorithms, those of our method reached an average increase of 6.42 %, 0.45%, 6.47%, and 7.23%. Moreover, the values achieved by our method in AG and EI on the RoadScene dataset are the best among all the values. In addition, we validated the effectiveness of each component in the network through ablation experiments. In the computational efficiency comparison experiments, the proposed method demonstrates significant advantages over most transformer-based and Mamba-based approaches. In semantic segmentation experiments, the proposed method outperforms the second-best approach by 0.42% in terms of mean intersection over union (mIoU), thereby demonstrating the effectiveness of our fusion algorithm in preserving multilevel semantic features. **Conclusion** In this study, we propose a Mamba-based multilevel fusion network that integrates a hierarchical feature extraction architecture with the F-Mamba module to preserve deep semantic features from the source images effectively while maintaining linear computational complexity. Experimental results show that, compared with 13 existing fusion methods, the proposed approach demonstrates superior performance in preserving fine-grained semantic features, thereby restoring target details and achieving high computational efficiency.

Key words: image fusion; multi-level; Mamba; feature aggregation; deep learning

0 引　言

单一模态成像设备受其成像原理制约，往往难以全面捕捉复杂场景信息。例如，红外图像能够在复杂环境下有效突出目标，但通常缺乏丰富的纹理和细节信息，空间分辨率有限。相比之下，可见光成像能提供高分辨率的场景纹理，却易受光照及天气条件影响（金伟其 等，2023）。红外与可见光图像融合旨在通过高效的特征提取与融合策略，整合不同模态图像中的互补信息，生成一幅目标显著、纹理细节信息丰富的融合图像。红外与可见光图像融合技术已广泛应用于目标检测（Wang 等，2023）、遥感监测（Amarsaikhan 等，2012）和安防监控（李国梁 等，2022）等领域。

现有红外与可见光图像融合方法可分为传统方法与基于深度学习的方法。传统的图像融合方法，如基于多尺度变换（陈木生，2016）、稀疏表示（杨培等，2021）、显著性分析（Zhang 等，2015）以及子空间分解（宫睿和王小春，2019）等，严重依赖于人工设计的特征提取规则与融合策略。这种依赖导致其自适应能力有限，在面对复杂多变的成像场景时，往往难以稳定地生成兼具显著目标和清晰纹理的高质量融合结果。

近年，深度学习在红外与可见光图像融合领域的广泛应用，一定程度上解决了传统方法的不足。根据网络架构区别，当前基于深度学习的融合方法可以分为基于自编码器的方法（auto-encoder，AE）、基于卷积神经网络（convolutional neural network，CNN）的方法、基于生成对抗网络（generative adversarial network，GAN）的方法以及基于Transformer的方法。

基于自编码器的方法通过编码器对源图像进行特征提取与编码，再经融合模块对特征进行融合，最后由解码器重建融合图像。例如，Li和Wu（2019）通过引入密集连接增强特征提取能力，并结合编码器—解码器架构，有效提升融合性能。基于卷积神经网络的方法以端到端方式自适应完成特征提取、图像融合和图像重建任务。例如，Xu等人（2022）针对图像融合领域存在的多任务兼容性不足及持续学习中的灾难性遗忘问题，提出统一的无监督图像融合框架U2Fusion。基于生成对抗网络的方法通过生成器与判别器的对抗训练机制，学习源图像的分布特征，以实现多模态图像信息的有效整合。例如，Liu等人（2022）提出基于目标感知的双对抗学习网络，可实现图像融合与目标检测的协同优化。

2020年，Dosovitskiy等人（2021）突破性地将Transformer架构引入计算机视觉领域，提出了视觉Transformer（vision Transformer，ViT）模型，通过其核心的自注意力机制，实现了对图像全局上下文的高效建模，从而弥补了卷积神经网络（CNN）在感受野方面的不足。此后，基于Transformer的融合方法相继涌现。Wang等人（2022）提出了一种基于Swin Transformer的红外与可见光图像融合网络SwinFuse，有效统一了红外图像的显著目标与可见光图像的丰富纹理。此外，Li等人（2024）提出了一种交叉注意机制融合网络，显著提升了红外与可见图像融合性能。然而，自注意力机制固有的二次方计算复杂度，使得Transformer模型在处理高分辨率图像时面临显著的计算瓶颈。

状态空间模型（state space models，SSM）的突破性进展（Gu和Dao，2023）为视觉任务提供了新范式。针对多模态图像融合中全局建模与计算效率难以兼顾的问题，Xie等人（2024）首次将改进的Mamba架构引入该领域，提出了动态特征增强融合方法FusionMamba，在多个任务上实现性能与效率的同步提升。Li等人（2024）提出了MambaDFuse模型，一种基于Mamba状态空间模型的双阶段框架，通过双级特征提取器和双阶段融合模块实现高效全局特征建模和跨模态互补信息整合。

上述方法在图像融合任务中均取得了较好的融合结果，但依然存在一些被忽略的问题。不同模态的源图像在特征分布和语义层级上存在显著差异，可见光图像主要侧重于宏观场景的全局语义表达，而红外图像则在细粒度语义信息的表达方面更具优势。现有融合方法普遍存在多层级语义信息的表征退化问题，且缺乏有效的跨层级特征交互机制，导致浅层细节与深层语义信息在融合过程中难以完全耦合。

针对上述问题，本文提出了一种基于Mamba的多层级红外与可见光图像融合网络，其主要贡献如下：1）构建了多层级特征融合网络框架，通过对多分辨率源图像进行全局特征建模与跨层级特征交互，实现了跨模态图像细粒度语义信息的有效保留；2）设计了F-Mamba模块，在仅具线性计算复杂度的前提下，精准捕获源图像长程依赖关系，提升了特征提取效率；3）针对不同层级间视觉—语义特征失配问题，设计了跨层级特征聚合模块（cross-level feature aggregation module，CFAM），在融合网络编码过程中，对不同层级的特征进行了有效增强；4）在多个公开数据集上的实验表明，本文所提方法在主观视觉效果和客观定量指标上均具有优越性。

1 融合方法

1.1 整体框架

本文提出了一种基于Mamba的多层级红外与可见光图像融合算法，旨在通过层次化特征提取与跨尺度信息融合提升图像融合任务的性能。

如图1所示，网络整体由特征提取、特征融合以及图像重建部分构成。由于红外图像与可见光图像在特征分布与语义信息上存在显著差异，融合网络的特征提取部分采用双分支结构独立提取源图像特征。为充分挖掘源图像的多分辨率信息，网络采用分层处理策略构建特征金字塔。

给定输入图像$\boldsymbol{I} \in \mathbf{R}^{H \times W \times C}$，首先通过降采样操作生成多尺度图像序列$\boldsymbol{I}_l (l = 1, 2, 3)$，其中第$l$层分辨率降为原始尺寸的$1/2^{l-1}$，采用可学习的卷积下采

