引用格式：WU Xiaocong, FENG Xin, HUANG Meina, et al. A Method for Infrared and Visible Image Fusion Based on CNN-Mamba Feature Extraction Combined with Information Selection[J]. Acta Photonica Sinica, 2026, 55(1):0110001
仵晓聪,冯鑫,黄美娜,等. 基于CNN-Mamba特征提取结合信息选择的红外与可见光图像融合方法[J].光子学报,2026,55(1):0110001

基于CNN-Mamba特征提取结合信息选择的红外与可见光图像融合方法

仵晓聪[1],冯鑫[1],黄美娜[1],赵彬[1],胡开群[1],沈瑜[2]
(1 重庆工商大学 机械工程学院 智能装备绿色设计与制造重庆市重点实验室, 重庆 400067)
(2 兰州交通大学 电子与信息工程学院, 兰州 730070)

摘 要:提出了一种基于CNN-Mamba特征提取结合信息选择的红外与可见光图像融合方法,旨在克服传统融合策略难以面对复杂场景和基于深度学习网络的融合策略复杂的问题。在图像编码阶段,引入了空间上下文感知模块(SCAM)来整合上下文特征信息并分配注意力权重,接着设计了一种CNN结合Mamba网络的特征提取模块(CMFE)对两种模态图像信息进行深层次提取。在融合阶段,提出了全新的自适应权重选择融合层以保留更多源信息。通过消融实验结果可以看出本文提出的网络结构与损失函数具有合理性。在对比实验中,本文在MSRS、Road-Scene、TNO和M3FD四个公开数据集上将所提出的方法与七种先进融合方法进行对比,结果显示所提方法不仅具有更好的主观视觉效果,还具有最好的客观评价指标。最后为验证本文所提出方法的后续应用能力,针对Road-Scene数据集中的行人和车辆进行目标检测任务评估,结果表明本文方法在行人与车辆检测方面均取得最多的最优指标,这充分验证了本文所提出算法的综合性能。

关键词:图像融合;归一化互信息;CMFE模块;信息熵;自适应权重选择融合层
中图分类号:TP391 **文献标识码**:A **doi**:10.3788/gzxb20265501.0110001

0 引言

图像融合旨在整合来源于不同成像条件下获取的多种影像数据,通过信息交互以弥补单一成像传感器的局限性,并获得更高质量的综合影像信息[1]。其中,红外与可见光图像融合是当前研究的热点之一。红外传感器通过捕获物体表面的热辐射进行成像,能有效反映物体的热量信息。然而红外图像易受拍摄噪声影响,且难以呈现丰富的物体细节和纹理信息[2]。相比之下,可见光传感器通过收集物体表面的反射光成像,能够提供符合人类视觉习惯的丰富细节和纹理信息。但其成像质量易受光照等客观条件影响,可能导致成像不完全[3]。因此如何有效融合红外与可见光图像,以获得细节丰富、纹理清晰且成像完整的融合结果,是一项具有重要意义的研究任务[4]。

目前图像融合主要分为传统图像融合和基于深度学习图像融合。传统图像融合算法主要有基于多尺度变换的方法[5]、基于稀疏表示的方法[6]、基于子空间的方法[7]、基于显著性的方法[8]和基于混合模型的方法[9]等。尽管这些传统方法在一定程度上取得了进展,但它们普遍存在特征提取不充分、融合策略过度依赖手动设计规则以及在复杂场景下融合性能欠佳等局限性[10]。这些问题随着深度学习的发展得到了有效解决。

基于深度学习图像融合主要分为:基于自编码器(Auto Encoder, AE)融合方法、卷积神经

基金项目:国家自然科学基金(62241106,61861025),甘肃省重点研发计划(24YFGA037),国家杰出青年科学基金(42325502)。
第一作者:仵晓聪,2023311001@ctbu.edu.cn
通讯作者:冯鑫,149495263@qq.com
收稿日期:2025-07-23;**录用日期**:2025-09-22

