计算机科学
COMPUTER SCIENCE

MTFuse:基于Mamba和Transformer的红外与可见光图像融合网络

丁政泽，聂仁灿，李锦涛，苏华平，徐航

引用本文

丁政泽，聂仁灿，李锦涛，苏华平，徐航. MTFuse:基于Mamba和Transformer的红外与可见光图像融合网络[J]. 计算机科学，2025，52(8)：188-194.

DING Zhengze, NIE Rencan, LI Jintao, SU Huaping, XU Hang. MTFuse:An Infrared and Visible Image Fusion Network Based on Mamba and Transformer [J]. Computer Science, 2025, 52(8): 188-194.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240600106

MTFuse:基于 Mamba 和 Transformer 的红外与可见光图像融合网络

丁政泽　聂仁灿　李锦涛　苏华平　徐　航
云南大学信息学院　昆明 650091
(dingzhengze@stu.ynu.edu.cn)

摘　要　红外与可见光图像融合旨在保留红外图像的热辐射信息和可见光图像的纹理细节,以表示成像场景并全面促进下游视觉任务。基于卷积神经网络的融合模型由于专注于局部卷积运算,在捕获全局图像特征方面遇到限制。基于 Transformer 的模型虽然在全局特征建模方面表现出色,但也面临着二次复杂性带来的计算挑战。选择性结构化状态空间模型(Mamba)在具有线性复杂性的远程依赖建模方面表现出了巨大的潜力,为解决上述问题提供了一条有希望的路径。为了高效建模图像远程依赖,设计了一个残差选择性结构化状态空间模块(RMB)提取全局特征。同时,为了对多模态图像之间的关系进行建模,设计了一个跨模态查询融合注意力模块(CQAM)用于特征的自适应融合。此外,设计了一个由两项组成的损失函数,包括梯度损失和亮度损失,旨在以无监督的方式训练所提出的模型。与大量其他先进的方法在融合质量的对比实验和消融实验上证明了所提出的方法的有效性。

关键词:选择性结构化状态空间模型;Transformer;无监督学习;红外与可见光图像融合
中图分类号　TP391

MTFuse:An Infrared and Visible Image Fusion Network Based on Mamba and Transformer

DING Zhengze,NIE Rencan,LI Jintao,SU Huaping and XU Hang
School of Information Science and Engineering,Yunnan University,Kunming 650091,China

Abstract　Infrared and visible image fusion aims to retain the thermal radiation information from infrared images and the texture details from visible images to represent the imaging scene and comprehensively promote downstream visual tasks. Fusion models based on convolutional neural networks(CNNs) encounter limitations in capturing global image features due to their focus on local convolutional operations. Although Transformer-based models excel in global feature modeling,they also face computational challenges posed by quadratic complexity. Recently,the selective structured state-space model(Mamba) has shown great potential in modeling long-range dependencies with linear complexity,providing a promising path to address the aforementioned issues. To efficiently model long-range dependencies in images,this paper designs a residual selective structured state space module(RMB) for extracting global features. Simultaneously,to model the relationship between multimodal images,a cross-modal query fusion attention module(CQAM) is designed for adaptive feature fusion. Furthermore,a loss function consisting of two terms,including gradient loss and brightness loss,is designed to train the proposed model in an unsupervised manner. Comparative experiments on fusion quality and efficiency with numerous other state-of-the-art methods and ablation studies demonstrate the effectiveness of the proposed MTFuse method.

Keywords　Selective structured state space model,Transformer,Unsupervised learning,Infrared and visible image fusion

1　引言

图像融合是图像处理领域的一个基本课题[1],旨在通过整合原始图像中的关键信息[2]来创建信息丰富的融合图像。红外可见光图像融合(Infrared and Visible Image Fusion,IVIF)专注于从各种传感器中提取和整合互补特征[3],以创建

到稿日期:2024-06-17　返修日期:2024-09-26
基金项目:国家自然科学基金(61966037);云南省基础研究计划重点项目(202301AS070025,202401AT070467);国家重点研发项目(2020YFA0714301);云南省科技厅项目基金资助项目(2012105AF150011);云南省教育厅科学研究基金项目(2024Y031)
This work was supported by the National Natural Science Foundation of China(61966037),Key Project of Yunnan Basic Research Program(202301AS070025,202401AT070467),National Key Research and Development Project of China(2020YFA0714301),Science and Technology Department of Yunnan Province project Fundation(202105AF150011) and Yunnan Provincial Department of Education Science Foundation(2024Y031).
通信作者:聂仁灿(rcnie@ynu.edu.cn)

融合图像。具体而言，可见光图像包含丰富的纹理细节，但对光照变化敏感；相比之下，红外图像揭示了对光照变化不敏感的热信息，但无法提供足够的纹理细节。IVIF 的目标是生成融合图像，既保留红外图像中的热辐射数据，又保留可见光图像中的复杂纹理细节[4]。这些融合图像克服了可见光图像对光照条件的敏感性以及红外图像固有噪声和低分辨率的限制，并可以为下游应用提供帮助[5]。

传统的图像融合方法主要包括基于多尺度变换的方法[6]、基于稀疏表示的方法[7]，以及其他方法[8]。近年来，基于深度学习的方法成为了主流，大多数的方法使用卷积神经网络(CNN)提取特征，融合特征并进行图像重建。