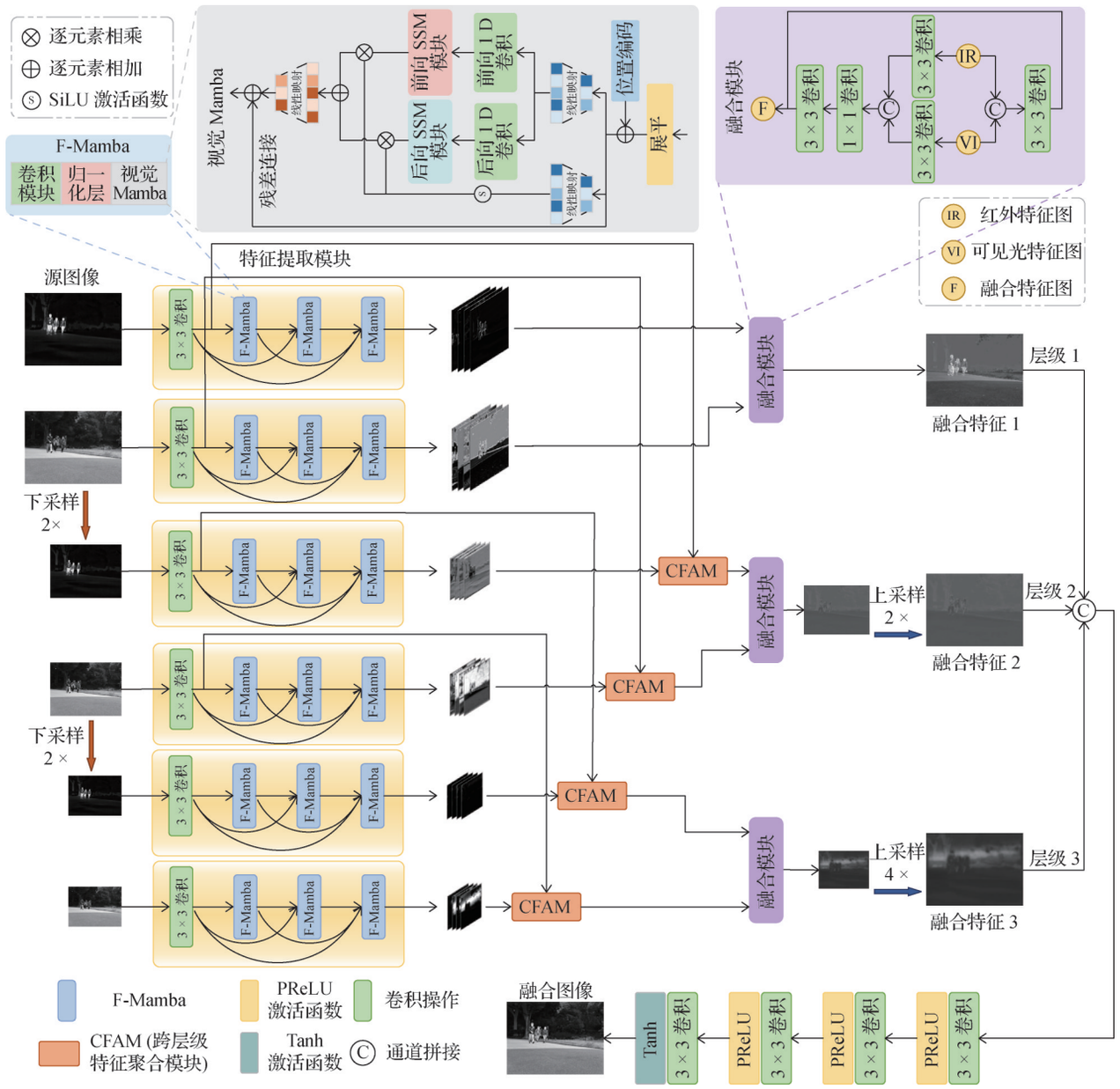

图 1　融合网络结构

Fig. 1　Structure of fusion network

样模块实现尺度变换，具体为

$$\boldsymbol{I}_l = D_l\left(\boldsymbol{I}_{l-1}\right) = C_{3\times3}^{s=2}\left(\boldsymbol{I}_{l-1}\right) \tag{1}$$

式中，$C_{3\times3}^{s=2}$ 表示步长为 2 的 3 × 3 卷积操作，D_l 表示第 l 级下采样操作。该设计在降低空间维度的同时保留局部结构信息，形成包含细粒度细节到粗粒度语义的多层次表征。

每个层级特征分别输入独立的分支网络进行处理，其中各分支共享相同的网络架构但具有独立参数，以此适应不同尺度的特征分布特性。各分支的特征提取过程采用统一架构，首先通过卷积层进行局部特征提取，计算过程为

$$\boldsymbol{F}_l^0 = C_{3\times3}\left(\boldsymbol{I}_l\right) \tag{2}$$

式中，$C_{3\times3}$ 表示 3 × 3 卷积操作。$\boldsymbol{F}_l^0 \in \mathbf{R}^{H\times W\times C}$ 表示第 l 层初始卷积特征。

随后，特征张量 $\boldsymbol{F}_l^0$ 依次通过 3 个 F-Mamba 模块进一步提取全局特征。同一特征层级内，3 个级联的 F-Mamba 模块通过密集连接将浅层细节特征与深层语义特征动态聚合。

对于第 l 层第 m 个 F-Mamba 模块的处理过程，其输出特征为

$$\boldsymbol{F}_l^m = \boldsymbol{F}_{\text{mamba}}\left(Concat\left(\boldsymbol{F}_l^0, \boldsymbol{F}_l^1, \cdots, \boldsymbol{F}_l^{m-1}\right)\right) \tag{3}$$

式中，$\boldsymbol{F}_{\text{mamba}}(\cdot)$表示F-Mamba模块特征提取过程，$\boldsymbol{F}_l^k(1 \leqslant k < m)$表示前序模块输出，$Concat(\cdot)$表示沿通道维度拼接。

本文所提出的融合网络通过跨层级特征聚合模块，实现了图像深层级语义信息的有效保留。具体而言，将高层级F-Mamba模块的输出特征和上一层级的初始卷积特征作为跨层级特征聚合模块的输入，从而得到该层级不同模态特征的最终输出，其计算过程为

$$\begin{cases} \boldsymbol{F}_l = \boldsymbol{F}_l^3 & l = 1 \\ \boldsymbol{F}_l = CFAM\left(\boldsymbol{F}_{l-1}^0, \boldsymbol{F}_l^3\right) & l = 2, 3 \end{cases} \tag{4}$$

式中，$\boldsymbol{F}_l$表示第l层最终输出特征，$\boldsymbol{F}_{l-1}^0$表示第$l-1$层初始卷积特征，$\boldsymbol{F}_l^3$表示第l层第3个F-Mamba模块输出特征，$CFAM(\cdot)$表示跨层级特征聚合操作。

在特征融合阶段，红外与可见光的特征图在3个不同层级上通过融合模块（fusion block）进行跨模态信息交互。首先将红外特征$\boldsymbol{F}_l^{ir} \in \mathbf{R}^{H \times W \times C}$与可见光特征$\boldsymbol{F}_l^{vi} \in \mathbf{R}^{H \times W \times C}$（其中$l$表示特征层级）分别输入参数独立的卷积层进行特征空间对齐，其计算过程为

$$\boldsymbol{F}_l^{ir'} = C_{3\times3}\left(\boldsymbol{F}_l^{ir}\right) \tag{5}$$

$$\boldsymbol{F}_l^{vi'} = C_{3\times3}\left(\boldsymbol{F}_l^{vi}\right) \tag{6}$$

随后，将双模态特征沿通道维度拼接，通过1×1卷积进行通道压缩，并将原始拼接特征以残差形式引入，其计算过程为

$$\boldsymbol{F}_l^{\text{concat}} = C_{1\times1}\left(Concat\left(\boldsymbol{F}_l^{ir'}, \boldsymbol{F}_l^{vi'}\right)\right) \tag{7}$$

$$\boldsymbol{F}_l^{\text{fuse}} = C_{3\times3}\left(\boldsymbol{F}_l^{\text{concat}}\right) + C_{3\times3}\left(Concat\left(\boldsymbol{F}_l^{ir}, \boldsymbol{F}_l^{vi}\right)\right) \tag{8}$$

式中，$C_{1\times1}$表示1×1卷积操作，$\boldsymbol{F}_l^{\text{fuse}} \in \mathbf{R}^{H \times W \times C}$表示不同层级的融合特征。不同层级的融合块输出的融合特征$\boldsymbol{F}_1^{\text{fuse}}$、$\boldsymbol{F}_2^{\text{fuse}}$、$\boldsymbol{F}_3^{\text{fuse}}$经过通道维度拼接后作为图像重建模块的输入。

图像重建模块包含4个级联卷积单元，每个卷积单元均由卷积层与非线性激活层构成，旨在将融合特征恢复为高质量的融合图像。

1.2 F-Mamba模块

基于Transformer的红外与可见光图像融合方法虽然能有效建模全局上下文信息，但其自注意力机制的二次计算复杂度严重制约了模型在高分辨率图像任务中的应用效率。为此，本文在网络特征提取阶段提出了F-Mamba模块，将卷积神经网络与视觉状态空间模型（SSM）相结合，通过F-Mamba模块的密集连接，在实现计算复杂度线性增长的同时保持全局特征提取能力。

如图1所示，输入特征图$\boldsymbol{X} \in \mathbf{R}^{H \times W \times C}$首先通过$3 \times 3$卷积层提取局部空间相关性，并经过批归一化（batch normalization，BN）层稳定特征分布，计算过程为

$$\boldsymbol{X}_{\text{loc}} = BN\left(C_{3\times3}(\boldsymbol{X})\right) \tag{9}$$

式中，$\boldsymbol{X}_{\text{loc}} \in \mathbf{R}^{H \times W \times C}$表示局部增强特征，$BN(\cdot)$表示批归一化。

随后，特征图经过展平操作转换为一维序列。然而，这一过程不可避免地会导致关键位置信息的丢失，而这些信息对于多模态图像融合任务至关重要。通过嵌入可学习位置编码，确保了在相同尺度的特征图中空间位置保持不变，计算过程为

$$\boldsymbol{S}_P = \boldsymbol{S} + \boldsymbol{P} \tag{10}$$

式中，$\boldsymbol{S}_P \in \mathbf{R}^{HW \times C}$表示位置编码特征，$\boldsymbol{S} \in \mathbf{R}^{HW \times C}$表示一维序列，$\boldsymbol{P} \in \mathbf{R}^{HW \times C}$表示可学习位置编码。

视觉Mamba单元采用双支路架构，其中主支路对位置编码特征进行线性投影后，通过一维卷积增强局部感知，再经过SiLU（sigmoid linear unit）激活函数得到视觉状态空间模型的输入特征，计算过程为

$$\boldsymbol{F}_{\text{SSM}} = SiLU\left(1\text{D}Conv\left(Linear\left(\boldsymbol{S}_p\right)\right)\right) \tag{11}$$

式中，$\boldsymbol{F}_{\text{SSM}}$表示双向状态空间模型的输入特征，$1\text{D}Conv(\cdot)$表示一维卷积操作，$Linear(\cdot)$表示线性投影。

如图2所示，双向状态空间模型采用双向扫描机制，通过前向扫描（行优先顺序）与后向扫描（逆向行序）对特征图进行空间遍历，将二维空间关系编码至隐状态空间，继而通过状态方程实现跨像素长程依赖的线性复杂度建模，同时避免传统注意力机制的高昂计算代价，前向SSM隐状态$\boldsymbol{h}_t^f \in \mathbf{R}^N$和后向SSM隐状态$\boldsymbol{h}_t^b \in \mathbf{R}^N$更新过程为

$$\begin{cases} \boldsymbol{h}_t^f = \bar{\boldsymbol{A}}^f \boldsymbol{h}_{t-1}^f + \bar{\boldsymbol{B}}^f \boldsymbol{x}_t \\ \boldsymbol{y}_t^f = \boldsymbol{C}^f \boldsymbol{h}_t^f + \boldsymbol{D}^f \boldsymbol{x}_t \end{cases} \tag{12}$$

$$\begin{cases} \boldsymbol{h}_t^b = \bar{\boldsymbol{A}}^b \boldsymbol{h}_{t-1}^b + \bar{\boldsymbol{B}}^b \boldsymbol{x}_t \\ \boldsymbol{y}_t^b = \boldsymbol{C}^b \boldsymbol{h}_t^b + \boldsymbol{D}^b \boldsymbol{x}_t \end{cases} \tag{13}$$

