智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于多尺度协调卷积与自适应加权的红外与可见光图像融合

刘诗怡, 刘金平, 黄丽娟, 蒋嘉豪, 宋殿义, 杨广益

引用本文:
刘诗怡, 刘金平, 黄丽娟, 等. 基于多尺度协调卷积与自适应加权的红外与可见光图像融合[J]. 智能系统学报, 2026, 21(1): 95-108.
LIU Shiyi, LIU Jinping, HUANG Lijuan, et al. Infrared and visible image fusion based on multi-scale coordinated convolution and adaptive weighting[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(1): 95-108.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202504002

DOI: 10.11992/tis.202504002
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20250903.1715.002

基于多尺度协调卷积与自适应加权的红外与可见光图像融合

刘诗怡[1], 刘金平[1], 黄丽娟[2], 蒋嘉豪[1], 宋殿义[3], 杨广益[4]
（1. 湖南师范大学 信息科学与工程学院, 湖南 长沙 410081; 2. 湖南省智能康复机器人与辅助设备工程技术研究中心, 湖南 长沙 410004; 3. 国防科技大学 军政基础教育学院, 湖南 长沙 410072; 4. 湖南省计量检测院, 湖南 长沙 410081）

摘　要: 针对当前基于卷积神经网络的图像融合模型在全局信息感知、高频细节保持及损失函数权重设定上的局限性, 提出一种集成卷积和多层感知器架构的多尺度协调网络, 以实现红外与可见光图像的高质量融合。提出一种卷积加权重排多层感知器模块, 通过模拟特征排列增强空间维度理解, 并结合自适应特征重加权机制有效整合全局信息。同时, 提出多尺度协调卷积模块, 利用中心差分卷积增强高频信息的保留能力, 并通过多尺度并行子网络优化多层次特征表达; 其内嵌的坐标注意力机制, 通过通道−空间联合调制强化互补信息并抑制冗余特征。此外, 还提出一种数据驱动的自适应权重策略, 基于图像特征统计量动态调整监督信号的贡献度, 降低调参复杂性并提升损失函数的自适应性。在 RoadScene、TNO 和 M^3FD 这 3 个公开数据集上的实验结果表明, 本文算法生成的融合图像在边缘保持、纹理过渡方面表现更优, 且在信息熵、标准差、空间频率、视觉信息保真度和平均梯度等指标上全面超越主流融合方法, 为红外与可见光图像融合提供了新的思路, 为图像融合领域的进一步发展打下了坚实的基础。

关键词: 图像融合; 红外图像; 可见光图像; 多尺度协调卷积; 卷积加权重排多层感知器; 坐标注意力; 自适应权重
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1673−4785(2026)01−0095−14

中文引用格式: 刘诗怡, 刘金平, 黄丽娟, 等. 基于多尺度协调卷积与自适应加权的红外与可见光图像融合 [J]. 智能系统学报, 2026, 21(1): 95–108.
英文引用格式: LIU Shiyi, LIU Jinping, HUANG Lijuan, et al. Infrared and visible image fusion based on multi-scale coordinated convolution and adaptive weighting[J]. CAAI transactions on intelligent systems, 2026, 21(1): 95–108.

Infrared and visible image fusion based on multi-scale coordinated convolution and adaptive weighting

LIU Shiyi[1], LIU Jinping[1], HUANG Lijuan[2], JIANG Jiahao[1], SONG Dianyi[3], YANG Guangyi[4]
(1. College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China; 2. Hunan Intelligent Rehabilitation Robot and Auxiliary Equipment Engineering Technology Research Center, Changsha 410004, China; 3. Basic Education College, National University of Defense Technology, Changsha 410072, China; 4. Hunan Institute of Metrology and Testing, Changsha 410081, China)

Abstract: To address the limitations of convolution neural networks-based image fusion models, such as restricted global information perception, high-frequency detail preservation, and the loss function weights configuration, this article proposes a convolution and multilayer perceptron-integrated multiscale coordinate network (CM-MCNet) for high-quality infrared and visible image fusion. In the encoder of CM-McNet, a convolutional weighted permute multilayer perceptron module is introduced to enhance spatial understanding by simulating feature permutation and integrates an adaptive feature reweighting mechanism to effectively capture global information. Meanwhile, a multiscale coordinate convolution (MsCConv) module is designed, leveraging the advantages of central difference convolution to enhance the retention and expression of high-frequency details. By incorporating multiscale parallel sub-networks, MsCConv ensures the comprehensive preservation of multi-level features. Moreover, the embedded coordinate attention mechanism jointly modulates channel and spatial dimensions, enhancing complementary information while suppressing redundancy. Furthermore, a data-driven adaptive loss weighting strategy is proposed, which can dynamically adjust the contribution of supervision signals based on image feature statistics. This reduces the complexity of hyperparameter tuning while ensuring the loss function more accurately reflects the characteristics of the source images. Experimental results on the RoadScene, TNO, and M^3FD public datasets demonstrate that CM-MCNet generates fused images with sharper edge preservation and more natural texture transitions. Additionally, our method achieves superior performance across various objective metrics, including information entropy, standard deviation, spatial frequency, visual information fidelity, and average gradient, outperforming existing state-of-the-art fusion methods. This work provides a novel perspective for infrared and visible image fusion and lays a solid foundation for further advancements in the field.
Keywords: image fusion; infrared image; visible image; multiscale coordinate convolution; convolutional multilayer perceptron; coordinate attention; adaptive weighting

收稿日期: 2025−04−01.　**网络出版日期**: 2025−09−04.
基金项目: 国家自然科学基金项目 (62371187); 湖南省自然科学基金项目 (2024JJ8309).
通信作者: 刘金平. E-mail: ljp@hunnu.edu.cn.

红外图像具备光照不变性, 适用于弱光和遮挡场景, 但细节表达能力较弱; 可见光图像则包含丰富的纹理和边缘信息, 但受光照影响较大。

红外和可见光图像融合 (infrared and visible image fusion, IVIF), 可整合来自不同传感器的成像信息, 有效克服单一模态的局限, 提升下游任务的鲁棒性与准确性[1]。当前, IVIF 在语义分割[2]、目标检测[3]、无人机导航[4]、图像复原[5] 等领域展现出广泛应用价值。

近年来, 基于卷积神经网络 (convolution neural network, CNN) 的 IVIF 方法成为主流, 大多采用编码器−解码器结构[6-8]。典型工作包括: Zhao 等[9] 提出的深度图像分解模型 (deep image decomposition for infrared and visible image fusion, DIDFuse), 通过图像分解实现特征级融合。Liang 等[10] 提出的基于自监督表示学习的图像分解模型 (DeFusion), 利用自监督学习在特征嵌入空间中分离共有与独特特征以提升融合效果。Liu 等[11] 提出一种基于显著性检测的图像融合框架 (saliency guided deep-learning framework for pixel-level image fusion, SGFusion), 采用双解码器处理多模态图像和多曝光图像。然而, 这些方法仍面临三大挑战:

1) CNN 受限于其相对狭窄的感受野, 难以捕获全局上下文信息以生成高质量的融合图像[12-13]。为捕获长程依赖, Tang 等[14] 提出了一种 Y 形动态 Transformer(YDTR), 其 dynamic Transformer 模块兼顾局部与全局特征提取。此外, Tang 等[15] 进一步提出了一种基于双注意力 Transformer 的 IVIF 模型 (infrared and visible image fusion via dual attention transformer, DATFuse), 通过并联 CNN 与 Transformer 以整合局部与全局特征。但由于自注意力机制的弱归纳偏置问题, 该类融合模型在捕捉局部特征方面的能力受限。

2) 前向传播过程中高频信息的损失会导致生成图像的细节与边缘模糊, 进而影响后续的计算机视觉任务[16]。为了有效保留细节信息, Ma 等[16] 提出了一种用于 IVIF 的生成对抗网络 (FusionGAN), 利用对抗学习实现了源图像优势特征的融合与细节信息的增强。Liu 等[17] 提出了一种注意力引导的全局−局部对抗学习网络 (attention-guided global-local adversarial learning, AGAL), 结合注意力引导和边缘损失优化融合质量。然而, 过度强调高频信息可能引入冗余, 影响全局一致性。

3) 当前的融合模型普遍依赖于手动/经验调整损失函数中的权衡参数, 以整合不同感知源的图像特征, 导致训练复杂度高且融合均衡性难以保证。为解决上述问题, Cheng 等[18] 提出了一种基于结构相似性变量的自适应权重机制。Liu 等[1] 则结合强度与结构相似性动态调整权重。然而, 上述方法仅考虑了源图像的比例重要性, 未充分挖掘各模态源图像自身的独特特征, 且难以适应于单一模态的损失项。

针对上述挑战, 本文提出一种集成卷积和多层感知器的多尺度协调红外与可见光图像融合网络 (convolution and multilayer perceptron-integrated multiscale coordinate network, CM-MCNet), 主要贡献可以总结如下:

1) 提出了卷积加权重排多层感知器 (convolutional weighted permute multilayer perceptron, CWPMLP) 模块, 通过模拟排列操作增强模型对空间维度的理解, 并利用重新加权模块确保模型能够根据全局特征的重要性来动态调整特征组合, 实现了全局上下文信息的有效整合。