然而，这些基于 CNN 的方法由于感受野有限，难以捕获全局上下文，因此生成高质量的融合图像较为困难。最近，一些基于 Transformer 的模型[9]或 Transformer 和 CNN 的组合在全局建模中表现出了出色的性能，但其自注意力机制的复杂度与令牌数量呈平方关系，导致计算开销显著增加。

为了解决上述问题，本文旨在设计一种高效捕获全局特征的红外与可见光图像融合网络。与现有方法不同，本文创新性地设计了一个残差选择性结构化状态空间模块(RMB)用于高效建模图像域内远程依赖，在线性计算复杂度下即可捕获全局上下文信息，克服了 CNN 感受野有限和 Transformer 计算量大的问题。此外，不同于大多数方法直接使用卷积神经网络进行特征融合，本文提出了一个跨模态查询融合注意力模块(CQAM)，通过自适应感知不同模态特征的关联性，实现了更优的跨域全局特征融合。最后，在多个公开数据集上的大量实验证明，所提出的融合方法在融合图像质量、计算效率等方面均优于现有方法，展现了其有效性和优越性。

2 相关工作

2.1 Transformer

Transformer 模型最先由 Vaswani 等[10]提出，并被广泛用于自然语言处理的各个任务中。Dosovitskiy 等[11]最先将这种结构应用于计算机视觉领域，提出了 Vision Transformer 架构并在分类任务中取得了优异的性能。Swin Transformer[12]通过引入分层结构和移位窗口机制来克服标准 Vision Transformer 无法处理不同尺寸图像和缺乏多尺度特征处理能力的局限性，并在密集预测任务中展示出不俗的性能。Restormer[13]通过对多头注意力和前馈网络等组件进行关键设计，使其能够在大图像上高效捕捉远程像素交互，在多个图像恢复任务上取得了最先进的结果。Lu 等[14]首先提出了一个统一的双流 Transformer 架构(ViLBERT)，用于联合处理跨模态的信息，并在多个视觉-语言任务上取得了显著的性能提升。这主要得益于不同模态的表示在 Transformer 中统一为令牌，并通过注意力机制进行融合，该机制同样也可以捕获长距离依赖，弥补 CNN 较难获取全局感受野的缺陷。

2.2 Mamba

Transformer 模型的注意力机制有效，无法对有限窗口之外的信息进行建模，并且计算、内存消耗随序列长度呈平方级增长[15]。针对这些缺点，Gu 等[16]提出了具有选择性机制的结构化状态空间模型(Mamba)。该模型在语言建模等任务中超越同尺寸 Transformer 的性能，且计算复杂度降至线性级别。Liu 等[17]将 Mamba 引入计算机视觉领域，在保证计算效率的同时有效提取图像全局特征，从而在多种视觉识别任务上取得了优异的性能。

3 网络介绍

3.1 状态空间模型

状态空间模型(SSMs)[18]通常被视为一个线性时不变系统，通过一个潜在状态 $\boldsymbol{y}(\boldsymbol{h})\in\mathbb{R}_N$ 来建立从刺激 $\boldsymbol{x}(t)\in\mathbb{R}_N$ 到响应 $\boldsymbol{y}(t)\in\mathbb{R}_N$ 的映射。该系统可以用线性常微分方程(ODE)来数学表达：

$$\boldsymbol{h}'(t)=\boldsymbol{A}\boldsymbol{h}(t)+\boldsymbol{B}\boldsymbol{x}(t) \tag{1}$$

$$\boldsymbol{y}(t)=\boldsymbol{C}\boldsymbol{h}(t) \tag{2}$$

其中，N 是状态大小，$\boldsymbol{A}\in\mathbb{R}_{N\times N}$。然而，由于 SSMs 是连续时间模型，将其直接用于机器学习会面临困难。为了解决这一问题，需要对 SSMs 进行离散化处理。零阶保持(ZOH)是一种常用的离散化方法，它将连续时间方程转换为离散时间方程：

$$\boldsymbol{h}_t=\bar{\boldsymbol{A}}\boldsymbol{h}_{t-1}+\bar{\boldsymbol{B}}\boldsymbol{x}_t \tag{3}$$

$$\boldsymbol{y}_t=\boldsymbol{C}\boldsymbol{h}_t \tag{4}$$

其中，$\bar{\boldsymbol{A}}=\exp(\Delta\boldsymbol{A})$ 和 $\bar{\boldsymbol{B}}=(\Delta\boldsymbol{A})^{-1}(\exp(\Delta\boldsymbol{A})-\boldsymbol{I})\cdot\Delta\boldsymbol{B}$ 是离散化状态参数，Δ 是离散化步长。然而，传统的 SSMs 假设系统参数对于不同的输入数据是固定不变的，这限制了它们的使用范围。为了克服这一局限性，Mamba 引入了选择性扫描机制，根据输入数据自适应地调整矩阵 $\boldsymbol{B}$、$\boldsymbol{C}$ 和 Δ 的参数。此外，Mamba 还采用了更加高效的硬件感知算法，进一步提升了其实际应用的潜力。

VMamba 在该基础上提出了二维状态空间模型(2D-SSM)。如图 1 所示，它按照不同的顺序对图像块进行 4 次扫描，使每一个图像块都能融合不同方向的全局信息。