式中，$\bar{\boldsymbol{A}}^f$和$\bar{\boldsymbol{A}}^b \in \mathbf{R}^{N \times N}$表示状态转移矩阵，$\bar{\boldsymbol{B}}^f$和$\bar{\boldsymbol{B}}^b \in \mathbf{R}^{N \times 1}$表示输入投影矩阵，$\boldsymbol{C}^f$和$\boldsymbol{C}^b \in \mathbf{R}^{1 \times N}$表示输出投影矩阵，$\boldsymbol{D}$为直接传递矩阵，用于将输入直接映射到输出。所有参数均通过梯度下降优化，得到前向

SSM输出 $\boldsymbol{Y}^f \in \mathbf{R}^{HW \times C}$ 和后向SSM输出 $\boldsymbol{Y}^b \in \mathbf{R}^{HW \times C}$。

次支路则由线性投影层与SiLU激活函数构成。两分支输出经过逐元素乘积实现特征交互后，再通过线性变换层进行特征重组，计算过程为

$$\boldsymbol{F}_{\text{gate}} = SiLU\left(Linear\left(\boldsymbol{S}_p\right)\right) \tag{14}$$

$$\boldsymbol{Y}_{\text{out}} = Linear\left(\left(\boldsymbol{Y}^f \otimes \boldsymbol{F}_{\text{gate}}\right) + \left(\boldsymbol{Y}^b \otimes \boldsymbol{F}_{\text{gate}}\right)\right) \tag{15}$$

式中，$\boldsymbol{F}_{\text{gate}}$ 表示次支路的输出特征，$\boldsymbol{Y}_{\text{out}}$ 表示双向状态空间模型的输出特征，⊗表示逐元素相乘。

最终，经过双向状态空间模型处理的特征序列加入残差连接并通过维度重构操作恢复至原始特征图尺寸，得到F-Mamba模型的最终输出，计算过程为

$$\boldsymbol{X}_{\text{out}} = Reshape\left(\boldsymbol{Y}_{\text{out}} + \boldsymbol{S}_p\right) \tag{16}$$

式中，$\boldsymbol{X}_{\text{out}} \in \mathbf{R}^{H \times W \times C}$ 表示F-Mamba模型的输出特征，$\boldsymbol{S}_p$ 表示位置编码特征。

F-Mamba模块通过卷积神经网络的局部感知能力与视觉状态空间模型的线性复杂度全局建模特性，实现源图像特征精确提取的同时，显著降低了高分辨率图像处理的计算复杂度，其计算效率优势将通过消融实验进行验证。

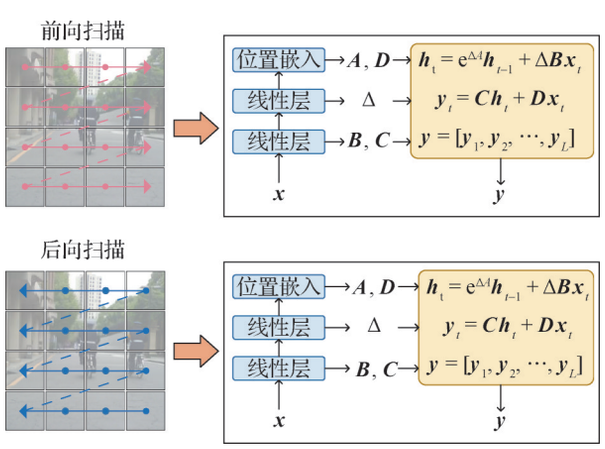

图2　双向状态空间模型示意图

Fig. 2　Schematic diagram of a bidirectional state space model

1.3　跨层级特征聚合模块

在图像融合任务中，单一尺度的特征提取存在显著局限性，难以同时捕获源图像的空间细节与语义信息，因此本文构建了多层级特征融合框架。低层级特征通常包含丰富的细节信息，视觉特征显著，而高层级特征往往包含细粒度的深层语义信息。然而，跨层级特征融合过程中存在视觉特征和语义特征难以完全耦合的问题。

针对上述问题，本文设计了跨层级特征聚合模块（CFAM），其核心思想通过分组交互机制以实现特征解耦与再耦合。

如图3所示，给定高层级特征 $\boldsymbol{F}_h \in \mathbf{R}^{H \times W \times C}$ 和经过双线性插值下采样对齐的低层级特征 $\boldsymbol{F}_l \in \mathbf{R}^{H \times W \times C}$，跨层级特征聚合模块首先沿通道维度将其划分为4组子特征 $\{\boldsymbol{F}_i^h, \boldsymbol{F}_i^l\}_{i=1}^4$。

每组特征通过通道拼接与层归一化实现初步聚合，计算过程为

$$\hat{\boldsymbol{F}}_i = LN\left(Concat\left(\boldsymbol{F}_i^h, \boldsymbol{F}_i^l\right)\right), i = 1, 2, 3, 4 \tag{17}$$

式中，$LN(\cdot)$ 表示层归一化。

为进一步增强聚合特征的表征能力，模块采用不同扩张率的扩张卷积对聚合特征进行深度提取，计算过程为

$$\tilde{\boldsymbol{F}}_i = DConv_{d_i}\left(\hat{\boldsymbol{F}}_i\right), d_i = 1, 3, 5, 7 \tag{18}$$

式中，$DConv_{d_i}(\cdot)$ 表示特征 $\hat{\boldsymbol{F}}_i \in \mathbf{R}^{H \times W \times C}$ 所对应的扩张率为 d_i 的3×3扩张卷积。这种设计通过渐进式扩张策略构建多级感受野，在保持参数量不变的前提下，有效捕获局部细节与全局上下文信息。

最终，4组特征通过通道拼接与1×1卷积得到输出特征，计算过程为

$$\boldsymbol{F}_{\text{out}} = C_{1\times1}\left(Concat\left(\tilde{\boldsymbol{F}}_1, \tilde{\boldsymbol{F}}_2, \tilde{\boldsymbol{F}}_3, \tilde{\boldsymbol{F}}_4\right)\right) \tag{19}$$

本文提出的融合网络架构中创新性地引入了跨层级特征聚合模块，通过多尺度特征通道的分组重构，构建了具有多粒度语义信息一致性的融合特征空间，实现了融合结果视觉与语义特征的深度对齐。

1.4　损失函数

在红外与可见光图像融合任务中，为了平衡多模态数据的互补特性并提升融合结果的视觉质量，本文通过结构相似度损失 L_{SSIM}、内容损失 L_{con} 及纹理损失 L_{texture} 的协同约束指导网络训练。

结构相似度损失旨在量化融合图像与源图像之间的全局结构相似性，其核心思想基于Wang等人（2004）提出的结构相似性指数（structural similarity index measure，SSIM）。

结构相似度损失 L_{SSIM} 通过结构相似性指数度量融合图像与源图像在亮度、对比度以及结构特征层面的匹配程度，具体为

$$L_{\text{SSIM}} = \frac{1}{2}\left(1 - S\left(\boldsymbol{I}_{\text{f}}, \boldsymbol{I}_{\text{ir}}\right)\right) + \frac{1}{2}\left(1 - S\left(\boldsymbol{I}_{\text{f}}, \boldsymbol{I}_{\text{vi}}\right)\right) \tag{20}$$

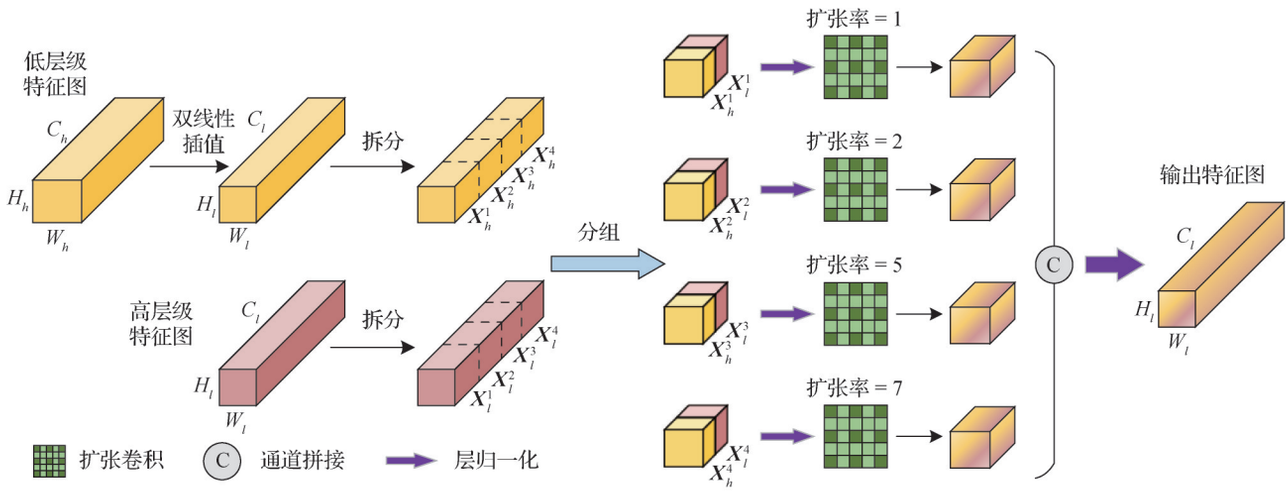

图3　跨层级特征聚合模块结构

Fig. 3　Structure of cross-level feature aggregation module

$$S(x,y)=\frac{(2\mu_x\mu_y+c_1)(2\delta_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\delta_x^2+\delta_y^2+c_2)} \tag{21}$$