2) 设计了多尺度协调卷积 (multiscale coordinate convolution, MsCConv) 模块, 结合中心差分卷积增强融合图像对高频信息的保留能力, 并通过高效多尺度注意力机制 (efficient multi-scale attention, EMA)[19] 与坐标注意力机制 (coordinate attention, CA)[20] 的并联架构, 实现了对多尺度信息的聚合以及对关键细节的进一步聚焦。

3) 引入数据驱动自适应权重设置策略, 通过计算信息保留度及各损失函数的优化程度以动态调整损失权重, 降低手动参数化设置的需求, 提高源图像和融合结果之间的细节一致性。

1　相关工作

1.1　基于编码−解码的 IVIF 模型

基于编码−解码 (Enc-Dec) 架构的 IVIF 模型构建一种通用且高效的融合框架, 结构如图 1 所示。

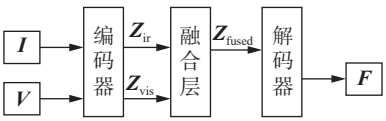


图 1　基于 Enc-Dec 的 IVIF 框架

Fig. 1　IVIF framework based on Enc-Dec

模型在编码阶段输入红外图像 $\boldsymbol{I}$ 与可见光图像 $\boldsymbol{V}$, 通过编码器进行特征处理, 生成能捕捉原始图像中关键信息与结构细节的特征表示图: $\boldsymbol{Z}_{\text{ir}}$ 和 $\boldsymbol{Z}_{\text{vis}}$。在特征融合阶段, 采用加权平均、特征拼接或基于注意力机制的动态权重调整等融合策略, 将 $\boldsymbol{Z}_{\text{ir}}$ 和 $\boldsymbol{Z}_{\text{vis}}$ 进行深度融合获得融合特征图 $\boldsymbol{Z}_{\text{fused}}$。在解码阶段, 使用解码器将融合后的特征表示 $\boldsymbol{Z}_{\text{fused}}$ 重建为高质量的融合图像 $\boldsymbol{F}$, 确保融合结果兼具丰富的细节与优良的视觉质量。尽管该架构在提升融合质量方面表现优异, 仍存在以下挑战: 1) 传统方法在局部细节与全局结构两者间权衡不足; 2) 损失函数权重参数依赖经验设置, 调优复杂, 易影响融合一致性与模型稳定性。

1.2　EMA 与 CA

注意力机制在深度学习与计算机视觉领域中发挥着关键作用, 尤其在多模态图像融合任务中, 通过自适应筛选关键特征, 有效增强了模型的特征表达能力与融合质量。因此, 在多模态融合架构中合理引入注意力机制成为提升性能的关键途径。

1) EMA

为避免传统融合方法中通用卷积降维导致的信息损失, Ouyang 等[19] 提出了如图 2 所示的 EMA 机制。

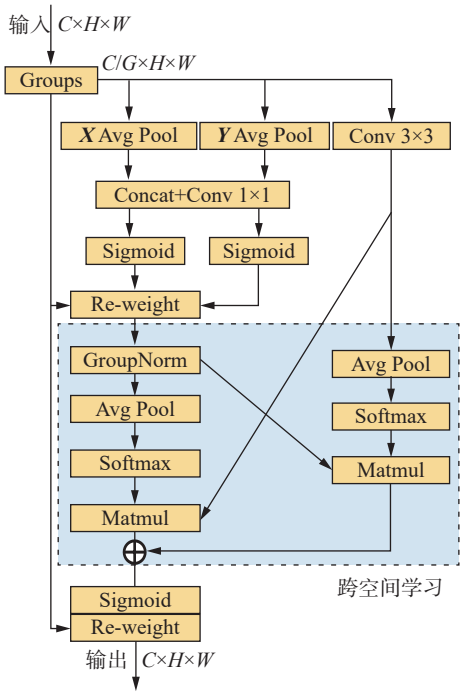


图 2　EMA 结构

Fig. 2　Architecture of EMA

EMA 将输入特征图 $\boldsymbol{X}\in\mathbf{R}^{H\times W\times C}$ 分为 G 个子特征组, 每个组学习不同的语义, 以加强语义区域的特征学习, 并有效抑制噪声。本文中关于 G 的设置参考了文献 [19] 中提出的经验原则 $G\ll C$, 并依据其在 ResNet-50 / CIFAR-100 上的消融实验选取 G=32 为默认值。其特征提取模块包含 3 条并行路径: 两条采用 1×1 卷积结合一维全局平均池化, 分别在水平与垂直方向进行通道编码; 第 3 条路径采用 3×3 卷积以获取多尺度特征, 提升跨通道信息的表达能力。该多尺度并行结构可有效编码通道间相关性, 并通过空间聚合策略增强融合特征的完整性。3 条路径的输出通过矩阵点积进行融合, 生成初始空间注意力图。随后, 在每个子特征组内采用 Sigmoid 激活计算注意力权重, 并对特征图加权聚合, 实现像素级关联建模与全局上下文整合, 从而提升融合质量。总体而言, EMA 在保留通道信息的同时增强了多层次特征表达能力。但需注意, 并行子网络在处理相似特征时可能引入信息冗余, 影响融合效果与计算效率。

2) CA

为增强模型定位和识别目标对象的能力, Hou 等[20] 提出了如图 3 所示的 CA 机制。

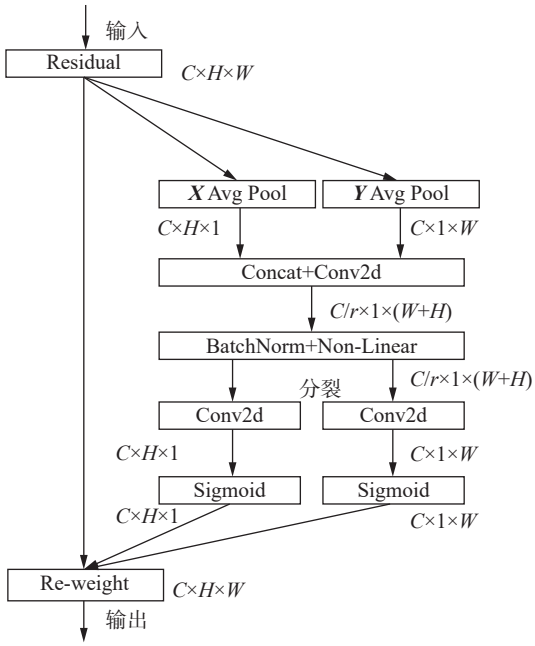


图 3　CA 结构图

Fig. 3　Architecture of CA

CA 对传统的全局平均池化操作的基础上, 引入水平−垂直分解策略, 以精确捕获目标对象的位置:

$$\boldsymbol{Z}_c^h(h)=\frac{1}{W}\sum_{0\leqslant i<W}\boldsymbol{X}_c(h,i)$$

$$\boldsymbol{Z}_c^w(w)=\frac{1}{H}\sum_{0\leqslant j<H}\boldsymbol{X}_c(j,w)$$

式中: $\boldsymbol{X}_c$ 为第 c 通道的输入, $\boldsymbol{Z}_c^h(h)$ 和 $\boldsymbol{Z}_c^w(w)$ 分别为高度 h 处和宽度 w 的第 c 通道的输出。

通过共享的 1 × 1 卷积变换函数 F_1, 对 $\boldsymbol{Z}_c^h(h)$ 和 $\boldsymbol{Z}_c^w(w)$ 进行联合变换得 $\boldsymbol{F}\in\mathbf{R}^{C/r\times(H+W)\times1}$:

$$\boldsymbol{F}=\delta\left(F_1\left(\left[\boldsymbol{Z}^h,\boldsymbol{Z}^w\right]\right)\right)$$

式中: $[\cdot,\cdot]$ 是空间维度拼接操作, δ 是非线性激活函数。

将 F 沿着空间维度进行分裂操作, 分为 $\boldsymbol{F}^h\in\mathbf{R}^{C/r\times H\times1}$ 和 $\boldsymbol{F}^w\in\mathbf{R}^{C/r\times1\times W}$, 再利用 1×1 卷积进行升维度操作, 并结合 Sigmoid 激活函数得到最后的注意力向量 $\boldsymbol{G}^h\in\mathbf{R}^{C\times H\times1}$ 和 $\boldsymbol{G}^w\in\mathbf{R}^{C\times1\times W}$, 即

$$\boldsymbol{G}^w=\sigma(F_w(\boldsymbol{F}^w))$$

$$\boldsymbol{G}^h=\sigma\left(F_h\left(\boldsymbol{F}^h\right)\right)$$

CA 的最终输出可表示为

$$\boldsymbol{Y}_c(i,j)=\boldsymbol{X}_c(i,j)\times\boldsymbol{G}_c^h(i)\times\boldsymbol{G}_c^w(j)$$

CA 将水平与垂直注意力映射分别施加到输入特征图上，使每个像素点能够感知其所在行列的全局信息，从而实现精准目标定位。

为缓解融合网络在前向传播中高频信息易丢失的问题，本文提出 MsCConv 模块。该模块引入中心差分卷积以增强对细节与纹理变化的捕捉能力，并结合 EMA 机制将部分通道重构为批处理维度，引入多尺度并行子网络以提升多层次特征的保留能力。为降低 EMA 中并行路径可能带来的冗余，引入并联的 CA 模块，进一步优化对关键区域的响应能力，提升细节聚焦效果。

2　所提出的 CM-MCNet

2.1　网络架构

DIDFuse[9] 作为典型的基于编码−解码 (Enc-Dec) 结构的图像融合方法，在 IVIF 任务中展现出优异性能。该方法通过优化激活函数与网络层配置，有效提升了融合图像的信息保真度与视觉质量，能够同时呈现高亮目标与细节纹理。