(Convolutional Neural Network, CNN)网络融合方法、生成式对抗网络(Generative Adversarial Network, GAN)融合方法、Transformer融合方法。其中,LI Hui等[11]提出了DenseFuse网络,采用AE方式利用DenseNet提取深层次特征,取得了不错的效果。但是DenseNet网络的每一层输出都会与后面一层的输入进行拼接,导致计算量十分巨大。此外,该方法使用红外与可见光直接做加法的融合策略,在面对较为复杂的场景是表现不佳。ZHANG Yu等[12]提出了一种通用卷积神经网络图像融合框架(Image Fusion Framework based on Convolutional Neural Network, IFCNN)融合网络,将图像融合任务表述为一个全卷积神经网络,从而实现端到端的训练。IFCNN网络采用逐元素融合规则来融合特征,这些规则相对简单且固定在面对某些复杂的场景时难以有效融合两种模态信息。MA Jiayi等[13]提出了FusionGAN,它通过生成器和鉴别器之间的对抗博弈来实现图像融合,旨在同时保留红外图像中的热辐射信息和可见光图像中的纹理细节。该方法生成了具有较好视觉效果和细节保留的融合图像。但是GAN网络存在损失函数难以收敛、训练时间较长等问题。近年来,Transformer模型成为研究热点。例如,WANG Zhishe等[14]提出SwinFuse融合网络,该方法主要利用SwinTransformer来提取全局特征,并设计基于行向量和列向量维度的融合策略。TANG Wei等[15]提出一种双注意力残差融合网络DATFuse。该网络通过设计双注意力残差模块精确捕捉源图像关键区域特征,采用Transformer模块建模长程依赖关系以保留全局互补信息。尽管这些基于Transformer深度学习的融合方法取得了较好的效果,但是其融合网络所设计的基于深度学习网络的融合策略过于复杂,需要大量训练时间。

针对传统融合策略在面对复杂场景时融合效果不佳且基于学习网络的融合策略通常过于复杂的问题,本文提出一种基于CNN-Mamba特征提取结合信息选择的红外与可见光图像融合方法。在编码阶段,引入空间上下文感知模块SCAM对全局上下文信息的整合和线性变换,进一步提高边缘、纹理信息与背景之间关系。随后设计了一种具有局部性和长距离依赖性的CMFE模块用于两种模态图像的特征提取。最后,本文还根据两种模态图像之间的归一化互信息提出了一种全新的自适应权重选择融合层,依据归一化互信息的大小关系设计了对应的融合规则,以应对不同场景下的信息交互,从而最大限度地保留源图像信息。在解码阶段设计了由四个CNN构成的解码器,通过逐步减少通道数量来确保信息不会丢失。通过大量的消融实验、对比实验以及后续任务实验证明,本文所提出的方法具有较好的综合性能。

1 融合网络结构

本文提出的基于CNN-Mamba特征提取结合信息选择的红外与可见光图像融合网络主要以自编码器为基本框架。相较于Transformer来说Mamba网络具有更低的复杂度,能够显著降低运算时间。所以本网络主要设计了一种改进的VMamba模块进行特征提取,并且提出了一种全新的自适应权重选择的融合规则。整个融合网络的具体结构如图1所示。

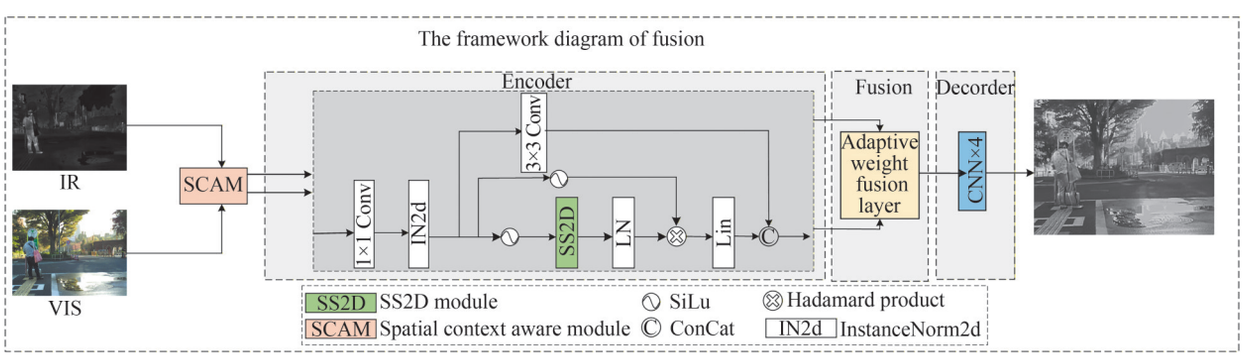


图1 本文所提出网络的主要结构

Fig. 1 The main structural diagram of the network proposed in the article

本文提出的红外与可见光图像融合网络主要包含三个阶段:图像编码阶段、图像融合阶段和图像解码阶段。在图像编码阶段,引入的SCAM模块旨在卷积特征图上同时建模通道注意力和空间注意力,并融合