式中，$\boldsymbol{I}_{\mathrm{f}}$、$\boldsymbol{I}_{\mathrm{ir}}$和$\boldsymbol{I}_{\mathrm{vi}}$表示融合图像、红外图像和可见光图像。$S(\cdot)$表示两幅图像之间SSIM值。$\mu_x$和$\mu_y$表示$x$和$y$的均值，$\delta_x$和$\delta_y$分别表示$x$和$y$的方差，$\delta_{xy}$表示$x$和$y$的协方差。

为约束融合结果与源图像在全局内容上的一致性，引入基于L1范数的内容损失函数。该损失直接衡量融合图像$\boldsymbol{I}_{\mathrm{f}}$与红外图像$\boldsymbol{I}_{\mathrm{ir}}$、可见光图像$\boldsymbol{I}_{\mathrm{vi}}$在像素空间的绝对差异。通过最小化像素级偏差，内容损失L_{con}能够有效抑制融合过程中的噪声干扰，同时保留红外图像的热辐射强度分布与可见光图像的亮度分布特性，具体为

$$L_{\mathrm{con}}=\frac{1}{HW}\left\|\left|\boldsymbol{I}_{\mathrm{f}}\right|-\max\left(\left|\boldsymbol{I}_{\mathrm{ir}}\right|,\left|\boldsymbol{I}_{\mathrm{vi}}\right|\right)\right\|_1 \tag{22}$$

式中，H和W分别表示图像的长和宽，$\|\cdot\|_1$表示L1范数，$\max(\cdot)$表示逐元素取最大值。

在红外与可见光图像融合任务中，通过最大化融合图像与源图像间的显著梯度响应，纹理损失L_{texture}可有效增强可见光图像中复杂场景的细节（如植被、建筑纹理）与红外图像中热目标的边缘锐度，具体为

$$L_{\mathrm{texture}}=\frac{1}{HW}\left\|\left|\nabla\boldsymbol{I}_{\mathrm{f}}\right|-\max\left(\left|\nabla\boldsymbol{I}_{\mathrm{ir}}\right|,\left|\nabla\boldsymbol{I}_{\mathrm{vi}}\right|\right)\right\|_1 \tag{23}$$

式中，∇表示Sobel梯度算子。

综上所述，本文网络训练过程中的总损失函数为$L_{\mathrm{total}}=\lambda_1L_{\mathrm{SSIM}}+\lambda_2L_{\mathrm{con}}+\lambda_3L_{\mathrm{texture}}$。其中，超参数$\lambda_1$、$\lambda_2$和$\lambda_3$分别设置为1、1和10。

2　实　验

2.1　实验设置

本文采用多光谱道路场景数据集MSRS（multi-spectral road scenarios）（Tang等，2022）作为基准数据集进行算法验证。该数据集包含1 444对高质量红外与可见光图像对，其空间分辨率均为480 × 640像素，采集环境涉及光照变化、天气变化和复杂背景等具有挑战性的场景。实验所使用的MSRS数据集被划分为训练集与测试集。训练集包含752对昼夜图像，测试集则包含361对图像。此外，还将LLVIP（visible-infrared paired dataset for low-light vision）数据集（Jia等，2021）的150对图像和RoadScene数据集的300对图像作为补充测试集，以验证融合网络的泛化能力。

所有实验均在配备NVIDIA L20 GPU和Intel Xeon Platinum 8457C CPU的硬件平台上完成。模型训练采用以下配置：使用Adam优化器，初始学习率固定为1×10^{-4}，batch_size为8，共训练80轮。所有定量评价指标均通过MATLAB 2023b计算获得。

对比实验中，本文选取了13种现有融合算法以验证本文方法的有效性，其中包括高雪琴等人（2020）提出的基于4阶偏微分方程的传统方法FPDE（fourth-order partial differential equation）以及12种基于深度学习的方法：GANMcC（generative adversarial network with multiclassification constraints）（Ma等，2020）、TarDAL（target-aware dual

adversarial learning)(Liu 等，2022)、IRFS(interactively reinforced fusion and saliency)(Wang 等，2023)、SwinFuse(residual swin Transformer fusion network)(Wang等,2022)、DATFuse(dual attention transformer)(Tang等,2023)、CrossFuse(Li和Wu,2024)、MBHFuse (multi-branch heterogeneous global and local infrared and visible image fusion)(Sun 等，2025)、U2Fusion(Xu 等,2022)、LRRNet(representation learning guided fusion network)(Li 等，2023)、MUFusion(general unsupervised image fusion network based on memory unit)(Cheng 等,2023)、DDBFusion(dual decomposition and bézier curves)(Zhang 等，2025)和 FusionMamba(dynamic feature enhancement for multimodal image fusion with Mamba)(Xie 等，2024)。

为客观评估融合图像质量,本文选取6种广泛使用的评价指标进行定量分析(孙彬 等,2023),分别是信息熵(entropy,EN)(Roberts等,2008)、空间频率(spatial frequency,SF)(Eskicioglu 和 Fisher,1995)、视觉保真度(visual information fidelity,VIF)(Han等,2013)、峰值信噪比(peak signal-to-noise ratio,PSNR)、平均梯度(average gradient,AG)(Cui 等，2015)和边缘强度(edge intensity,EI)(Xydeas和Petrović,2000)。

EN表征图像信息量的丰富程度,其值越大表明融合结果包含的纹理细节越丰富;SF反映图像灰度值的变化率,其数值越高表明图像边缘纹理越清晰;VIF基于人眼视觉系统特性,量化融合图像与源图像之间的信息保真度;PSNR通过最大信号强度与噪声强度的比值衡量图像重建质量,数值越高表示失真越小;AG评估图像局部特征的锐利程度,直接反映细节表达能力;EI则通过Sobel算子检测边缘能量分布,客观表征图像轮廓和结构特征的强度。

2.2 对比实验分析

2.2.1 MSRS数据集融合结果分析

如图4和图5所示,为评估模型在MSRS数据集上的融合性能,本文选取了典型场景进行主观视觉对比,其中“01008N”为夜晚场景。

如图4定性对比结果所示,除FPDE、MBHFuse、U2Fusion、LRRNet、DDBFusion及本文方法外,其余算法均未能有效保留人物面部细节特征。FPDE、U2Fusion和DDBFusion方法因整体亮度不足,导致蓝色矩形标注区域呈现明显的低照度特征,背景细节显著弱化。LRRNet方法在红外特征提取层面存在局限性,人物目标的热辐射信息呈现不完整,致使融合结果出现对比度不足的缺陷。值得关注的是,MBHFuse与本文方法在可见光背景区域(蓝色矩形标注区域)实现了与源图像一致的视觉保真度,但MBHFuse在人物轮廓边缘引入了明显的噪声干扰。

如图5定性对比结果所示,FPDE、TarDAL、IRFS、CrossFuse、MUFusion以及FusionMamba方法的融合结果在红色标注区域出现显著的特征退化现象,细节信息大量丢失,车牌号码难以辨认。GANMcC和DATFuse在该区域存在对比度失衡问题,导致车牌字符的轮廓信息部分丢失。与本文方法相比,SwinFuse、U2Fusion、LRRNet和DDBFusion方法由于整体亮度不足,蓝色标注区域中自行车的纹理细节信息存在不同程度的丢失。MBHFuse方法在红色标注区域存在过曝光现象,并在车牌边缘区域产生结构性伪影,导致图像视觉质量下降。

相较于现有13种主流融合算法,本文方法在图4与图5所示的典型昼夜场景中均能实现最优的细节保留与特征表达。这种优势主要得益于本文构建的基于Mamba多层级特征融合网络框架,通过跨层级特征聚合模块,在特征提取过程中实现多尺度空间特征与深层语义特征的有效整合。

为客观评估本文方法与主流算法的融合性能差异,本文选取6项评价指标对13种对比方法和本文方法进行定量比较。表1展示了各方法在MSRS数据集上的定量对比结果,该数据集包含361对测试样本,表中数值为全量测试数据的平均计算结果。

如表1所示,本文算法在EN、SF、VIF、PSNR、AG和EI这6项指标上均取得最优值。特别是在衡量信息保真度的VIF、PSNR以及反映边缘强度的EI等关键指标上,本文方法相较于次优模型优势明显。主观视觉分析与客观量化指标结果一致表明,本文所提融合算法通过多层级特征融合网络框架在保持源图像细节纹理的同时能够有效抑制噪声,显著提高融合图像质量,相较于现有主流方法具有优越性。