受其启发，本文在编码−解码框架基础上构建 CM-MCNet，核心组件包括 CWPMLP 与 MsCConv。CM-MCNet 整体结构如图 4 所示。

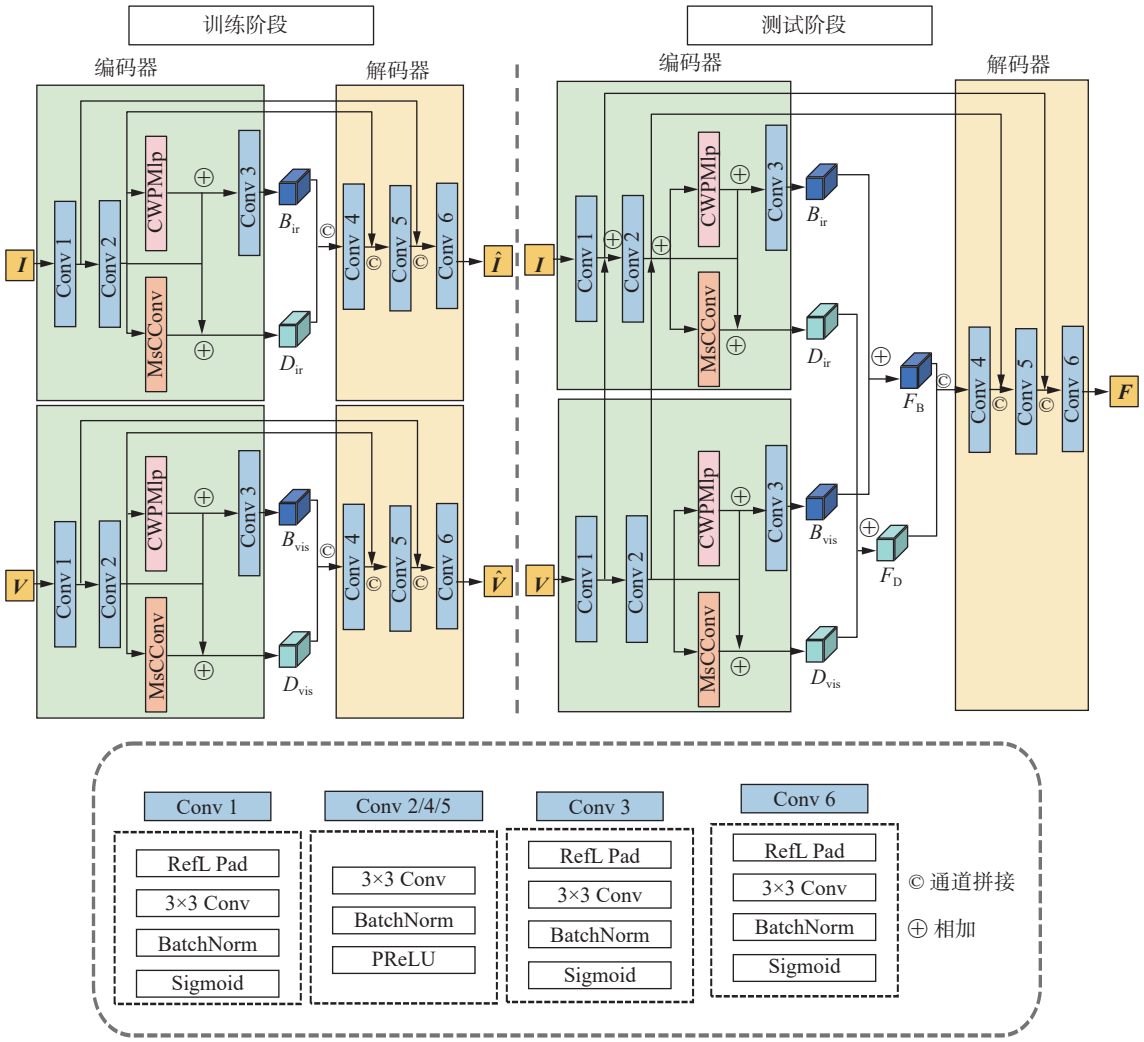


图 4　CM-MCNet 模型结构

Fig. 4　Architecture of CM-MCNet

训练阶段，编码器接收红外图像 $\boldsymbol{I}$ 或可见光图像 $\boldsymbol{V}$，提取背景与细节特征，并在通道维度上将得到的细节和背景特征拼接融合；解码器负责将融合特征重构为图像 $\hat{\boldsymbol{I}}$ 或 $\hat{\boldsymbol{V}}$。为缓解多层卷积导致的细节损失并提升训练效率，解码器在最后两层卷积操作中引入编码初期生成的特征图，实现特征复用。同时，利用多尺度跳跃连接融合不同层级特征，协同高层语义与底层细节，实现多尺度精细重构。

在编码器中，CWPMLP 模块通过模拟排列操作增强空间表征能力，并引入动态重加权机制更好地整合全局上下文信息。MsCConv 则利用中心差分卷积提升对高频信息的感知能力，并结合多尺度特征与关键细节聚合机制，强化对局部与全局细节的联合建模，从而显著提升特征质量。此外，训练过程引入数据驱动的自适应损失权重

策略, 以自动调节各子损失的权重, 有效降低人工调参成本, 同时挖掘源图像潜在结构, 为模型优化提供更精确的指导。

在测试阶段, 红外与可见光图像分别输入预训练编码器, 提取出的背景与细节特征通过元素级加法完成融合, 随后输入解码器进行图像重建。解码器利用反卷积层恢复图像的像素与细节, 并通过 Sigmoid 函数归一化输出, 确保像素值合理分布。最终融合图像兼具红外图像的热辐射信息与可见光图像的细节分辨能力, 实现对多模态信息的有效集成与增强, 呈现更加完整与清晰的视觉效果。

2.2　CWPMLP 模块

为在保持局部信息的同时, 增强模型对全局上下文的理解, 本文设计了 CWPMLP 模块。该模块结构如图 5 所示。不同于传统的 MLP 类模型[21], CWPMLP 以三维令牌表示作为输入。CWPMLP 由 3 个分支组成, 分别专注于编码图像沿高度、宽度以及通道维度的信息。

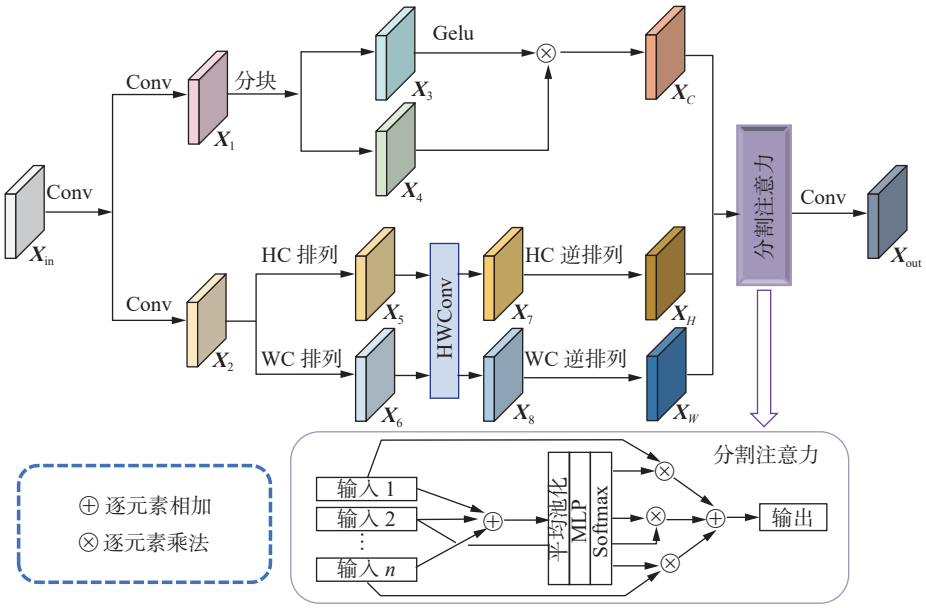


图 5　CWPMLP 结构图

Fig. 5　Architecture of CWPMLP

在通道维度上, 采用深度可分离卷积和激活函数相结合的策略来编码信息得到通道维度特征图 $\boldsymbol{X}_C$, 以提高计算效率并增强通道维度信息的非线性表示能力。$\boldsymbol{X}_C$ 可表示为

$$\boldsymbol{X}_C = \mathrm{GELU}(\boldsymbol{X}_4) \otimes \boldsymbol{X}_5 \tag{1}$$