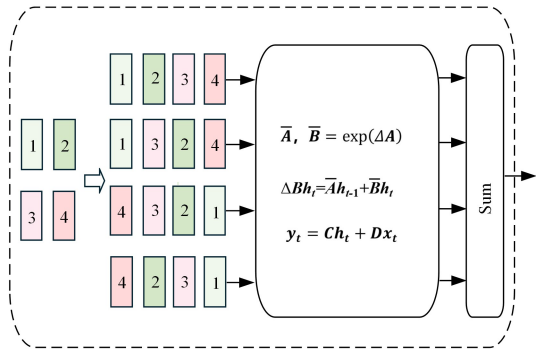


图 1 2D-SSM 的四向扫描机制

Fig. 1 Four-directional scanning mechanism of 2D-SSM

3.2 网络框架

本文网络结构如图 2 所示，先使用卷积分别提取模态内局部特征，之后利用残差选择性结构化状态空间模块(RMB)提取全局特征，并把两个模态的特征输入跨模态查询融合注意力模块(CQAM)中得到融合特征，并通过卷积重建融合图像。

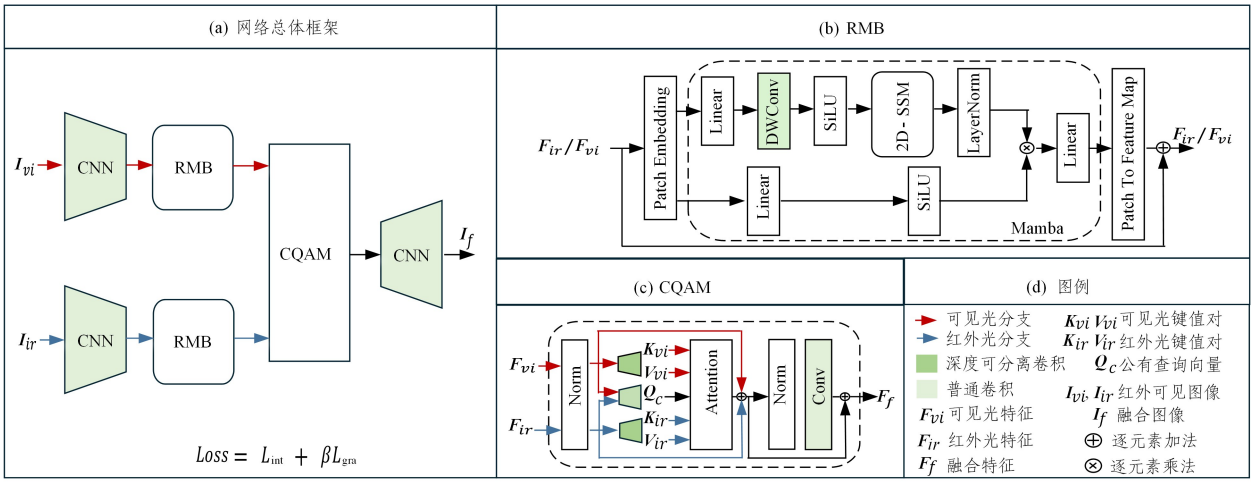


图 2 MTFuse 网络框架

Fig. 2 Framework of MTFuse

3.3 残差选择性结构化状态空间模块

为了解决 Transformer 提取特征过程中对输入尺寸不灵活，计算量大的问题，本文使用 RMB 建模图像的远程依赖。在红外和可见光分支分别经过卷积神经网络得到浅层局部特征$\boldsymbol{F}_{ir}$，$\boldsymbol{F}_{vi}$后，将其分别送入 RMB 提取全局特征，如图 2(b)所示。

每个 RMB 由线性层、深度可分离卷积、SiLU 激活函数、2D-SSM 块和归一化层等组成。

$\boldsymbol{F}_{ir}$或$\boldsymbol{F}_{vi}$先经过 Patch Embedding 层将特征图分割为 4 * 4 的小块并通过卷积提高隐藏层维度，如图 3 所示。其中，CHW 代表原特征图的尺寸，P 代表 patch 的尺寸，L 则代表了隐藏层维度。接着分别送入两个支路，其中一条支路包括线性层、深度可分离卷积、SiLU 激活函数、2D-SSM 块和归一化层，而另一支路包括线性层和 SiLU 激活函数。然后，将两个支路的输出结果相乘并经过线性层。最后，Mamba 处理后的特征经过 Reshape 和转置卷积恢复为特征图原有的形状。值得注意的是，为了确保浅层局部特征的丰富细节不被忽视，本文设计了残差连接。