2.2.2 LLVIP数据集融合结果分析

为验证所提融合模型的泛化性能,本研究在LLVIP数据集“200292”和“190183”典型场景开展定

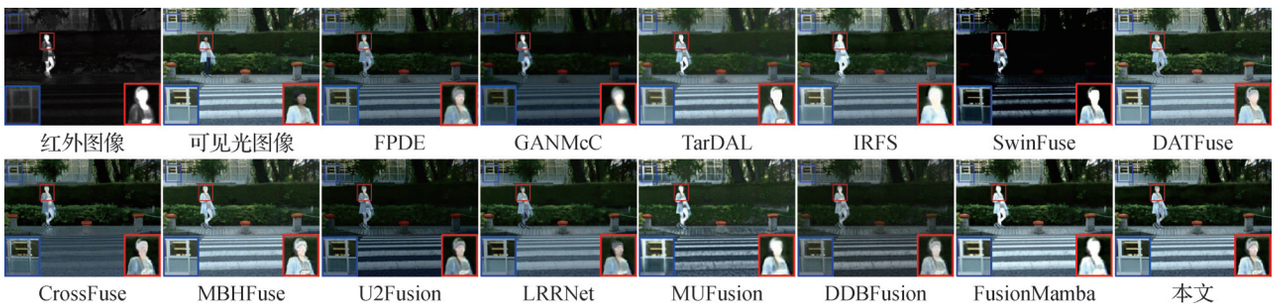

图4　MSRS数据集“00123D”融合结果视觉对比

Fig. 4　Visual comparison of fusion results on the MSRS dataset “00123D”

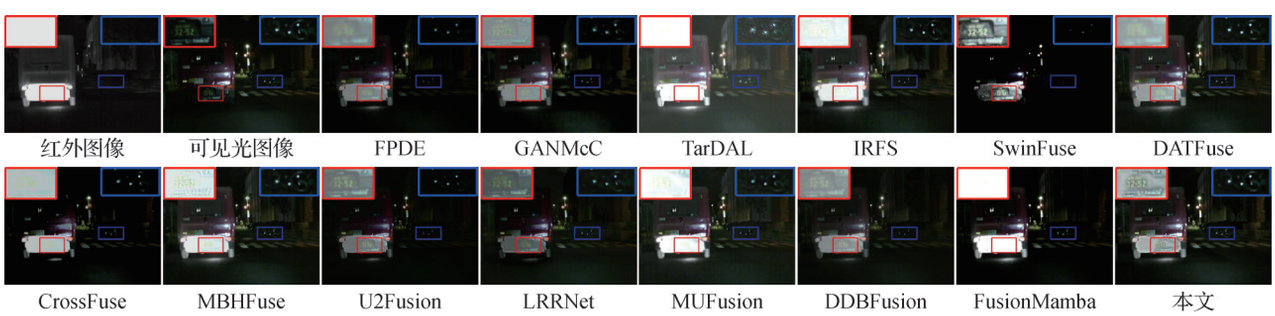

图5　MSRS数据集“01008N”融合结果视觉对比

Fig. 5　Visual comparison of fusion results on the MSRS dataset “01008N”

表1　MSRS数据集上不同融合算法的定量指标

Table 1　Quantitative metrics of fusion algorithms on the MSRS dataset

方法	EN	SF	VIF	PSNR/dB	AG	EI
FPDE	5.989 8	7.489 7	0.393 4	18.675 6	2.687 9	28.041 1
GANMcC	6.117 3	5.640 5	0.429 6	17.952 7	1.994 9	21.309 7
TarDAL	6.344 3	9.872 9	0.544 0	14.001 9	3.114 9	32.231 0
IRFS	6.484 7	9.118 2	0.611 5	15.892 8	2.918 2	30.591 1
SwinFuse	4.233 5	9.452 7	0.303 9	16.214 1	1.935 6	20.433 6
DATFuse	6.479 6	<u>10.916 9</u>	0.623 9	17.457 1	<u>3.573 9</u>	<u>37.406 3</u>
CrossFuse	5.048 0	10.361 7	0.368 8	17.696 3	2.937 3	31.066 1
MBHFuse	6.535 6	10.896 3	0.660 7	17.746 4	3.529 9	36.991 1
U2Fusion	5.039 6	7.225 0	0.333 0	18.507 5	2.262 1	24.112 1
LRRNet	6.191 5	8.454 7	0.414 7	18.334 3	2.639 7	28.004 4
MUFusion	5.963 9	8.759 7	0.618 5	15.726 8	3.117 6	34.231 9
DDBFusion	5.905 2	7.891 6	0.391 6	17.719 5	2.590 3	26.864 1
FusionMamba	<u>6.594 6</u>	8.893 7	<u>0.731 0</u>	<u>19.775 6</u>	3.126 8	34.261 0
本文	**6.733 4**	**11.087 5**	**0.765 7**	**20.032 3**	**3.667 0**	**39.695 4**

注：加粗、下划线字体表示各列最优、次优结果。

性对比实验。

如图6所示，在红色矩形标记区域中，GANMcC、DATFuse和CrossFuse方法生成的融合结果人物面部细节特征显著丢失。由于亮度不足，TarDAL和SwinFuse方法融合图像在蓝色矩形标注区域出现暗部汽车纹理信息完全缺失的问题。

MBHFuse方法在人物轮廓边缘处产生明显噪声，影响视觉质量。MUFusion方法因红外特征过度注入，造成人物面部细节特征丢失并伴随边缘模糊。相比于本文方法，IRFS方法对比度较低，路面细节纹理难以突出。尽管LRRNet、DDBFusion与本文方法在背景层面均取得良好的融合效果，但本文方法在人物面部及衣物等细节特征保留方面展现出显著优势。

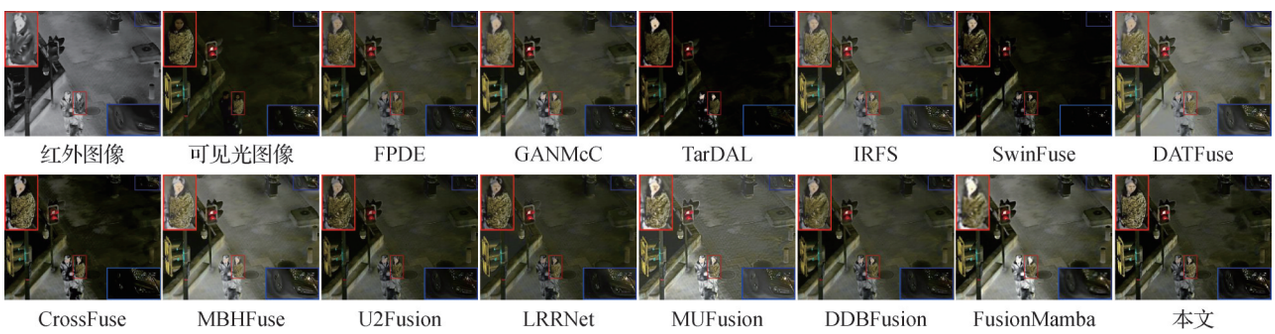

图6　LLVIP数据集“200292”融合结果视觉对比

Fig. 6　Visual comparison of fusion results on the LLVIP dataset “200292”

如图7所示，红色矩形标记区域中FPDE、GANMcC、DATFuse以及CrossFuse方法的融合结果均出现了人物面部细节缺失，特征信息丢失严重。TarDAL和SwinFuse方法融合结果存在亮度失衡问题，导致衣物褶皱和边缘轮廓缺失并出现伪影。MUFusion方法多模态信息保留失衡，使得人物面部局部泛白且边缘模糊。MBHFuse方法融合结果人物眉眼及发丝区域引入明显噪声。相比之下，本文方法在该区域不仅完整保留了人物的面部轮廓和衣物褶皱纹理，而且面部表情层次分明，细节清晰度显著提升。在蓝色矩形标记区域中，受过度曝光环境的影响，SwinFuse、DATFuse以及FusionMamba方法汽车车灯纹理信息几乎完全丢失。相较之下，本文方法通过多层级特征提取策略，在恢复源图像矩形标记区域细节的同时，还有效保留了路面砖块的纹理特征，使整个场景视觉效果更为自然。

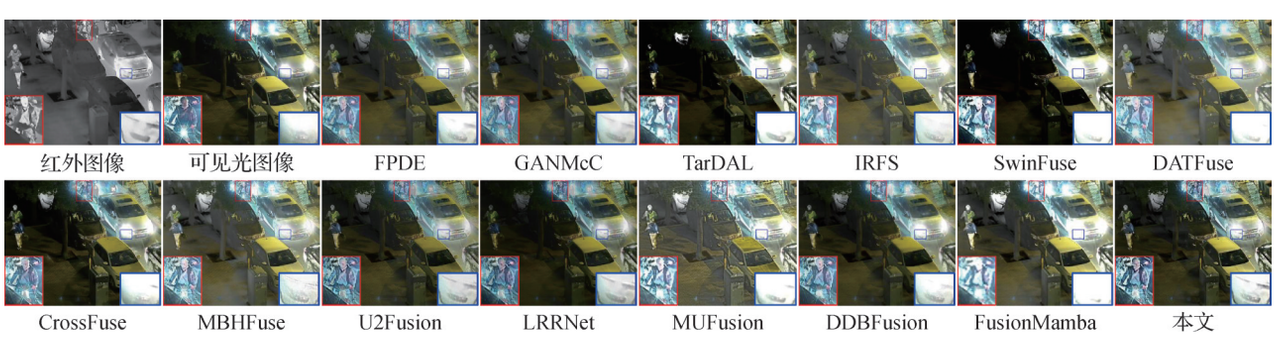

图7　LLVIP数据集“190183”融合结果视觉对比

Fig. 7　Visual comparison of fusion results on the LLVIP dataset “190183”

基于LLVIP测试集的客观评价结果如表2所示。本文方法在SF、PSNR、AG和EI上表现最佳，VIF也达到次优水平。而MBHFuse算法由于融合结果中产生的轮廓噪声，其EN值偏高。上述定量结果验证了本文方法能够有效生成目标特征显著、轮廓结构清晰的融合图像。