式中 GELU 为高斯误差线性单元激活函数。为有效编码沿高度维度的空间信息, 本文首先在高度分支上进行模拟排列操作, 具体步骤如下: 1) 对给定输入 $\boldsymbol{X}_2 \in \mathbf{R}^{H\times W\times C}$ 进行高度通道 (HC) 排列操作: 首先沿着通道维度分为 S 段, 得到 $\bar{\boldsymbol{X}}_2 \in \mathbf{R}^{H\times W\times S\times N}$。其中 N 满足 $C = N\times S$。对每个片段沿高度与通道维度进行排列重组, 得到 $\bar{\boldsymbol{X}}_2^{\mathrm{T}} \in \mathbf{R}^{S\times W\times H\times N}$。最后沿着通道维度进行拼接, 输出 $\boldsymbol{X}_6 \in \mathbf{R}^{S\times W\times (H\times N)}$。2) 利用共享卷积层 HW 卷积对排列后的张量进行线性变换得到 $\boldsymbol{X}_5 \in \mathbf{R}^{S\times W\times (H\times N)}$。3) 对变换后的张量执行 HC 逆排列操作, 输出高度维度特征图 $\boldsymbol{X}_H \in \mathbf{R}^{H\times W\times C}$。类似地, 在宽度分支中执行与上述相同的操作, 得到宽度维度特征图 $\boldsymbol{X}_W \in \mathbf{R}^{H\times W\times C}$ 以编码沿宽度维度的空间信息。通过模拟排列操作, CWPMLP 模块中的高度与宽度分支分别增强了模型对空间维度的建模能力。该操作通过重新排列特征张量, 引入非局部依赖关系, 从而有效捕捉图像中的长程语义信息。相比传统 MLP 先展平空间维度后进行线性投影而导致位置信息丢失, CWPMLP 分别沿高度与宽度方向建模, 保留了更丰富的位置信息。此外, 为确保模型能够根据全局特征的重要性动态调整特征组合, 并克服自注意力机制弱归纳偏差造成细节纹理丢失, 将分割注意力[22] 分别应用于 $\boldsymbol{X}_C$、$\boldsymbol{X}_H$ 和 $\boldsymbol{X}_W$, 重新校准不同分支的重要性, 从而更好地整合全局上下文信息。最后, 通过卷积层对 3 个分支的输出进行融合, 得到最终输出 $\boldsymbol{X}_{\mathrm{out}} \in \boldsymbol{R}^{H\times W\times C}$。综上所述, CWPMLP 模块通过多层次的特征处理和动态权重调整, 显著加强了局部特征的表达, 并促进了模型对全局上下文信息的整合, 使得模型能够学习到更全面、有效的特征表示。

2.3　MsCConv 模块

为了增强融合图像在纹理、边缘和高频信息的保留和表达, 本文设计了 MsCConv 模块, 其结构如图 6 所示。

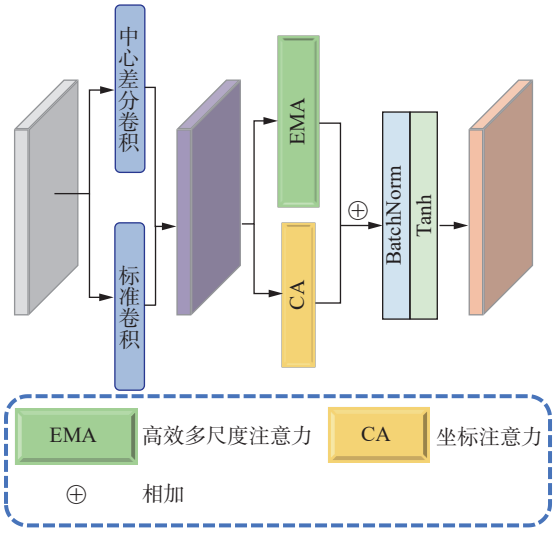


图 6　MsCConv 结构图

Fig. 6　Architecture of MsCConv

MsCConv 结合中心差分卷积与标准卷积的加权和，并采用包含 EMA 模块和 CA 模块的并联架构。具体流程如下：首先，对输入到 MsCConv 的图像进行采样处理，提取出多个图像块。然后，将采样获得的图像块中的每个元素减去中心元素，这一过程称为中心差分。接着，通过普通卷积操作得到输出结果。如果进行了中心差分步骤，则该过程被视为中心差分卷积；若未进行此步骤，则视为标准卷积。上述卷积过程可表示为

$$y(p_0)=\theta\sum_{p_n\in R}w(p_n)\cdot(x(p_0+p_n)-x(p_0))+(1-\theta)\cdot\sum_{p_n\in R}w(p_n)\cdot x(p_0+p_n) \tag{2}$$

式中：p_0为中心元素位置，p_n表示p_0的邻核，θ表示控制着差分卷积以及标准卷积对最终输出的影响程度，w代表权重。

在 IVIF 任务中，标准卷积与差分卷积各具优势但也存在一定局限性。标准卷积通过直接聚合局部邻域信息，能够有效提取图像的基础结构特征。然而，其对光照变化较为敏感，泛化能力相对较弱，且在精细结构建模方面存在一定不足，难以准确捕捉目标边缘等关键细节。相比之下，差分卷积通过对相邻像素间的差异建模，在提取边缘与纹理等高频特征方面表现出显著优势。其生成的空间差分特征对光照变化具有更强的鲁棒性，能够在不同环境条件下保持稳定的特征表达能力，并更好地保留图像中的细粒度结构信息，对提升融合图像质量具有重要意义。因此，本文采用中心差分卷积协同利用标准卷积与差分卷积的优势，在保留全局上下文信息的同时，增强对局部细节的理解，以提供更丰富的特征表示。

为增强融合图像高频信息的保留和表达能力，同时兼顾多尺度特征提取，本文设计了一个由 EMA 和 CA 模块组成的并联架构。将通过卷积提取的特征输入上述并联架构，然后对两个注意力路径的特征进行融合，最后经过 BatchNorm 归一化与 Tanh 激活函数处理得到最终的输出结果。在 EMA 模块的特征提取过程中，设计了 3 条并行路径，用于获取分组特征图的注意力权重描述符：第 1、2 条路径采用 1×1 卷积分支，辅以沿水平方向与垂直方向的一维全局平均池化，提取空间维度上的全局统计信息。随后，通过共享的 1×1 卷积实现通道维度的交互与重构，增强通道间的信息整合能力。第 3 条则采用 3×3 卷积分支以捕捉更大感受野下的空间语义信息，从而丰富多尺度特征的表达能力，增强网络对局部纹理与全局结构的适应性。在特征编码过程中，EMA 模块通过将部分通道维度重构为批处理维度，引入跨通道交互机制，避免了通用卷积操作导致的特征维度压缩问题，进而保留更加完整且具有辨识力的表示。然而，EMA 中多个并行子路径可能引入功能冗余，增加特征冗余度。针对这一问题，本文将 EMA 模块与 CA 模块并联集成，利用后者在通道和空间坐标维度的协同建模能力，进一步提升特征选择的精度。CA 模块沿空间维度进行方向分解，实现对通道注意力与空间结构信息的协同感知，进而强化模型对关键区域的响应能力。通过非线性变换机制，CA 模块可有效建模远程依赖关系，优化特征聚焦区域，尤其适用于处理背景复杂、对比度较低的多模态图像场景[23]。本文设计的 EMA-CA 并联架构使得融合模型能够充分利用多尺度特征与高频信息，在保留互补信息的同时消除冗余，有效提升融合图像的质量与表达能力。

2.4　数据驱动的自适应损失权重

对于 IVIF，需要学习到最小化源图像与融合图像的相似度差异。损失函数主要由分解损失、结构相似损失、强度相似损失和梯度损失四大关键部分组成，可综合表示为

$$L_{\text{total}}=L_{\text{decomp}}+\alpha_1L_{\text{SSIM}}+\alpha_2L_{\text{MSE}}+\alpha_3L_{\text{grad}}$$

式中：L_{decomp}表示分解损失，L_{SSIM}表示结构相似性损失，L_{MSE}表示强度相似性损失，L_{grad}表示梯度损失。

分解损失包括最小化背景特征图的差异并最大化细节特征图差距，以从源图像中提取共同特征，同时捕捉红外和可见光图像之间的不同特征。因而，L_{decomp}可表示为

$$L_{\text{decomp}}=\Phi\left(\|B_{\text{vis}}-B_{\text{ir}}\|_2^2\right)-\beta_1\Phi\left(\|D_{\text{vis}}-D_{\text{ir}}\|_2^2\right) \tag{3}$$

式中：B_{vis}、D_{vis}为可见光图像V的背景和细节特征图，B_{ir}、D_{ir}为红外图像I的背景和细节特征图；β_1为

调谐参数; Φ(·)为 tanh 函数, 用于将间隔约束到(−1、1) 内。

Wang 等[24] 提出的 SSIM 是衡量图像结构相似性的重要指标, L_{SSIM} 可表示为

$$L_{\mathrm{SSIM}}=\sigma^{a}\left(1-\mathrm{SSIM}\left(I,\hat{I}\right)\right)+\sigma^{b}\left(1-\mathrm{SSIM}\left(V,\hat{V}\right)\right)$$

式中 σ^{a} 与 σ^{b} 是用于平衡红外图像与可见光图像对 L_{SSIM} 贡献的比例权重。

利用 L2 范数构建强度损失, L_{MSE} 可表示为

$$L_{\mathrm{MSE}}=\gamma^{a}\left\|I-\hat{I}\right\|+\gamma^{b}\left\|V-\hat{V}\right\|$$

式中 γ^{a} 与 γ^{b} 是用于平衡红外图像与可见光图像对 L_{MSE} 贡献的比例权重。

由于可见光图像具有丰富的纹理, 因此通过梯度稀疏惩罚对可见光图像的重建进行正则化, 以保证纹理的一致性。利用 L1 范数构建梯度损失。L_{grad} 可表示为

$$L_{\mathrm{grad}}=\left\|\nabla V-\nabla\hat{V}\right\|_{1}$$

其中∇代表梯度算子。

在 Li 等[25] 提出的方法中, 式 (1) 和式 (2) 中的 $\{\sigma^{a},\sigma^{b}\}$ 和 $\{\gamma^{a},\gamma^{b}\}$ 被经验地设置为固定值。在 Liu 等[1] 提出的方法中, 它们被设置为用不同模态平均梯度占比计算的比例权重。然而, 以上方式都不足以充分挖掘源图像特征, 且无法相应地调节只利用单独模态梯度损失项。