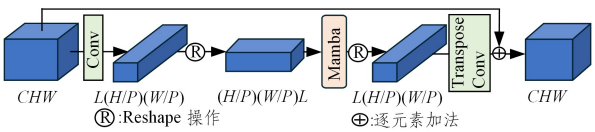


图 3 RMB 的 Patch Embedding 与反操作

Fig. 3 Patch Embedding and inverse operation in RMB

3.4 跨模态查询融合注意力模块

为了解决传统卷积在特征融合阶段利用跨模态信息不充分的问题，本文设计了跨模态查询融合注意力模块，如图2(c)所示。

对于红外特征$\boldsymbol{F}_{ir}$和可见特征$\boldsymbol{F}_{vi}$，通过如下公式得到公有查询向量$\boldsymbol{Q}_c$，红外键值对向量$\boldsymbol{K}_{ir}$，$\boldsymbol{V}_{ir}$和可见键值对向量$\boldsymbol{K}_{vi}$，$\boldsymbol{V}_{vi}$：

$$\boldsymbol{Q}_c = Conv(\boldsymbol{F}_{ir} + \boldsymbol{F}_{vi}) \tag{5}$$

$$\boldsymbol{K}_{ir}, \boldsymbol{V}_{ir} = DWConv(\boldsymbol{F}_{ir}) \tag{6}$$

$$\boldsymbol{K}_{vi}, \boldsymbol{V}_{vi} = DWConv(\boldsymbol{F}_{vi}) \tag{7}$$

其中，$Conv()$代表卷积层，而 $DWConv()$则代表深度可分离卷积。通过使用这两种卷积代替原注意力的线性层，在实现高效计算的同时保留图像的空间信息。接着，融合特征$\boldsymbol{F}_f$通过如下方式计算：

$$\boldsymbol{F}_f = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_c\boldsymbol{K}_{ir}^{\mathrm{T}} \otimes \boldsymbol{Q}_c\boldsymbol{F}_{vi}^{\mathrm{T}}}{\sqrt{d}}\right)(\boldsymbol{V}_{ir} \otimes \boldsymbol{V}_{vi}) \tag{8}$$

其中，$\otimes$代表通道拼接，d 代表向量维度。每个模态的向量$\boldsymbol{K}_{ir}$，$\boldsymbol{F}_{vi}$与公有查询向量$\boldsymbol{Q}_c$做点乘运算，经过激活函数处理后，得到每个模态特征对融合贡献的注意力图，并最终得到融合特征$\boldsymbol{F}_f$。通过这种方式，红外特征$\boldsymbol{F}_{ir}$和可见特征$\boldsymbol{F}_{vi}$的重要信息被自适应地聚合进融合特征$\boldsymbol{F}_f$。

3.5 损失函数

为了使融合图像$\boldsymbol{I}_f$既包含来自红外图像的目标信息，也包含来自可见光图像的纹理细节，本文采用逐元素的最大值选择运算符来捕捉给定源图像的最大强度，并使融合图像维持最佳强度分布。具体来说，强度损失定义如下：

$$L_{\mathrm{int}} = \| \boldsymbol{I}_f - \max(\boldsymbol{I}_{ir}, \boldsymbol{I}_{vi}) \|_1 \tag{9}$$

其中，max()代表逐元素的最大值选择，$\| \|_1$表示矩阵的l_1范数。纹理损失被设计用于约束融合结果以保持输入图像的空间细节，其公式为：

$$L_{\mathrm{gra}} = \| |\nabla \boldsymbol{I}_f| - \max((|\nabla \boldsymbol{I}_{ir}|, |\nabla \boldsymbol{I}_{vi}|)) \|_1 \tag{10}$$

其中，$|\cdot|$和∇代表求模与梯度运算。因此，本文的总损失函数可以表达为：

$$L = L_{\mathrm{int}} + \beta L_{\mathrm{gra}} \tag{11}$$

其中，β为权重系数，本文设β为 1。

4 实验

本章将对实验的具体细节、对比方法的选择和评价指标进行详细介绍。随后，本文将展示与几种先进方法的比较结果，包括定性对比、定量对比和运行效率分析，以证明网络的有效性。另外，通过消融实验，本章将验证 MTFuse 在网络设计选择方面的有效性。

4.1 实验细节

本文使用 PyTorch 框架在 NVIDIA 4090D GPU 上实现了 MTFuse 网络。训练过程中，将批大小(Batch Size)设置为 30，学习率设为 10^{-4}，在训练阶段，图像被裁剪为 64 * 64 大小，并使用 Adam 优化器训练 500 轮。

4.2 数据集和测试指标

本文在 3 个广泛使用的 IVIF 数据集上验证了所提出方法的有效性，包括 MSRS，TNO 和 M3FD。每个数据集包含不同场景下成对配准的红外和可见光图像。MSRS 数据集包含 1 083 对用于进行训练的图像对和 361 对用于测试的图像对，用于模型训练和评估。TNO 数据集的全部 40 张图像和 M3FD 数据集的 300 张图像被用于进一步测试。

此外，为了定量衡量融合结果，本文使用了 7 个常用的指标来证明 MTFuse 的有效性，包括互信息(MI)[19]、视觉信息保真度(VIF)[20]、基于梯度的融合度量(Qabf)[21]和结构相似性(SSIM)[22]、空间频率(SF)[23]、平均梯度(AG)[24]以及峰值信噪比(PSNR)[25]。MI 评估从源图像到融合图像的信息量传递；VIF 量化融合图像符合人类视觉感知的保真度；Qabf 衡量从源图像传递到融合图像的边缘信息保留情况；SSIM 评估融合结果与源图像之间的结构相似性；SF 反映图像灰度的变化率，可以体现图像的清晰程度；AG 主要反映图像中细节间的区别和纹理变化；而峰值信噪比(PSNR)是融合图像中峰值功率与噪声功率的比值[5]，反映了融合过程中的失真情况。除 PSNR 外，指标越高，融合质量越好。