2.2.3　RoadScene数据集融合结果分析

为全面评估所提融合模型的泛化能力，本文基于RoadScene数据集中的“05548”典型过曝光场景开展了定性对比实验，融合对比结果如图8所示。除FPDE、GANMcC、U2Fusion、LRRNet及本文方法外，其余方法在过曝光场景中红色标定区域的车辆尾部及车牌轮廓均难以清晰辨识。GANMcC与LRRNet的融合结果因亮度偏低，导致背景细节存在一定程度的丢失。FPDE和U2Fusion方法的融合结果中可见光图像的强光照信息被红外图像的轮廓信

表2　LLVIP数据集上不同融合算法的定量指标

Table 2　Quantitative metrics of fusion algorithms on the LLVIP dataset

方法	EN	SF	VIF	PSNR/dB	AG	EI
FPDE	6.682 2	8.146 3	0.381 2	<u>17.694 1</u>	2.582 1	26.178 6
GANMcC	6.682 4	6.063 7	0.381 8	17.273 4	1.806 0	18.866 0
TarDAL	6.341 8	6.416 5	0.300 5	15.150 2	1.955 9	20.335 8
IRFS	6.910 7	10.662 3	0.513 0	17.208 0	2.723 2	28.149 6
SwinFuse	6.589 7	10.499 4	0.524 9	15.165 5	2.664 1	27.353 5
DATFuse	6.902 8	8.683 0	0.368 4	15.898 6	2.041 7	21.039 2
CrossFuse	6.759 2	12.275 9	0.421 8	16.836 8	3.279 8	33.576 8
MBHFuse	<u>7.198 8</u>	<u>13.737 4</u>	0.601 8	15.839 7	<u>3.726 3</u>	<u>37.944 4</u>
U2Fusion	6.640 3	8.389 1	0.426 8	17.477 5	2.618 7	27.283 6
LRRNet	6.315 8	8.800 4	0.318 9	17.668 1	2.326 1	23.657 0
MUFusion	6.830 1	8.926 5	**0.651 0**	15.190 1	3.172 1	34.315 8
DDBFusion	6.691 3	9.391 4	0.389 3	17.279 4	2.689 4	26.975 0
FusionMamba	**7.276 6**	5.218 9	0.563 9	17.082 7	1.919 3	21.300 3
本文	6.916 7	**14.619 4**	<u>0.602 0</u>	**17.773 2**	**3.967 5**	**40.688 2**

注：加粗、下划线字体表示各列最优、次优结果。

图8　RoadScene数据集“05548”融合结果视觉对比

Fig. 8　Visual comparison of fusion results on the RoadScene dataset “05548”

息所抑制，未能有效凸显场景的典型特征。相比之下，本文所提方法在过曝光条件下不仅显著突出了关键目标的轮廓特征，同时更好地保留了场景信息。

在RoadScene数据集上对比方法及本文方法的定量指标如表3所示。由于RoadScene数据集中图像原始尺寸不一致，为满足模型输入要求并保证公平比较，所有图像在测试时均统一调整至固定尺寸。融合完成后，为与原始输入进行一致比较，利用线性插值将融合结果上采样至原尺寸，一定程度上影响了部分指标的表现。尽管如此，本文方法在AG和EI两项指标上仍取得最优值，在EN上也达到次优水平。定量分析结果表明，所提出的融合方法具有良好的跨场景泛化能力。基于多个数据集的对比实验结果，验证了本文算法在融合性能上的优越性，同时也表明该模型具有较强的泛化能力。本文所提网络模型构建了基于Mamba架构多层级特征融合网络框架，提高模型计算效率的同时，实现了对跨模态图像细粒度语义信息的有效保留。

2.3　消融实验分析

本节针对融合网络中F-Mamba模块、多层级特征融合网络框架以及跨层级特征聚合模块进行了消融实验研究。消融实验基于MSRS数据集的完整测试集进行，且所有对比实验均保持完全一致的参数

表3　RoadScene数据集上不同融合算法的定量指标

Table 3　Quantitative metrics of fusion algorithms on the RoadScene dataset

方法	EN	SF	VIF	PSNR/dB	AG	EI
FPDE	6.900 0	14.218 3	0.341 4	16.342 3	5.905 6	60.478 4
GANMcC	7.236 7	9.018 7	0.454 6	13.377 7	3.778 4	40.446 2
TarDAL	7.257 4	10.422 6	0.432 5	15.474 8	4.167 9	45.023 9
IRFS	7.000 3	10.238 0	0.414 8	**16.404 5**	3.900 0	40.978 5
SwinFuse	**7.509 8**	16.090 5	0.693 8	14.304 8	6.031 0	63.018 8
DATFuse	6.709 8	11.394 4	0.242 1	15.781 5	4.031 1	41.056 4
CrossFuse	7.306 2	14.150 1	0.315 2	13.491 6	5.061 8	53.187 7
MBHFuse	7.068 5	**16.762 8**	0.355 3	15.393 7	5.822 5	58.216 9
U2Fusion	7.262 2	15.067 3	0.550 0	15.769 6	6.159 9	65.102 9
LRRNet	7.131 5	12.373 8	0.377 4	12.187 0	4.640 0	48.657 1
MUFusion	7.340 7	13.406 1	**0.754 5**	14.866 6	6.093 3	66.891 5
DDBFusion	6.969 5	13.520 8	0.491 1	14.895 5	5.155 6	53.329 6
FusionMamba	7.015 6	14.407 1	0.634 4	13.899 9	5.577 1	59.843 5
本文	7.491 2	15.129 7	0.519 8	14.814 4	**6.378 2**	**68.214 0**

注：加粗、下划线字体表示各列最优、次优结果。

设置。通过图9定性结果与表4定量指标的综合分析，验证各模块对融合性能的贡献。

多层级特征融合网络框架的消融实验（w/o multi-level）将本文方法与仅保留最高层级融合特征的网络框架进行对比，以验证多层级网络架构在特征提取和深层语义保留方面的优势。如图9（c）所示，与完整模型相比，单层级特征融合框架的融合结果整体亮度不足，细粒度语义特征丢失。

为验证F-Mamba模块的贡献，F-Mamba模块的消融实验（w/o F-Mamba）将该模块替换为卷积层，而保持网络其余部分不变。如图9（d）所示，完整模型在红色矩形区的细节还原能力明显更强。这是F-Mamba模块特有的全局感受野，使其在捕捉跨区域关联性方面优于局部操作的卷积层，从而增强了对源图像细节信息的表征能力。

跨层级特征聚合模块的消融实验（w/o CFAM）去除网络架构中跨层级特征聚合模块，其余网络结构均保持不变，重新训练后进行测试。如图9（e）所示，由于特征提取过程中缺乏跨层级特征交互机制，导致融合过程中不同层级视觉特征与语义特征难以完全耦合，融合结果在红色矩形标注区域中部分纹理细节缺失。

如表4所示，本文方法通过引入多层级特征融合网络框架，实现了源图像细粒度语义信息的有效保留。因此，相比于w/o multi-level模型，完整模型在EN、VIF和EI指标上均有大幅度提升。定量对比表明，F-Mamba模块在提升融合质量的同时，显著优化了模型的计算效率。此外，相较于w/o CFAM模

（a）红外图像　（b）可见光图像　（c）w/o F-Mamba　（d）w/o multi-level　（e）w/o CFAM　（f）本文

图9　消融实验结果视觉对比

Fig. 9　Visual comparison of ablation experiments results

（（a）infrared image；（b）visible image；（c）w/o F-Mamba；（d）w/o multi-level；（e）w/o CFAM；（f）ours）

表4　消融实验定量指标

Table 4　Quantitative metrics of ablation experiments

方法	EN	SF	VIF	PSNR/dB	AG	EI
w/o multi-level	6.078 3	10.106 1	0.652 1	14.374 6	3.159 2	33.716 1
w/o F-Mamba	6.690 8	10.721 5	0.739 6	15.299 9	3.498 6	37.870 3
w/o CFAM	6.607 2	10.525 1	0.707 9	18.702 4	3.348 6	36.016 1
本文	**6.733 4**	**11.087 5**	**0.765 7**	**20.032 3**	**3.667 0**	**39.695 4**

注：加粗、下划线字体表示各列最优、次优结果。

型，本文方法在VIF和AG指标上的显著提升，验证了跨层级特征聚合模块（CFAM）在整合与增强不同层级特征方面的有效性。

为量化F-Mamba模块的效率优势，本文通过构建对比基准（base Transformer）：将本文模型中的F-Mamba模块替换为vision Transformer模块，系统评估了本文方法与对比模型在不同分辨率输入下的浮点运算量（floating point number operations，FLOPs）和推理时间（inference time），实验结果如表5所示。不同分辨率的测试结果表明，两种模型的计算效率呈现出显著差异。当输入分辨率从32×32像素增至128×128像素时，本文模型的计算量仅从0.34 G线性增长至5.52 G，而base Transformer模型架构的计算量则显著增加，在分辨率为128×128像素时即达到20.20 G，接近本文模型架构计算量4倍。输入图像分辨率提高时，本文模型的推理时间增长平缓，而base Transformer模型的推理时间大幅增加。

结合图10的融合结果分析，引入F-Mamba模块，既能有效缓解传统Transformer模型随分辨率平方增长的计算负担，又能通过状态空间模型的序列建模优势增强全局特征的表征能力，在细节纹理保留方面具有明显优势。