为了解决上述问题, 在梯度损失项中引入代表可见光图像本身梯度特征的项 μ。此时 L_{grad} 为

$$L_{\mathrm{grad}}=\mu\left\|\nabla V-\nabla\hat{V}\right\|_{1}$$

由于平均梯度 (average gradient, AG) 反映了图像的基本强度变化, 所以其被认为很好地匹配了梯度损失目标。因此本文采用 AG 来更新 μ, 其中 AG 可表示为

$$G_{\mathrm{A}}=g(P)=\frac{1}{HW}\left(\|\nabla_{h}P\|_{1}+\|\nabla_{v}P\|_{1}\right)$$

式中: P 代表图像, $\nabla_{h}I$ 和 $\nabla_{v}I$ 分别表示图像在水平和垂直方向上的一阶微分, H 和 W 分别是高度和宽度。

μ 可表示为

$$\mu=\frac{1}{1+\exp(g(V))}$$

SSIM 的设计初衷是模拟人类视觉系统的感知特性。通过引入 AG 作为权衡因子, 可以在保持整体结构相似性的同时, 更加强调那些具有更高清晰度和更多细节的部分。因此, AG 也很好地匹配了 SSIM 损失目标。为了进一步挖掘源图像特征并适应于 L_{SSIM}, σ^{a} 与 σ^{b} 计算公式为

$$\sigma^{a}=\frac{\mu\cdot\exp(g(I))}{\left[\exp(g(I))+\exp(g(V))\right]}$$

$$\sigma^{b}=\frac{\mu\cdot\exp(g(V))}{\left[\exp(g(I))+\exp(g(V))\right]}$$

另一方面, 为了实现不同损失项之间的互补, 提高整体性能, 本文采用与 Liu 等[1] 提出的方法相同的计算方式, 即使用图像熵 (entropy, EN) 来更新 L_{MSE} 的权重参数 γ 以融合具有高对比度的图像。EN 可表示为

$$E_{\mathrm{N}}=\varepsilon(P)=-\sum_{x=0}^{L-1}p_{x}\log_{2}p_{x}$$

式中: P 代表图像, L 表示给定图像的灰度级, p_{x} 表示像素位于相应灰度级的概率。正如公式 (3) 所示, EN 衡量了图像的信息量, 并且是在像素级别计算的, 它与 L_{MSE} 密切相关。由于具有更多信息的模态应该获得更高的权重以最大化有意义的特征, γ^{a} 与 γ^{b} 的更新公式为

$$\gamma^{a}=\frac{\exp(\varepsilon(I))}{\exp(\varepsilon(I))+\exp(\varepsilon(V))}$$

$$\gamma^{b}=\frac{\exp(\varepsilon(V))}{\exp(\varepsilon(I))+\exp(\varepsilon(V))}$$

3　实验验证与结果分析

3.1　数据描述与预处理

为评估 CM-MCNet 的有效性, 实验采用了 3 个公开的红外与可见光图像融合数据集: RoadScene[26]、TNO[27] 和 M^{3}FD[28]。其中, RoadScene 聚焦真实驾驶场景, 包含 221 对来自实际行车视频的高质量图像; TNO 主要用于军事安全, 提供 261 对多光谱图像; M^{3}FD 涵盖道路、校园、雾霾等复杂环境, 共计 4200 对图像。考虑到 RoadScene 图像质量高、场景典型, 本文选其作为主训练集, 示例如图 7 所示。

(a) RoadScene 数据集的红外图像　(b) RoadScene 数据集的可见光图像　(c) TNO 数据集的红外图像　(d) TNO 数据集的可见光图像　(e) M3FD 数据集的红外图像　(f) M3FD 数据集的可见光图像

图 7　3 个数据集示例图片

Fig. 7　Example images from three datasets

在训练前，将所有图像转为灰度图以消除颜色干扰，并统一进行中心裁剪 (128×128) 以聚焦关键区域，减少边缘冗余。上述预处理操作有助于提升模型对关键特征的提取能力，加快训练收敛，并增强实验的准确性与可重复性。

3.2　实验设置与评价指标

为了确保模型能在合理的时间内收敛，训练阶段采用 Adam 优化器，并引入 MultiStepLR 策略[29]动态调整学习率。实验设置中，批量大小为 8，训练周期为 120 轮。模型基于 PyTorch 框架实现，运行于 Python 环境下，硬件为 NVIDIA RTX 3080 GPU。

训练数据使用 RoadScene 训练集 (190 对)，验证集为 NIR-Street[30](50 对)；测试阶段则在 RoadScene(30 对) 和 TNO(40 对) 上评估模型性能，以检验其在不同场景下的泛化能力。

评价指标为熵、标准差 (standard deviation, SD)、空间频率 (spatial frequency, SF)、视觉信息保真度 (visual information fidelity, VIF) 和平均梯度。这 5 个指标能够从信息量、清晰度、细节丰富度及视觉质量等多个维度来评估融合图像的质量，能够全面地反映出方法在红外与可见光图像融合任务中的表现。这些度量的细节可以在文献 [31] 中找到。这些指标值越高说明生成的融合图像越好。

为确保评估的客观性与公正性，测试阶段未对 RoadScene 和 TNO 数据集进行微调。

3.3　实验结果与分析

3.3.1　参数实验

为增强模型的空间建模能力，CWPMLP 中引入通道划分超参数 S，以调控空间特征重排的粒度，实现性能与效率的平衡。在 NIR-Street 数据集上，S 尝试取值为 2、4、8、16 和 32，并选取 EN、SD、SF、AG、VIF 及浮点运算数作为评价指标。实验结果如表 1 所示。

表 1　参数 S 调整下模型性能指标的定量比较

Table 1　Quantitative comparison of model performance indicators under parameter S adjustment

S	EN	SD	SF	VIF	AG	浮点运算数/10^9
2	7.04	61.53	20.86	0.99	6.44	8.60
4	7.07	62.45	21.46	1.00	6.59	8.75
8	7.11	63.43	22.03	1.00	6.68	9.05
16	7.09	63.22	22.25	1.01	6.67	9.66
32	7.08	63.51	22.21	1.01	6.65	10.87

从表 1 可以看出 S 从 2 增至 8 时，融合性能持续提升，说明适度划分有助于细粒度建模；而当 S 超过 8 后，指标趋于饱和甚至下降，表明过度切分削弱通道协同建模能力。同时，浮点运算数随 S 增大而上升，计算开销增加。综合考虑，本文最终将 S 设置为 8，以兼顾融合效果与计算效率。

为充分发挥标准卷积与中心差分卷积的互补优势，MsCConv 模块引入超参数 θ 以调控两者权重，实现全局上下文与局部细节的协同建模。在 NIR-Street 数据集上对 $\theta \in [0.0,1.0]$ 进行系统实验，评价指标包括 EN、SD、SF、VIF 和 AG，比较结果如表 2 所示。

表 2　参数 θ 调整下模型性能指标的定量比较

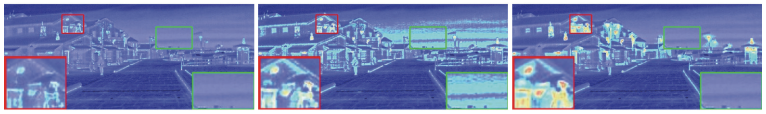
Table 2　Quantitative comparison of model performance indicators under parameter θ adjustment

θ	EN	SD	SF	VIF	AG
0.0	6.99	61.28	19.84	1.01	6.07
0.1	7.01	61.47	20.39	1.01	6.20
0.2	7.02	61.95	20.75	1.01	6.42
0.3	7.03	62.04	21.10	1.01	6.38
0.4	7.04	62.49	21.69	1.00	6.57
0.5	7.07	62.75	22.01	1.00	6.61
0.6	7.09	62.96	21.80	1.00	6.64
0.7	7.11	63.43	22.03	1.00	6.68
0.8	7.11	63.21	22.04	0.98	6.69
0.9	7.12	63.38	22.06	0.97	6.71
1.0	7.13	62.82	22.07	0.97	6.71

由表 2 可知随着 θ 的增加，EN、SD、SF 与 AG 整体提升，表明适度引入差分卷积有助于增强细节与对比度。然而，当 θ 超过 0.7 时，上述指标趋于饱和，且 SD、VIF 指标出现轻微回落，表明过度强调差分卷积引入高频噪声与伪边缘，损害结构连续性与视觉保真度。综合来看，θ 为 0.7 时性能最优，兼顾结构一致性与细节表现。

3.3.2　消融实验

为对 CWPMLP 和 MsCConv 进行有效性分析，本文使用 Selvaraju 等[32] 提出的类梯度激活热力图 (gradient-weighted class activation mapping, Grad-CAM) 对 CM-MCNet 网络编码路径进行可视化。从 RoadScene 数据集中选择一组典型的红外与可见光图像并将可视化结果展示在图 8 中。