4.3 对比实验

MSRS，M3FD 和 TNO 数据集部分图片的融合结果如图 4 和图 5 所示。为了更清晰地展示比较结果，背景纹理信息和伪彩处理后的热辐射信息在图片中被红框和绿框标记。另外，MTFuse 与其他对比方法的平均指标结果如表 1 和表 2 所列，最佳和次佳的结果分别用粗体和下划线标识。

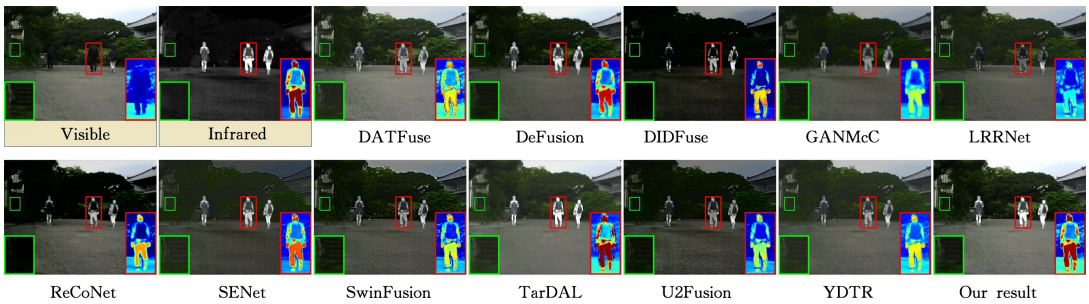


图 4 MSRS 定性对比结果(电子版为彩图)

Fig. 4 Qualitative comparison results on MSRS dataset

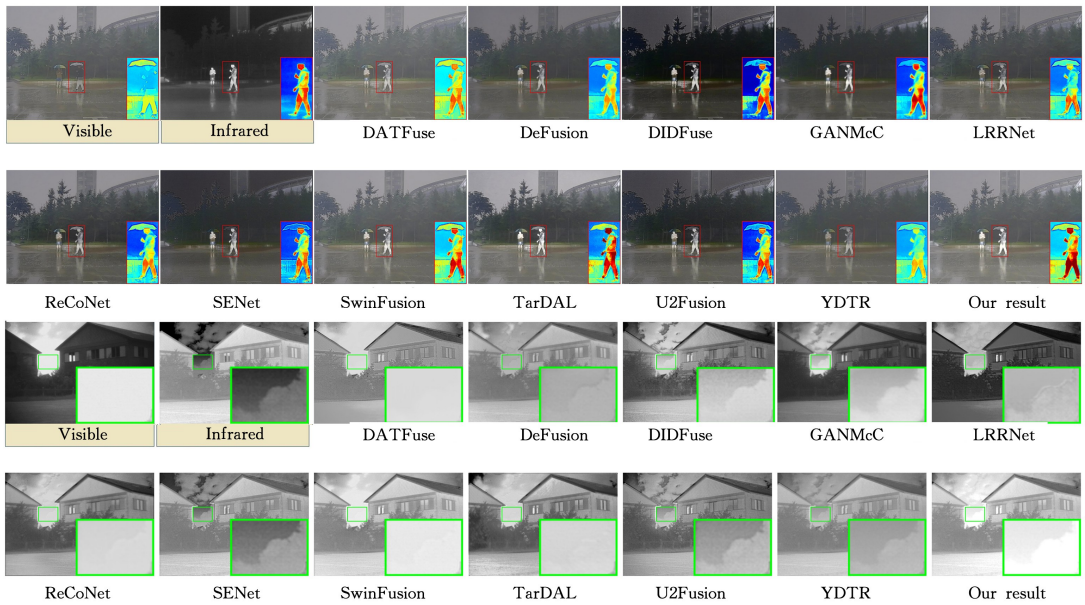


图 5 M3FD 和 TNO 数据集定性对比结果(电子版为彩图)

Fig. 5 Qualitative comparison results on M3FD and TNO dataset

红外图像和可见光图像之间存在显著的模态差异，因此图像融合的关键在于，如何在保留红外图像中不同亮度层次的关键信息的同时，确保可见光图像中颜色和纹理的真实性。通过在 MSRS 数据集上对比多种方法，发现除了 DeFusion，TarDAL，SDNet 和提出的方法外，其余方法均不同程度地出现了红外热特征丢失的现象。具体而言，SDNet 过度强调热辐射信息，导致图像整体亮度下降，影响了视觉效果；SwinFusion 与之类似，且目标显著性信息也有所欠缺。TarDAL

则过于注重亮度，致使背景颜色和纹理出现失真；DeFusion生成的图像整体偏暗，在保留低亮度场景细节方面的能力有限。与上述方法相比，提出的方法在保持行人目标与背景对比突出的同时，很好地保留了可见光图像的强度分布和纹理特征，实现了红外热特征信息与可见光图像视觉质量的平衡，展现出了优异的融合效果。