综上所述，通过对消融实验的全面分析，定性与定量结果证明了本文提出的各个模块在融合性能提升方面的有效性。

2.4　运行效率分析

为系统评估所提出方法的计算效率，本研究将其与当前主流的红外与可见光图像融合方法在计算量（FLOPs）和参数量（parameters）方面进行了对比实验。所有测试均在统一硬件环境（NVIDIA GeForce RTX 3090 GPU）下进行，输入图像尺寸为512×512像素。

表5　本文方法与对比模型计算量和推理时间比较

Table 5　Comparison of FLOPs and inference time between the proposed method and baseline module

方法	输入分辨率/像素	FLOPs/G	推理时间/ms
base Mamba	32×32	0.34	25.23
base Transformer	32×32	0.40	22.78
base Mamba	64×64	1.38	25.50
base Transformer	64×64	3.10	26.17
base Mamba	128×128	5.52	26.22
base Transformer	128×128	20.20	55.97

图10　本文方法与基于Transformer模型融合结果对比

Fig. 10　Comparison of the fusion result between the proposed method and module based on Transformer

如表6所示，在计算复杂度方面，本文方法的FLOPs为88.27 G，显著低于多数基于Transformer的方法，如SwinFuse、MBHFuse和DDBFusion，同时也低于同为Mamba架构的FusionMamba。在参数量方面，本文方法参数量相比于FusionMamba方法具有显著优势，并优于部分基于Transformer的方法（如CrossFuse、DDBFusion）。本文方法运行效率优势主要源于本文所采用的多层级特征提取与融合策略。通过构建特征金字塔并采用可学习的卷积下采样模块，在保留多尺度信息的同时有效控制了特征图的空间尺寸，从而降低了后续模块的计算负担。此外，F-Mamba模块通过选择性状态空间建模机制，在保持全局感受野的同时避免了Transformer中自注意力机制的二次复杂度，进一步提升了计算效率。

综上所述，本文方法在维持优异融合性能的基础上，显著提升了模型的运行效率，不仅缓解了基于Transformer方法常见的高计算复杂度问题，同时实现了参数规模的有效控制，为红外与可见光图像融合任务提供了一种高效且轻量化的解决方案。

2.5 语义分割实验分析

由于语义分割任务更多地关注融合图像中的高层语义信息，其性能表现可有效反映图像融合结果对语义信息的保持能力。本文选用MSRS数据集进行实验分析，该数据集同时包含可见光与红外模态图像数据，并提供9类语义分割标注。实验通过计算融合图像在语义分割任务中的平均交并比（mean intersection over union，mIoU）指标，定量评估不同融合方法对后续语义理解任务的支撑能力。

定性实验结果如图11所示，除FPDE和本文方法，其余方法的融合结果均未能分出红色矩形标注区域中curve。此外，本文方法在红色矩形标注区域中路肩的分割结果最接近Ground Truth的分割结果。定量实验结果如表7所示，本文方法在background、bike和bump类别中均取得交并比最优值，在curve类别取得次优值，同时在mIoU指标上为最优值。

表 6　不同融合方法的模型计算量和参数量比较

Table 6　Comparison of FLOPs and parameter for different fusion methods

方法	FLOPs/G	参数量/M
GANMcC	4.48	3.91
TarDAL	77.77	0.30
IRFS	62.92	0.24
SwinFuse	176.33	0.34
DATFuse	4.74	0.01
CrossFuse	20.81	1.61
MBHFuse	115.69	0.30
U2Fusion	172.68	0.66
LRRNet	12.09	0.05
MUFusion	48.134	0.56
DDBFusion	967.68	3.67
FusionMamba	105.90	8.55
本文	88.27	1.11

综上所述，在MSRS数据集上的语义分割实验充分表明，本文方法通过多层级融合网络框架实现了跨模态图像细粒度语义信息的有效保留，在下游任务中具有显著优势。

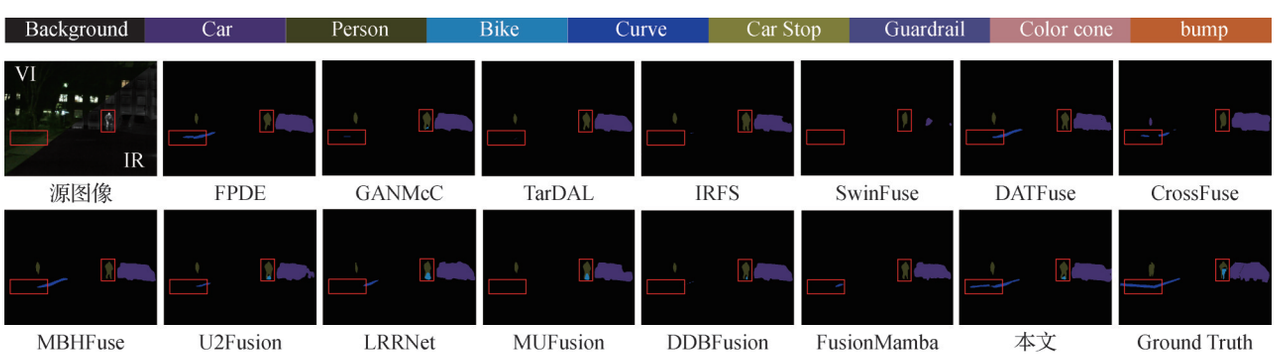

图 11　不同融合方法语义分割结果视觉对比

Fig. 11　Visual comparison of semantic segmentation results from different fusion methods

3 结　论

本文提出了一种基于Mamba的多层级红外与可见光图像融合算法，通过构建多层级特征融合框架，显著增强了模型深层细粒度语义信息的保留能力。此外，本文在特征提取阶段创新性引入密集连接F-Mamba模块，在有效捕获长程依赖特征的同时实现了网络结构的轻量化设计。由于不同层级红外与可见光图像所包含的视觉特征和语义特征存在差异，通过本文设计的跨层级特征聚合模块，可以充分整合源图像的互补特征，有效解决不同层级融合结果视觉信息与语义信息的表征失配问题。实验结果表明，本文算法有效改善了现有方法（尤其在夜间等复杂场景下）难以保留源图像细粒度语义信息以及基于Transformer的融合网络参数冗余、训练困难的

表7　不同融合方法语义分割结果定量指标

Table 7　Quantitative metrics of the semantic segmentation results from different fusion methods

方法	Background	Car	Person	Bike	Curve	Car Stop	Guardrail	Color cone	bump	mIoU
FPDE	0.985 2	0.911 0	0.735 5	0.715 0	0.585 3	0.751 4	0.813 7	0.631 4	0.717 6	0.760 7
GANMcC	0.985 1	0.908 2	0.738 6	0.701 4	0.573 1	0.763 8	0.844 7	0.645 4	0.720 5	0.764 5
TarDAL	0.985 2	0.908 1	0.743 8	0.710 0	0.573 0	0.743 7	**0.861 5**	0.634 9	0.722 0	0.764 7
IRFS	0.985 3	0.911 1	0.735 8	0.713 4	0.552 4	**0.767 3**	0.809 7	0.657 9	0.744 2	0.764 1
SwinFuse	0.970 6	0.811 5	0.474 6	0.536 8	0.199 8	0.531 8	0.655 4	0.384 1	0.501 6	0.562 9
DATFuse	0.986 0	0.913 1	0.746 0	0.712 8	0.605 4	0.752 7	0.814 3	0.661 1	0.784 4	0.775 1
CrossFuse	0.980 9	0.890 2	0.668 4	0.676 4	0.415 5	0.731 7	0.762 9	0.571 8	0.484 4	0.686 9
MBHFuse	0.986 2	**0.915 4**	**0.749 8**	0.716 4	**0.624 4**	0.765 6	0.841 0	**0.663 8**	0.731 1	0.777 1
U2Fusion	0.984 9	0.904 8	0.743 6	0.710 5	0.547 0	0.742 2	0.806 7	0.634 3	0.714 3	0.754 2
LRRNet	0.984 6	0.909 3	0.709 9	0.679 3	0.506 7	0.754 8	0.833 4	0.634 5	0.793 9	0.756 3
MUFusion	0.985 0	0.907 1	0.733 7	0.707 2	0.595 4	0.759 4	0.834 0	0.647 0	0.678 5	0.760 8
DDBFusion	0.984 5	0.908 0	0.724 8	0.704 9	0.534 6	0.741 7	0.858 9	0.630 3	0.693 2	0.753 4
FusionMamba	0.979 7	0.868 0	0.657 8	0.642 5	0.467 7	0.657 5	0.804 3	0.474 5	0.618 5	0.685 6
本文	**0.986 3**	0.909 9	0.745 1	**0.717 3**	0.616 5	0.755 1	0.856 9	0.641 7	**0.794 1**	**0.780 3**

注：加粗、下划线字体表示各列最优、次优结果。

问题。在MSRS和LLVIP数据集上的对比实验表明，与其他13种现有融合方法相比，本文所提方法在提升视觉效果以及行人、车辆等关键目标的边缘特征保持方面展现出显著优势。

本文提出的方法虽然在特征表征以及计算效率方面取得显著提升，但仍存在亟待改进之处：当前框架未能建立与分割或检测网络的级联优化机制，导致特征融合过程缺乏高层次语义先验的引导，可能影响关键目标区域的完整性表征，同时限制了融合结果在高层视觉任务中的迁移能力。未来工作拟构建基于多任务协同的级联优化框架，通过引入高层视觉任务网络，并将其结果作为约束嵌入融合过程，使网络在保留纹理细节的同时，强化目标区域的语义一致性。