(a) 基线　(b) CM-MCNet-A　(c) CM-MCNet-B

图 8　CM-MCNet 编码阶段热力图

Fig. 8　Heat maps of CM-MCNet during the encoding phase

红框与绿框分别对应放大的目标区域与背景区域。其中 CM-MCNet-A 表示在基线模型中引入 CWPMLP 模块, CM-MCNet-B 则在前者基础上进一步集成 MsCConv 模块。从图中可以看出, CWPMLP 的引入显著提升了模型对大范围区域的响应能力, 表明该模块在整合全局上下文信息方面具有积极作用, 有助于模型学习更加全面的特征表示。然而, CM-MCNet-B 的热力图显示模型对天空等平坦区域也产生了较高响应, 说明存在冗余信息干扰, 目标与背景区分能力仍有提升空间。进一步引入 MsCConv 后, 模型对图像中的边缘与纹理区域表现出更高关注度, 且对无意义平坦区域的响应降低, 说明 MsCConv 有效增强了对关键高频细节的感知能力, 实现了对冗余信息的抑制与互补信息的保留。

此外, 基于 RoadScene 数据集中 30 对图像进行消融实验, 结果如表 3 所示。

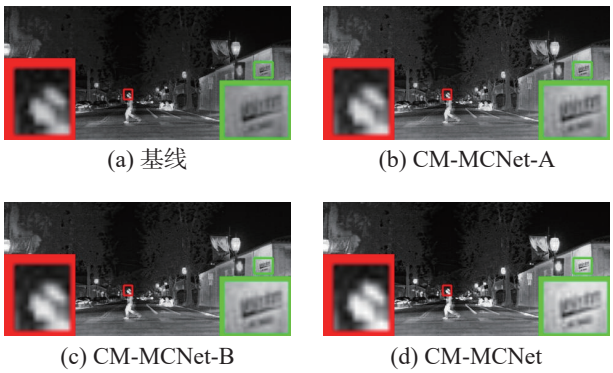
表 3　4 种融合模型的定量对比结果

Table 3　Quantitative comparison results of four fusion models

方法编号	CWPMLP	MsCConv	数据驱动 自适应损失	EN	SD	SF	VIF	AG
1	×	×	×	7.38	53.20	15.57	0.58	5.78
2	√	×	×	7.48	56.16	17.25	0.64	6.36
3	√	√	×	7.49	58.55	18.03	0.65	6.62
4	√	√	√	7.50	60.01	18.96	0.65	6.87

从表 3 可以看出引入 CWPMLP 后, EN、SD、SF、VIF、AG 分别提升 1.364%、5.56%、10.79%、10.34%、10.03%, 表明其有助于保留关键信息并提升融合一致性。进一步引入 MsCConv 与自适应损失权重后, SF、SD、AG 分别提升 4.26%、4.52%、4.09%, VIF、AG 也小幅改善, 验证其在增强细节与对比度方面的有效性。综上, 所提模块能显著提升融合图像质量。

另外, 从 RoadScene 数据集中选出一组典型的融合结果, 以展示 4 种模型的融合性能, 结果如图 9 所示。



(a) 基线　(b) CM-MCNet-A　(c) CM-MCNet-B　(d) CM-MCNet

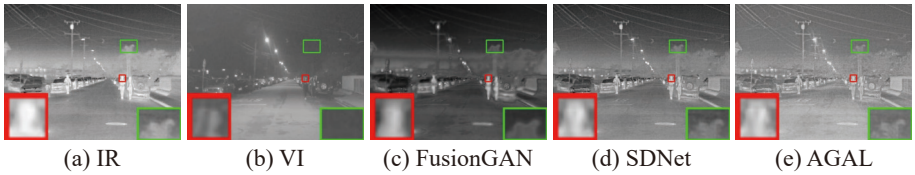
图 9　消融实验的可视化对比结果

Fig. 9　Qualitative comparison of ablation studies

从图 9 可以看出, CM-MCNet 生成的融合图像在红外目标显著性与可见光纹理清晰度方面表现更优, 整体对比度更高、结构更均衡。基线模型存在红框内目标不明显、绿框中纹理模糊等问题。引入 CWPMLP 后, 得益于其模拟排列操作增强了空间建模能力, 该问题显著改善。此外重加权机制根据全局特征重要性动态调整特征组合, 提升上下文整合效果。进一步引入 MsCConv 模块与自适应损失权重后, 图像边缘与纹理逐渐清晰, 说明 MsCConv 增强了高频信息保留与多尺度聚合能力, 而自适应权重优化了细节一致性。

3.3.3　对比实验

为对比 CM-MCNet 与代表性 IVIF 方法的性能, 以 RoadScene 为基准进行模型训练, 在 RoadScene 测试集及 TNO 测试集上进行了定性与定量实验。对比方法包括 FusionGAN[16]、SDNet[33]、AGAL[17]、ReCoNet[34]、DATFuse[15]、IRFS[35]、FusionMamba[36]。为了对融合结果进行直观的比较, 从 Roadscene 和 TNO 中分别选取了一组典型图像进行主观对比。结果如图 10~11 所示。红框与绿框分别对应放大的红外目标与可见光细节区域。

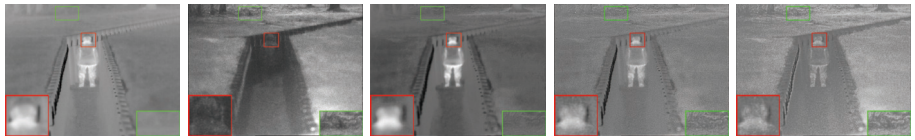


(a) IR　(b) VI　(c) FusionGAN　(d) SDNet　(e) AGAL

(f) ReCoNet　(g) DATFuse　(h) IRFS　(i) FusionMamba　(j) CM-MCNet

图 10　不同方法在 RoadScene 数据集上的可视化对比

Fig. 10　Qualitative comparison of different methods on the RoadScene dataset



(a) IR　(b) VI　(c) FusionGAN　(d) SDNet　(e) AGAL

(f) ReCoNet　(g) DATFuse　(h) IRFS　(i) FusionMamba　(j) CM-MCNet

图 11　不同方法在 TNO 数据集上的可视化对比

Fig. 11　Qualitative comparison of different methods on the TNO dataset

从图中可以看出，ReCoNet、DATFusion、IRFS 较好地保留了可见光细节，维护了场景信息与纹理结构；但对红外目标表现不足，图像对比度较低，红外信息表达不充分。而 FusionGAN 成功保留红外图像中的典型目标特征，但纹理信息损失严重，边缘模糊。SDNet、AGAL、FusionMamba 未表现出明显倾向，融合结果在细节表达上仍存在不足，如行人轮廓模糊、树木纹理不清，且 ReCoNet、DATFuse、IRFS 整体色调偏暗，目标与背景区分不明显。相比之下，CM-MCNet 融合结果中红外目标清晰突出，细节丰富，视觉效果更契合人眼感知特性。

进一步对不同融合方法在 RoadScene 与 TNO 数据集上性能进行定量比较，结果如表 4 所示。不同融合方法的参数量对比如表 5 所示。

表 4　不同融合方法在 RoadScene 与 TNO 数据集上的性能对比

Table 4　Performance and parameter comparison of different fusion methods on RoadScene and TNO datasets

数据集	融合方法	EN	SD	SF	VIF	AG
RoadScene	FusionGAN[16]	7.03	46.84	9.80	0.38	3.31
	SDN[33]	7.28	53.97	17.26	0.63	6.01
	AGAL[17]	6.82	38.09	17.12	0.53	6.03
	ReCoNet[34]	6.95	45.75	9.86	0.55	3.64
	DATFuse[15]	6.63	36.25	13.04	0.61	4.08
	IRFS[35]	6.90	41.52	11.13	0.58	3.71
	FM[36]	7.06	58.51	13.89	0.57	4.78
	CM-MCNet	7.50	60.01	18.96	0.65	6.87
TNO	FusionGAN[16]	6.58	40.14	6.88	0.42	2.41
	SDN[33]	6.70	43.59	12.80	0.58	4.60
	AGAL[17]	6.42	34.26	12.02	0.55	4.38
	ReCoNet[34]	6.82	50.97	7.95	0.54	3.17
	DATFuse[15]	6.45	35.51	10.53	0.68	3.57
	IRFS[35]	6.62	40.25	9.45	0.59	3.14
	FM[36]	7.08	51.48	8.39	0.55	3.22
	CM-MCNet	6.90	52.77	13.16	0.63	4.61

表 5　不同融合方法的参数量对比

Table 5　Comparison of the number of parameters for different fusion methods

融合方法	FusionGAN[16]	SDN[33]	AGAL[17]	ReCoNet[34]	DATFuse[15]	IRFS[35]	FM[36]	基线	CM-MCNet-A	CM-MCNet-B	CM-MCNet
参数量/10^6	0.93	0.07	1.59	0.01	0.26	0.24	0.77	0.26	0.46	0.49	0.49

在 RoadScene 数据集上, CM-MCNet 在所有指标上均取得最佳表现, EN、SD、SF、VIF、AG 相较次优结果分别提升 2.93%、2.56%、3.79%、3.08%、12.23%。在 TNO 数据集上, 所提方法在 SD、SF、AG 指标上取得最优值, 相较次优值分别提升 2.51%、3.95%、0.22%。EN 指标略低于 Fusion-Mamba, 主要是由于后者在自然场景中更倾向于保留可见光图像中的高频噪声和背景细节, 从而提升了信息熵值, 但这并未带来结构感知的同步提升。VIF 指标则略低于 DATFuse, 是由于该方法在自然场景丰富的 TNO 数据集中更偏向可见光图像, 因而在视觉信息保真度上占优。尽管如此, 综合 2 个数据集和多项评价指标, CM-MCNet 在融合质量与结构保真之间实现了更优平衡, 展现出更强的综合性能与鲁棒性。