在M3FD数据集上，大部分方法不同程度地引入了红外背景噪声，导致图像整体对比度信息出现退化。尽管DATFuse和YDTR还原了来自可见光图像的背景信息，但它们缺失了红外图像中的显著目标信息，特别是边缘信息的损失较为严重。TarDAL将部分背景噪声引入融合图像，导致边缘不够显著，如雨伞部分。SwinFusion和本文方法在背景亮度方面表现相似，但本文方法使人物目标更加突出。在TNO数据集上，11种对比方法普遍存在天空亮度衰减或云层信息缺失的问题。具体而言，尽管SwinFusion和TarDAL在一定程度上保持了天空区域的亮度，但它们未能有效捕捉和重建红外图像中蕴含的云层信息。而MTFuse不仅能够充分保留红外图像对夜间背景微弱信息的刻画，同时还能准确地维持场景的整体亮度。

各种方法在不同数据集上的平均指标结果如表1和表2所列。在MSRS数据集上，如表1所列，MTFuse在MI，VIF，Qabf，SSIM和SF这5个指标上均达到了最优。这些指标的最佳结果表明，提出的MTFuse方法能够将源图像中最多的像素信息、结构信息和边缘信息有效地转移到融合图像中，同时生成具有最佳视觉保真度的图像。由表2可知，MTFuse在M3FD数据集上的MI，VIF，QABF，SSIM和PSNR指标均优于其他方法，而在TNO数据集上的MI，SSIM，PSNR指标也超过了其他比较方法。此外，MTFuse在所有数据集上的MI和SSIM指标均排名第一，这说明了MTFuse能够充分捕获和融合两种模态的互补特征。

表1 在MSRS数据集上不同方法的定量结果

Table 1 Average quantitative results of MSRS dataset for different methods

Method	MI	VIF	QABF	SSIM	SF	AG	PSNR
DIDFuse[26]	1.61	0.31	0.2	0.24	9.64	2.00	64.14
U2Fusion[27]	1.35	0.51	0.39	0.71	9.08	2.87	66.36
GANMcC[28]	1.66	0.61	0.32	0.76	5.92	2.21	66.01
SDNet[29]	1.14	0.48	0.36	0.66	8.69	2.72	64.81
SwinFusion[9]	1.53	0.66	0.55	0.87	10.80	3.52	66.95
YDTR[30]	1.82	0.54	0.32	0.65	7.40	2.23	64.10
TarDAL[31]	1.49	0.42	0.18	0.47	9.69	3.04	**61.07**
DeFusion[32]	2.16	0.77	0.54	0.94	7.98	2.61	66.05
ReCoNet[33]	2.16	0.71	0.50	0.85	9.98	3.00	64.51
DATFuse[34]	2.70	0.91	0.64	0.91	10.93	**3.58**	63.27
LRRNet[35]	2.03	0.54	0.45	0.43	8.47	2.65	64.74
MTFuse	**4.62**	**1.01**	**0.67**	**1.03**	**11.39**	3.71	64.36

表2 在M3FD和TNO数据集上不同方法的定量结果

Table 2 Average quantitative results of M3FD and TNO dataset for different methods

Method	M3FD							TNO						
	MI	VIF	QABF	SSIM	SF	AG	PSNR	MI	VIF	QABF	SSIM	SF	AG	PSNR
DIDFuse[26]	2.14	0.70	0.50	0.83	**14.08**	4.88	61.93	2.35	0.60	0.40	0.87	11.28	4.29	61.75
U2Fusion[27]	1.9	0.66	0.54	0.92	13.38	**4.90**	63.56	1.84	0.61	0.42	0.86	11.63	4.94	62.75
GANMcC[28]	1.93	0.53	0.31	0.82	7.80	2.71	63.13	1.53	0.54	0.31	0.88	6.56	2.78	62.03
SDNet[29]	2.22	0.56	0.51	0.93	13.88	4.77	62.98	1.55	0.57	0.42	0.97	11.55	4.62	62.15
SwinFusion[9]	2.46	0.57	0.50	0.7	13.92	4.60	61.15	2.25	**0.73**	**0.51**	1.01	10.59	4.19	61.12
YDTR[30]	2.17	0.61	0.47	0.91	10.25	3.33	62.69	1.73	0.63	0.42	0.99	7.37	2.74	62.67
TarDAL[31]	2.38	0.54	0.29	0.87	12.60	4.22	62.39	1.86	0.53	0.32	0.88	10.30	3.85	62.51
DeFusion[32]	2.32	0.65	0.44	0.94	7.46	2.60	63.48	1.78	0.6	0.41	0.96	6.17	2.55	63.40
ReCoNet[33]	2.11	0.59	0.49	0.86	10.55	3.93	62.71	1.78	0.57	0.39	0.88	7.24	3.19	62.92
DATFuse[34]	2.79	0.62	0.48	0.88	10.65	3.48	61.24	2.21	0.68	0.37	0.75	**15.05**	**5.34**	61.60
LRRNet[35]	1.92	0.55	0.48	0.77	10.92	3.63	62.91	1.95	0.51	0.31	0.6	11.72	4.35	61.06
MTFuse	**3.42**	**0.72**	**0.56**	**1.00**	13.61	4.56	**60.43**	**2.68**	0.70	0.49	**1.03**	11.39	4.20	**60.13**