参考文献（References）

Amarsaikhan D, Saandar M, Ganzorig M, Blotevogel H H, Egshiglen E, Gantuyal R, et al. 2012. Comparison of multisource image fusion methods and land cover classification. International Journal of Remote Sensing, 33(8): 2532-2550 [DOI: 10.1080/01431161.2011.616552]

Chen M S. 2016. Image fusion of visual and infrared image based on NSCT and compressed sensing. Journal of Image and Graphics, 21(1): 39-44 (陈木生. 2016. 结合NSCT和压缩感知的红外与可见光图像融合. 中国图象图形学报, 21(1): 39-44) [DOI: 10.11834/jig.20160105]

Cheng C Y, Xu T Y and Wu X J. 2023. MUFusion: a general unsupervised image fusion network based on memory unit. Information Fusion, 92: 80-92 [DOI: 10.1016/j.inffus.2022.11.010]

Cui G M, Feng H J, Xu Z H, Li Q and Chen Y T. 2015. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. Optics Communications, 341: 199-209 [DOI: 10.1016/j.optcom.2014.12.032]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. 2020. An image is worth 16 × 16 words: Transformers for image recognition at scale [EB/OL]. [2025-06-11]. https://arxiv.org/pdf/2010.11929.pdf

Eskicioglu A M and Fisher P S. 1995. Image quality measures and their performance. IEEE Transactions on Communications, 43(12): 2959-2965 [DOI: 10.1109/26.477498]

Gao X Q, Liu G, Xiao G, Bavirisetti D P and Shi K L. 2020. Fusion algorithm of infrared and visible images based on FPDE. Acta Automatica Sinica, 46(4): 796-804 (高雪琴, 刘刚, 肖刚, Bavirisetti D P, 史凯磊. 2020. 基于FPDE的红外与可见光图像融合算法. 自动化学报, 46(4): 796-804) [DOI: 10.16383/j.aas.2018.c180188]

Gong R and Wang X C. 2019. Infrared and visible image fusion based on BEMD and W-transform. Journal of Image and Graphics, 24(6):

987-999（宫睿，王小春. 2019. BEMD分解和W变换相结合的红外与可见光图像融合. 中国图象图形学报，24(6)：987-999）［DOI：10.11834/jig.180530］

Gu A and Dao T. 2024. Mamba：linear-time sequence modeling with selective state spaces//Proceedings of the 1st Conference on Language Modeling. Philadelphia，USA：COLM：1-56

Han Y，Cai Y Z，Cao Y and Xu X M. 2013. A new image fusion performance metric based on visual information fidelity. Information Fusion，14(2)：127-135［DOI：10.1016/j.inffus.2011.08.002］

Jia X Y，Zhu C，Li M Z，Tang W Q and Zhou W L. 2021. LLVIP：a visible-infrared paired dataset for low-light vision//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal，Canada：IEEE：3489-3497［DOI：10.1109/ICCVW54120.2021.00389］

Jin W Q，Li L and Wang X. 2023. Research and application of thermal imaging mode and image processing technology. Acta Optica Sinica，43(15)：#1510001（金伟其，李力，王霞. 2023. 热成像模式及其图像处理技术的研究与应用. 光学学报，43(15)：#1510001［DOI：10.3788/AOS230740］

Li G L，Xiang W H，Zhang S L and Zhang B X. 2022. Infrared and visible image fusion algorithm based on residual network and attention mechanism. Unmanned Systems Technology，5(2)：9-21（李国梁，向文豪，张顺利，张博勋. 2022. 基于残差网络和注意力机制的红外与可见光图像融合算法. 无人系统技术，5(2)：9-21）［DOI：10.19942/j.issn.2096-5915.2022.2.012］

Li H and Wu X J. 2019. DenseFuse：a fusion approach to infrared and visible images. IEEE Transactions on Image Processing，28(5)：2614-2623［DOI：10.1109/TIP.2018.2887342］

Li H and Wu X J. 2024. CrossFuse：a novel cross attention mechanism based infrared and visible image fusion approach. Information Fusion，103：#102147［DOI：10.1016/j.inffus.2023.102147］

Li H，Xu T Y，Wu X J，Lu J W and Kittler J. 2023. LRRNet：a novel representation learning guided fusion network for infrared and visible images. IEEE Transactions on Pattern Analysis and Machine Intelligence，45(9)：11040-11052［DOI：10.1109/TPAMI.2023.3268209］

Li Z，Pan H，Zhang K，Duan Y，Yu K and Gong M. 2024. Mambadfuse：a Mamba-based dual-phase model for multi-modality image fusion［EB/OL］.［2025-06-11］. https://arxiv.org/pdf/2404.08406.pdf

Liu J Y，Fan X，Huang Z B，Wu G Y，Liu R S，Zhong W，et al. 2022. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans，USA：IEEE：5792-5801［DOI：10.1109/CVPR52688.2022.00571］

Ma J Y，Zhang H，Shao Z F，Liang P W and Xu H. 2021. GANMcC：a generative adversarial network with multiclassification constraints for infrared and visible image fusion. IEEE Transactions on Instrumentation and Measurement，70：#5005014［DOI：10.1109/TIM.2020.3038013］

Roberts J W，Van Aardt J A and Ahmed F B. 2008. Assessment of image fusion procedures using entropy，image quality，and multispectral classification. Journal of Applied Remote Sensing，2(1)：#023522［DOI：10.1117/1.2945910］

Sun B，Gao Y X，Zhuge W W and Wang Z X. 2023. Analysis of quality objective assessment metrics for visible and infrared image fusion. Journal of Image and Graphics，28(1)：144-155（孙彬，高云翔，诸葛吴为，王梓萱. 2023. 可见光与红外图像融合质量评价指标分析. 中国图象图形学报，28(1)：144-155）［DOI：10.11834/jig.210719］

Sun Y C，Dong M L，Yu M X and Zhu L Q. 2025. MBHFuse：a multi-branch heterogeneous global and local infrared and visible image fusion with differential convolutional amplification features. Optics and Laser Technology，181：#111666［DOI：10.1016/j.optlastec.2024.111666］

Tang L F，Yuan J T，Zhang H，Jiang X Y and Ma J Y. 2022. PIAFusion：a progressive infrared and visible image fusion network based on illumination aware. Information Fusion，83-84：79-92［DOI：10.1016/j.inffus.2022.03.007］

Tang W，He F Z，Liu Y，Duan Y S and Si T Z. 2023. DATFuse：infrared and visible image fusion via dual attention transformer. IEEE Transactions on Circuits and Systems for Video Technology，33(7)：3159-3172［DOI：10.1109/TCSVT.2023.3234340］

Wang D，Liu J Y，Liu R S and Fan X. 2023. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection. Information Fusion，98：#101828［DOI：10.1016/j.inffus.2023.101828］

Wang Z，Bovik A C，Sheikh H R and Simoncelli E P. 2004. Image quality assessment：from error visibility to structural similarity. IEEE Transactions on Image Processing，13(4)：600-612［DOI：10.1109/TIP.2003.819861］

Wang Z S，Chen Y L，Shao W Y，Li H and Zhang L. 2022. SwinFuse：a residual Swin transformer fusion network for infrared and visible images. IEEE Transactions on Instrumentation and Measurement，71：#5016412［DOI：10.1109/TIM.2022.3191664］

Xie X Y，Cui Y W，Tan T，Zheng X B and Yu Z T. 2024. FusionMamba：dynamic feature enhancement for multimodal image fusion with Mamba. Visual Intelligence，2(1)：#37［DOI：10.1007/s44267-024-00072-9］

Xu H，Ma J Y，Jiang J J，Guo X J and Ling H B. 2022. U2Fusion：a unified unsupervised image fusion network. IEEE Transactions on Pattern Analysis and Machine Intelligence，44(1)：502-518［DOI：10.1109/TPAMI.2020.3012548］

Xydeas C S and Petrović V. 2000. Objective image fusion performance measure. Electronics Letters，36(4)：308-309［DOI：10.1049/el：20000267］

Yang P，Gao L F and Zi L L. 2021. Image fusion method of convolution

sparsity and detail saliency map analysis. Journal of Image and Graphics, 26(10): 2433-2449 (杨培, 高雷阜, 訾玲玲. 2021. 卷积稀疏与细节显著图解析的图像融合. 中国图象图形学报, 26(10): 2433-2449) [DOI: 10.11834/jig.200205]

Zhang B H, Lu X Q, Pei H Q and Zhao Y. 2015. A fusion algorithm for infrared and visible images based on saliency analysis and non-subsampled Shearlet transform. Infrared Physics and Technology, 73: 286-297 [DOI: 10.1016/j.infrared.2015.10.004]

Zhang Z Y, Li H, Xu T Y, Wu X J and Kittler J. 2025. DDBFusion: an unified image decomposition and fusion framework based on dual decomposition and Bézier curves. Information Fusion, 114: #102655 [DOI: 10.1016/j.inffus.2024.102655]

作者简介

杨天宇,男,硕士研究生,主要研究方向为图像处理与模式识别。E-mail: 2023211474@stu.ppsuc.edu.cn

霍宏涛,通信作者,男,教授,博士生导师,主要研究方向为图像处理与模式识别、遥感应用技术、图像取证。E-mail: huohongtao@ppsuc.edu.cn

郭宝峰,男,博士研究生,主要研究方向为模式识别和图像融合。E-mail: 2019111024@ppsuc.edu.cn

郑博文,男,硕士研究生,主要研究方向为图像处理和目标识别。E-mail: 2022211499@stu.ppsuc.edu

刘晓文,男,博士研究生,主要研究方向为模式识别和计算机视觉。E-mail: liu_x_wen@163.com