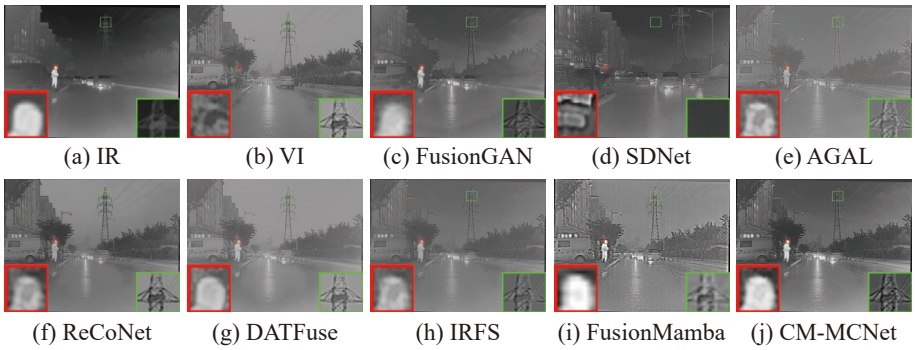
尽管 CM-MCNet 引入 CWPMLP 与 MsCConv 模块以增强多尺度建模与全局感知能力, 参数量略有增加, 但仍控制在合理范围内, 未超出部署资源限制。结合其显著的融合性能提升, CM-MCNet 在性能与复杂度之间实现了良好的权衡。

3.4　IVIF 应用

3.4.1　在多场景图像上的应用

为进一步验证 CM-MCNet 在 IVIF 任务中的泛化能力, 本文将其扩展至多场景图像, 并在 M^3FD 数据集上开展系统实验。选取 800 对图像用于训练, 100 对验证, 100 对测试, 训练策略与评价指标与 RoadScene 实验一致, 并与多种代表性方法进行定性与定量对比。

从图 12 的可视化结果可见, FusionGAN 生成图像中行人轮廓模糊, 树木纹理不清; SDNet、AGAL、FusionMamba 在细节表现上有所改进, 但依然丢失部分可见光信息; ReCoNet、DATFuse 和 IRFS 色调偏暗, 目标与背景区分不明显。相比之下, CM-MCNet 融合图像中红外目标清晰, 热辐射信息表达充分, 行人轮廓与树叶纹理等细节更加突出, 视觉感知更佳。



(a) IR　(b) VI　(c) FusionGAN　(d) SDNet　(e) AGAL

(f) ReCoNet　(g) DATFuse　(h) IRFS　(i) FusionMamba　(j) CM-MCNet

图 12　不同方法在 M3FD 数据集上的可视化对比

Fig. 12　Qualitative comparison of different methods on the M3FD dataset

表 6 给出了各方法在 M^3FD 数据集上的定量结果。CM-MCNet 在 EN、SD、SF、VIF4 项指标上均获最优, 分别提升 6.48%、5.04%、2.67%、5.41%; AG 指标略低于 AGAL, 仅差 0.09%。AGAL 在该指标上表现更优, 主要得益于其强调边缘的融合策略。总体而言, CM-MCNet 在多场景图像融合中表现出更强的综合能力与结构感知水平。

表 6　M^3FD 定量对比实验结果

Table 6　Quantitative comparison results on M^3FD

方法	EN	SD	SF	VIF	AG
FusionGAN[16]	6.72	39.54	9.14	0.46	3.15
SDN[33]	6.78	42.72	14.67	0.61	5.19
AGAL[17]	6.55	34.51	15.30	0.70	5.71
ReCoNet[34]	6.58	38.85	10.53	0.63	4.01
DATFuse[15]	6.32	30.47	10.42	0.68	3.44
IRFS[35]	6.52	33.92	10.47	0.67	3.50
FM[36]	6.91	42.84	11.21	0.41	4.64
CM-MCNet	7.25	45.00	15.72	0.74	5.66

3.4.2　极端场景下的应用

为评估模型在极端条件下的鲁棒性, 本文设

计了低光照与运动模糊两类退化实验，模拟真实环境中图像质量下降的情况。在低光照实验中对原始图像亮度进行压缩并添加高斯噪声；在运动模糊实验中施加方向性模糊核进行卷积退化。处理后图像输入各融合模型并进行定量评估。结果如表 7 所示。

表 7　低光照及运动模糊条件下融合性能的定量评估

Table 7　Quantitative evaluation of fusion performance under low-light and motion blur conditions

条件类别	方法	EN	SD	SF	VIF	AG
低光照	FusionGAN[16]	6.91	39.94	8.76	0.26	3.81
	SDN[33]	7.14	41.20	16.45	0.39	6.35
	AGAL[17]	6.57	30.06	17.64	0.35	6.94
	ReCoNet[34]	6.76	38.18	8.87	0.37	3.58
	DATFuse[15]	6.53	31.06	13.84	0.38	4.69
	IRFS[35]	6.96	41.28	14.82	0.38	4.95
	FM[36]	7.05	42.85	11.48	0.40	4.27
	CM-MCNet	7.11	43.91	18.00	0.39	6.62
运动模糊	FusionGAN[16]	6.75	43.69	8.11	0.93	2.71
	SDN[33]	6.85	49.91	10.67	1.21	3.68
	AGAL[17]	6.48	36.18	10.20	1.09	3.47
	ReCoNet[34]	6.59	43.06	6.00	1.08	2.07
	DATFuse[15]	6.36	34.97	8.15	1.05	2.38
	IRFS[35]	7.16	52.34	8.32	1.26	3.04
	FM[36]	6.61	61.03	9.30	1.10	3.25
	CM-MCNet	6.99	53.56	9.64	1.25	3.23

表 7 表明，在低光照场景下，所提方法在所有指标上都能取得最优或次优值，说明所提方法在低光照条件下依旧表现良好。在运动模糊场景下，尽管所提方法并未在所有指标上取得最优，但其在 3 项关键指标中排名第 2，其余指标亦位居前列，表现出优异的综合稳定性。相较于部分方法在单一指标上的波动表现，所提方法在多种评估维度上均衡、鲁棒，验证了其在复杂模糊环境中的适应能力。

3.4.3　在语义分割中的应用

语义分割作为计算机视觉中的经典任务，旨在精确划分图像中的不同语义区域，是衡量多模态图像融合效果的重要下游手段之一。本节重点探讨图像融合在提升语义分割性能方面的作用。

本文在 MSRS 数据集[37] 上进行了多模态语义分割实验。该数据集包含了 9 个物体类别 (背景、汽车、人、自行车、弯道、停车标志、护栏、交通锥和减速带) 的语义信息。数据集的划分遵循文献 [37]。实验中使用了 DeeplabV3+[38] 模型，并通过交并比 (IoU) 来比较模型的有效性。为了说明图像融合对下游任务的积极作用，对比了在融合图像上进行语义分割的效果与在单独的红外图像和可见光图像上的效果。

此外，为了展示本文方法相对于其他方法的优越性，还与两种先进的融合方法进行了对比。所有模型均使用交叉熵损失进行监督，并使用随机梯度下降在 100 个历元上以批处理大小为 8 进行训练，其余设置均遵循其原始配置。语义分割的定量实验结果如表 8 所示。

表 8　MSRS 数据集上多模态分割的 IoU 值

Table 8　IoU values for MM segmentation on the MSRS dataset　%

方法	背景	汽车	人	自行车	弯道	停车标志	护栏	交通锥	减速带	mIOU
IR	84.7	67.8	56.4	51.8	34.6	39.3	42.2	40.2	48.4	51.7
VI	90.5	75.6	45.4	59.4	37.2	51.0	46.4	43.5	50.2	55.4
SDN[33]	97.3	78.4	62.5	61.7	35.7	49.3	52.4	42.2	52.9	59.2
ReCoNet[34]	97.4	81.0	59.9	61.4	41.0	51.3	54.4	47.4	55.9	61.1
CM-MCNet	97.8	84.2	62.4	61.2	34.0	61.7	68.9	42.4	66.0	64.3

可以看出，所提方法在背景、汽车、停车标志、护栏和减速带五类语义类别上取得最高 IoU，并在平均交并比 (mIoU) 上排名第一，显示出卓越的分割性能。这一优势主要归因于两方面：一是引入的 CWPMLP 模块显著增强了全局信息的提取能力，提升了模型对复杂场景中多类目标的理解能力；二是 MsCConv 模块强化了高频信息的表达能力，使得融合图像在纹理、边缘细节等方面更为清晰，有效提升了语义分割的精度。此外，该模块在过滤模态间冗余信息的同时，充分保留了互补特征，进一步促进了整体语义理解。

4　结束语

为了解决传统的 IVIF 模型在局部细节保留与全局结构整合之间的平衡难题，本文提出了一种新颖的多尺度协调网络模型，即 CM-MCNet。通过引入 CWPMLP 模块和 MsCConv 模块，CM-MCNet 有效地增强了全局上下文信息的整合以及细节信息的保留。此外，MsCConv 模块中 EMA 与 CA 并联架构进一步提升了多尺度特征的表达

能力。同时, 设计了一种数据驱动自适应损失权重, 以克服传统损失函数中经验权重设置的局限性。主观评价和客观评价结果表明, 提出的方法在图像的信息量、清晰度、细节丰富度以及视觉质量等方面具有良好的性能。更重要的是, CM-MCNet 在多场景 IVIF 及语义分割任务中也表现出优越性能, 这不仅验证了该方法的有效性和广泛适用性, 还为其在更多领域的应用提供了可能性。然而, 尽管 CM-MCNet 在 IVIF 中取得了显著进展, 但在极端照明条件下, 光照因素对融合效果的影响尚未得到充分考虑。未来的研究将集中于开发基于光照感知的方法, 以期进一步提高融合网络在各种环境下的鲁棒性。

参考文献:

[1] LIU Jinyuan, LIN Runjia, WU Guanyao, et al. Coconet: coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion[J]. International journal of computer vision, 2024, 132(5): 1748−1775.