表3是不同方法的运行效率对比，虽然ReCoNet的运算量和参数量最低，但这是通过牺牲融合效果实现的，该方法在所有数据集和指标上均未取得最佳表现。和其他基于Transformer的方法相比，MTFuse的优势在于运算量和参数量较小，如SwinFusion，YDTR和DeFusion的运算量分别是本文方法的13倍、4倍和3倍。同时，SwinFusion和DeFusion的参数量分别是本文方法的2.5倍和21倍。尽管如此，MTFuse在参数量和运算量较低的情况下，依然在不同数据集上实现了定性和定量的最优性能，这证明了MTFuse在捕获图像长程依赖关系方面的高效性和有效性。

表3 不同方法的计算复杂度和参数量对比

Table 3 Comparison of computational complexity and parameter count among different methods

Method	Flops	Para.
DIDFuse[26]	24.503×10^9	0.261×10^6
U2Fusion[27]	43.170×10^9	0.659×10^6
GANMcC[28]	122.230×10^9	1.864×10^6
SDNet[29]	8.813×10^9	0.067×10^6
SwinFusion[9]	63.731×10^9	0.927×10^6
YDTR[30]	20.581×10^9	0.107×10^6
TarDAL[31]	19.443×10^9	0.297×10^6
DeFusion[32]	15.265×10^9	7.874×10^6
ReCoNet[33]	$\mathbf{0.347\times10^9}$	$\mathbf{0.002\times10^6}$
DATFuse[34]	1.185×10^9	0.011×10^6
LRRNet[35]	1.511×10^9	0.049×10^6
MTFuse	5.031×10^9	0.371×10^6

4.4 消融实验

本节旨在通过消融实验验证所提出不同网络结构的有效性，共提出了 3 个消融实验，如表 4 所列，分别是去掉 RMB 中的 Mamba；使用通道数减半的卷积的融合层替代 CQAM；用 Swin 块代替 Mamba，并保持相同的隐藏层维度和结构。

实验结果表明，除 PSNR 外，所有指标在这些降级版本中均出现了显著下降。这一结果有力地证明了 MTFuse 中各个组件的设计对于提高融合质量的重要性。值得注意的是，用 Swin 块代替 Mamba 后，虽然模型参数量略有减少，但模型的运算效率却下降了 3 倍以上，同时各项评估指标也出现了不同程度的下降。这一现象充分彰显了 Mamba 在图像融合任务中的关键作用，即对全局互补信息进行高效建模和利用，从而显著提升了模型的计算效率和融合性能。

表 4 消融实验的平均指标

Table 4 Comparison of the proposed method and its degraded versions

Experiment	VIF	MI	QABF	SSIM	SF	AG	PSNR	Flops	Para.
w/o Mamba	1.0307	4.6021	0.6692	1.0036	10.7184	3.3659	**63.1519**	**3.949×10^9**	**0.1069×10^6**
CQAM→CNN	0.9849	2.9345	0.6079	0.8977	11.0682	3.6675	64.1545	4.5136×10^9	0.3633×10^6
Mamba→Swin	1.0134	4.5988	0.6732	1.0002	11.2102	3.6004	64.4655	15.9751×10^9	0.2440×10^6
MTFuse	**1.0324**	**4.6190**	**0.6734**	**1.0103**	**11.3887**	**3.7057**	64.3583	5.0305×10^9	0.3713×10^6

消融实验的定性对比结果如图 6 所示，除本文方法外，其他所有降级版本在红外显著性信息的捕获和保留方面均存在不同程度的缺失。这说明 MTFuse 每一个组件都起着重要作用，最终实现了对红外微弱热信号的精准感知和完整保留。

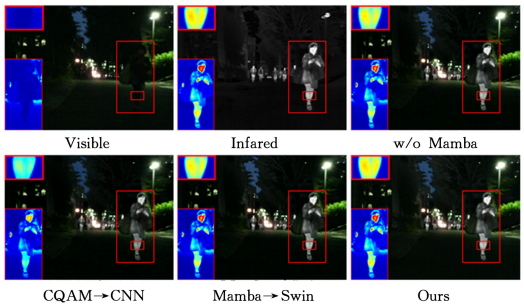


图 6 消融实验的定性对比结果

Fig. 6 Qualitative comparison results of ablation study

结束语 本文提出了一种基于 Mamba 和 Transformer 的红外与可见光图像融合网络 MTFuse。为了高效建模图像远程依赖，设计了一个残差选择性结构化状态空间模块(RMB)提取全局特征。同时，为了对多模态图像之间的关系进行建模，设计了一个跨模态查询融合注意力模块(CQAM)用于特征的自适应融合。此外，设计了一个由两个项组成的损失函数，包括梯度损失和亮度损失，旨在以无监督的方式训练所提出的模型。与大量其他先进的方法在融合质量的对比实验和消融实验证明了所提出的 MTFuse 方法的有效性。但本文目前只对红外与可见光图像进行讨论，如何将该方法拓展到其他融合任务上亟需进一步讨论。在未来的工作中，我们计划在更多的图像融合场景下，系统地评估和优化 MTFuse 的性能，并探索其与领域知识和先验信息相结合的可能性，以期进一步提升其泛化能力和实用价值。

参 考 文 献

[1] CHEN H，DENG L，ZHU L，et al. ECFuse：Edge-Consistent and Correlation-Driven Fusion Framework for Infrared and Visible Image Fusion [J]. Sensors，2023，23(19)：8071.