[2] 蓝鑫, 谷小婧. 基于域适应互增强的多模态图像语义分割[J]. 计算机工程与设计, 2022, 43(9): 2584−2593.
LAN Xin, GU Xiaojing. Multi-modal image semantic segmentation based on domain adaptation and mutual enhancement[J]. Computer engineering and design, 2022, 43(9): 2584−2593.

[3] 黎瑞虹, 付志涛, 张韶琛, 等. 基于多注意力机制的红外与可见光图像夜间目标检测[J]. 红外技术, 2024, 46(12): 1371−1379.
LI Ruihong, FU Zhitao, ZHANG Shaochen, et al. Nighttime object detection in infrared and visible images based on multi-attention mechanism[J]. Infrared technology, 2024, 46(12): 1371−1379.

[4] SUN Yiming, CAO Bing, ZHU Pengfei, et al. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning[J]. IEEE transactions on circuits and systems for video technology, 2022, 32(10): 6700−6713.

[5] 张心祎, 谭耀, 邢向磊. 基于物理先验的深度特征融合水下图像复原[J]. 智能系统学报, 2023, 18(6): 1185−1196.
ZHANG Xinyi, TAN Yao, XING Xianglei. Deep feature fusion for underwater-image restoration based on physical priors[J]. CAAI transactions on intelligent systems, 2023, 18(6): 1185−1196.

[6] 张志超, 左雷鹏, 邹捷, 等. 基于多模态图像信息的变电设备红外分割方法[J]. 红外技术, 2023, 45(11): 1198−1206.
ZHANG Zhichao, ZUO Leipeng, ZOU Jie, et al. Segmentation method of substation equipment infrared based on multimodal image information[J]. Infrared technology, 2023, 45(11): 1198−1206.

[7] 杨爱萍, 刘瑾, 邢金娜, 等. 基于内容特征和风格特征融合的单幅图像去雾网络[J]. 自动化学报, 2023, 49(4): 769−777.
YANG Aiping, LIU Jin, XING Jinna, et al. Content feature and style feature fusion network for single image dehazing[J]. Acta automatica sinica, 2023, 49(4): 769−777.

[8] 李景景, 杜梅, 孙滨. 基于卷积神经网络的红外与可见光图像融合方法[J]. 激光杂志, 2024, 45(2): 135−139.
LI Jingjing, DU Mei, SUN Bin. Infrared and visible image fusion method based on convolutional neural network[J]. Laser journal, 2024, 45(2): 135−139.

[9] ZHAO Zixiang, XU Shuang, ZHANG Chunxia, et al. DIDFuse: Deep image decomposition for infrared and visible image fusion[EB/OL]. (2020−03−20) [2025−08−20]. https://arxiv.org/abs/2003.09210.

[10] LIANG Pengwei, JIANG Junjun, LIU Xianming, et al. Fusion from decomposition: A self-supervised decomposition approach for image fusion[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 719−735.

[11] LIU Jinyang, DIAN Renwei, LI Shutao, et al. SGFusion: A saliency guided deep-learning framework for pixel-level image fusion[J]. Information fusion, 2023, 91: 205−214.

[12] ZHAO Zixiang, BAI Haowen, ZHANG Jiangshe, et al. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 5906−5916.

[13] 罗小同, 杨汶锦, 曲延云, 等. 基于全局局部协同的非均匀图像去雾方法[J]. 自动化学报, 2024, 50(7): 1−12.
LUO Xiaotong, YANG Wenjin, QU Yanyun, et al. Dehazeformer: nonhomogeneous image dehazing with collaborative global-local network[J]. Acta automatica sinica, 2024, 50(7): 1−12.

[14] TANG Wei, HE Fazhi, LIU Yu. YDTR: Infrared and visible image fusion via Y-shape dynamic transformer[J]. IEEE Transactions on Multimedia, 2022, 25: 5413−5428.

[15] TANG Wei, HE Fazhi, LIU Yu, et al. DATFuse: Infrared and visible image fusion via dual attention transformer[J]. IEEE transactions on circuits and systems for video technology, 2023, 33(7): 3159−3172.

[16] MA Jiayi, YU Wei, LIANG Pengwei, et al. FusionGAN: A generative adversarial network for infrared and visible image fusion[J]. Information fusion, 2019, 48: 11−26.

[17] LIU Jinyuan, SHANG Jingjie, LIU Risheng, et al. Attention-guided global-local adversarial learning for detail-preserving multi-exposure image fusion[J]. IEEE transactions on circuits and systems for video technology, 2022, 32(8): 5026−5040.

[18] CHENG Chunyang, XU Tianyang, WU Xiaojun. MUFusion: A general unsupervised image fusion network based on memory unit[J]. Information fusion, 2023, 92: 80−92.

[19] OUYANG Daliang, HE Su, ZHANG Guozhong, et al. Efficient multi-scale attention module with cross-spatial

learning[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes: IEEE, 2023: 1−5.
[20] HOU Qibin, ZHOU Daquan, FENG Jiashi. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. Virtual: IEEE, 2021: 13713−13722.
[21] TU Zhengzhong, Talebi H, ZHANG Han, et al. Maxim: Multi-axis mlp for image processing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 5769−5780.
[22] ZHANG Hang, WU Chongruo, ZHANG Zhongyue, et al. Resnest: split-attention networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 2736−2746.
[23] 刘金平, 吴娟娟, 张荣, 等. 基于结构重参数化与多尺度深度监督的 COVID-19 胸部 CT 图像自动分割[J]. 电子学报, 2023, 51(5): 1163−1171.
LIU Jinping, WU Juanjuan, ZHANG Rong, et al. Toward automated segmentation of COVID-19 chest CT images based on structural reparameterization and multi-scale deep supervision[J]. Acta electronica sinica, 2023, 51(5): 1163−1171.
[24] WANG Zhou, BOVIK A C, SHEIKHJ H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE transactions on image processing, 2004, 13(4): 600−612.
[25] LI Hui, WU Xiaojun. CrossFuse: a novel cross attention mechanism based infrared and visible image fusion approach[J]. Information fusion, 2024, 103: 102147.
[26] XU Han, MA Jiayi, LE Zhuliang, et al. Fusiondn: a unified densely connected network for image fusion[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020, 34(7): 12484−12491.
[27] TOET A, HOGERVORST M A. Progress in color night vision[J]. Optical engineering, 2012, 51(1): 010901−010901.
[28] LIU Jinyuan, FAN Xin, HUANG Zhanbo, et al. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 5802−5811.
[29] XIE Xiangning, LIU Yuqiao, SUN Yanan, et al. BenchENAS: a benchmarking platform for evolutionary neural architecture search[J]. IEEE transactions on evolutionary computation, 2022, 26(6): 1473−1485.
[30] BROWN M, SüSSTRUNK S. Multi-spectral SIFT for scene category recognition[C]//The 24th IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs: IEEE, 2011: 177−184.
[31] SINGH S, SINGH H, BUENO G, et al. A review of image fusion: Methods, applications and performance metrics[J]. Digital signal processing, 2023, 137: 104020.
[32] SELVRAJU R R, MICHAEL C, ABIISHEK D, et al. Grad-cam: visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 618−626.
[33] ZHANG Hao, MA Jiayi. SDNet: a versatile squeeze-and-decomposition network for real-time image fusion[J]. International journal of computer vision, 2021, 129(10): 2761−2785.
[34] HUANG Zhanbo, LIU Jinyuan, FAN Xin, et al. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 539−555.
[35] WANG Di, LIU Jinyuan, LIU Risheng, et al. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection[J]. Information fusion, 2023, 98: 101828.
[36] XIE Xinyu, CUI Yawen, TAN Tao, et al. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba[J]. Visual intelligence, 2024, 2(1): 37.
[37] TANG Linfeng, YUAN Jiteng, ZHANG Hao, et al. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware[J]. Information fusion, 2022, 83: 79−92.
[38] CHEN L C , ZHU Yukun, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer nature, 2018: 801−818.

作者简介：

刘诗怡，硕士研究生，主要研究方向为机器学习、计算机视觉和图像处理。E-mail: liushiyi@hunnu.edu.cn。

刘金平，教授，博士生导师，主要研究方向为机器学习、模式识别、工业过程监测、故障诊断、计算机视觉。主持、参与国家和省部级科研课题 10 余项，获国家发明专利授权 20 项。发表学术论文 80 余篇。E-mail: ljp@hunnu.edu.cn。

黄丽娟，讲师，主要研究方向为智能控制、机器学习和工业过程控制。主持、参与省部级和市厅级科研课题 5 项，获国家发明专利授权 6 项。E-mail: huanglijuan@csmzxy.edu.cn。