[2] KAUR H，KOUNDAL D，KADYAN V. Image fusion techniques：a survey [J]. Archives of Computational Methods in Engineering，2021，28(7)：4425-4447.

[3] ZHAO W，XIE S，ZHAO F，et al. Metafusion：Infrared and visible image fusion via meta-feature embedding from object detection[C]// Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[4] ZHAO Z，XU S，ZHANG J，et al. Efficient and model-based infrared and visible image fusion via algorithm unrolling [J]. IEEE Transactions on Circuits and Systems for Video Technology，2021，32(3)：1186-1196.

[5] MA J，MA Y，LI C. Infrared and visible image fusion methods and applications：A survey [J]. Information Fusion，2019，45：153-178.

[6] TANG W，LIU Y，CHENG J，et al. A phase congruency-based green fluorescent protein and phase contrast image fusion method in nonsubsampled shearlet transform domain [J]. Microscopy Research and Technique，2020，83(10)：1225-1234.

[7] ZHANG Q，LIU Y，BLUM R S，et al. Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images：A review [J]. Information Fusion，2018，40：57-75.

[8] KONG W，LEI Y，ZHAO H. Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization [J]. Infrared Physics & Technology，2014，67：161-172.

[9] MA J，TANG L，FAN F，et al. SwinFusion：Cross-domain long-range learning for general image fusion via swin transformer [J]. IEEE/CAA Journal of Automatica Sinica，2022，9(7)：1200-1217.

[10] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need [C]// Advances in Neural Information Processing Systems. 2017.

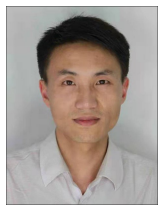
[11] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An image is worth 16x16 words：Transformers for image recognition at scale [J]. arXiv：2010.11929，2020.

[12] LIU Z，LIN Y，CAO Y，et al. Swin transformer：Hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[13] ZAMIR S W，ARORA A，KHAN S，et al. Restormer：Efficient transformer for high-resolution image restoration[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and

Pattern Recognition. 2022.
[14] LU J, BATRA D, PARIKH D, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks [C]// Advances in Neural Information Processing systems. 2019.
[15] SUN Y, DONG L, HUANG S, et al. Retentive network: A successor to transformer for large language models [J]. arXiv: 2307.08621, 2023.
[16] GU A, DAO T. Mamba: Linear-time sequence modeling with selective state spaces [J]. arXiv: 2312.00752, 2023.
[17] LIU Y, TIAN Y, ZHAO Y, et al. Vmamba: Visual state space model [J]. arXiv: 2401.10166, 2024.
[18] HAMILTON J D. State-space models [J]. Handbook of Econometrics, 1994, 4: 3039-3080.
[19] ZHAO D, SHU X, ZHANG L, et al. Sensor interrogation technique using chirped fibre grating based Sagnac loop [J]. Electronics Letters, 2002, 38(7): 312-313.
[20] HAN Y, CAI Y, CAO Y, et al. A new image fusion performance metric based on visual information fidelity [J]. Information Fusion, 2013, 14(2): 127-135.
[21] XYDEAS C S, PETROVIC V. Objective image fusion performance measure [J]. Electronics Letters, 2000, 36(4): 308-309.
[22] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity [J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.
[23] ESKICIOGLU A M, FISHER P S. Image quality measures and their performance [J]. IEEE Transactions on Communications, 1995, 43(12): 2959-2965.
[24] CUI G, FENG H, XU Z, et al. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition [J]. Optics Communications, 2015, 341: 199-209.
[25] JAGALINGAM P, HEGDE A V. A review of quality metrics for fused image [J]. Aquatic Procedia, 2015, 4: 133-142.
[26] ZHAO Z, XU S, ZHANG C, et al. DIDFuse: Deep image decomposition for infrared and visible image fusion [J]. arXiv: 2003.09210, 2020.
[27] XU H, MA J, JIANG J, et al. U2Fusion: A unified unsupervised image fusion network [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(1): 502-518.
[28] MA J, ZHANG H, SHAO Z, et al. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion [J]. IEEE Transactions on Instrumentation and Measurement, 2020, 70: 1-14.
[29] ZHANG H, MA J. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion [J]. International Journal of Computer Vision, 2021, 129(10): 2761-2785.
[30] TANG W, HE F, LIU Y. YDTR: Infrared and visible image fusion via Y-shape dynamic transformer [J]. IEEE Transactions on Multimedia, 2022, 25: 5413-5428.
[31] LIU J, FAN X, HUANG Z, 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. 2022.
[32] LIANG P, JIANG J, LIU X, et al. Fusion from decomposition: A self-supervised decomposition approach for image fusion[C]// European Conference on Computer Vision. Springer, 2022.
[33] HUANG Z, LIU J, FAN X, et al. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion[C]// European Conference on Computer Vision. Springer, 2022.
[34] TANG W, HE F, LIU Y, 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.
[35] LI H, XU T, WU X J, et al. LRRNet: A novel representation learning guided fusion network for infrared and visible images [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 11040-11052.

DING Zhengze, born in 2000, postgraduate. His main research interests include deep learning and image fusion.



NIE Rencan, born in 1982, Ph.D, professor, doctoral supervisor. His main research interests include neural networks, image processing and machine learning.

(责任编辑:喻藜)