两者信息以增强高频与低频信息之间的关联，以获取更加显著的图像特征。接着设计了用于特征提取的CMFE模块，该步骤的目的是在局部和大范围感受野内充分从源图像中提取特征并将其存储到潜空间。在图像融合阶段，通过对特征提取得到的潜空间信息计算信息熵以及互信息。若归一化互信息值大于某设定阈值则说明两种模态信息高度相似，融合权重各取一半。若归一化互信息小于某设定阈值，说明两种模态信息差距较大，信息熵更高的模态将获得更大的权重，以确保其包含的更多信息得以保留。在图像解码阶段，设计了由三个CNN层构成的解码器，通过逐层卷积与非线性激活，将多模态融合后的高维特征逐步压缩至单通道输出。在保持空间分辨率不变的同时，逐级减少通道数量，实现了特征容量与空间细节的平衡。该过程不仅有效去除冗余信息，还保留了最具代表性的结构与细节，从而生成紧凑且高质量的融合图像表征。通过这种方式，解码器能够逐渐剔除冗余或模态间的重复信息，保留更具代表性和判别性的特征表达。最终在 Tanh 激活的作用下，输出的单通道图像具备更为紧凑且有效的信息表征，最终生成融合图像。

1.1 SCAM模块

SCAM模块旨在利用全局空间上下文特征信息，同时建模通道注意力和空间注意力，从而提升对边缘、纹理信息等细节的检测能力[16]。该模块通常主要用于遥感图像目标检测任务中，可以提升细小信息与背景信息联系。本文首次将该模块引入到图像融合任务中，而该模块在融合任务中可以进一步提升后续特征提取的质量。该模块的具体结构如图2所示。

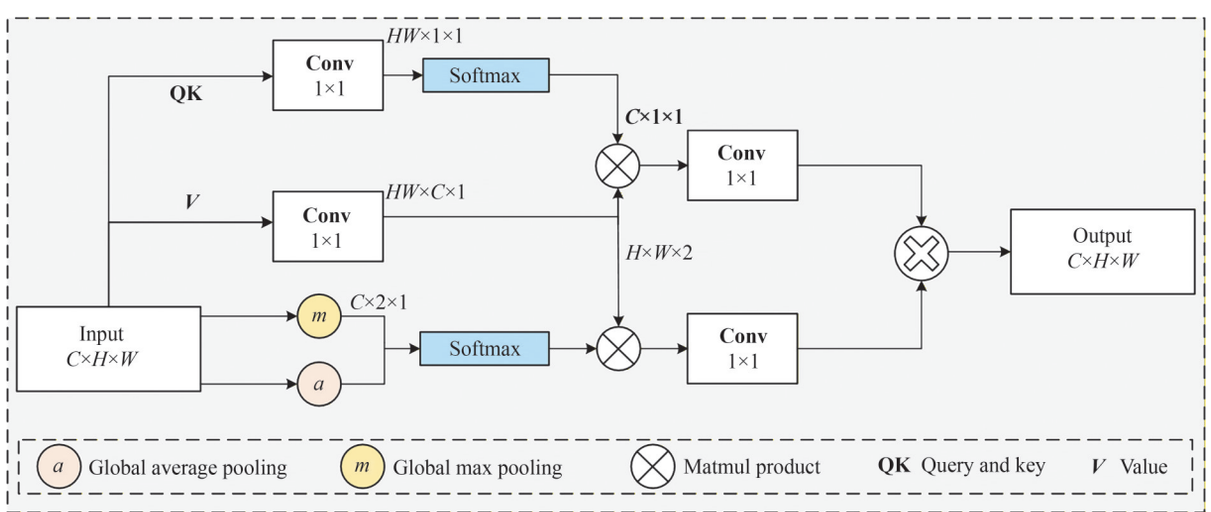


图2 SCAM模块结构图

Fig. 2 The structure diagram of SCAM module

SCAM模块由三个主要分支构成。第一个分支分别进行全局平均池化和全局最大池化操作，接着结合Softmax函数用于整合整体信息。第二分支针对每个Token的实际特征Value使用1×1Conv卷积生成特征图并求得线性变换结果。第三个分支使用1×1Conv卷积来简化查询和键的倍数。然后，将第二分支分别与第一分支和第三分支矩阵相乘，得到的结果分别表示跨通道和跨空间的上下文信息。最后，针对这两个分支进行Hadamard积操作得到最终的输出。这一过程可以表示为

$$Q_i^j = P_i^j + a_i^j \sum_{j=1}^{N_i} \left[\frac{\exp\left(\boldsymbol{\omega}_{qk} P_i^j\right)}{\sum_{n=1}^{N_i} \exp\left(\boldsymbol{\omega}_{qk} P_i^n\right)} \cdot \boldsymbol{\omega}_v P_i^j \right] \tag{1}$$

$$a_i^j = \frac{\exp\left(\left[\mathrm{avg}(P_i);\ \max(P_i)\right] P_i^j\right)}{\sum_{n=1}^{N_i} \exp\left(\left[\mathrm{avg}(P_i);\ \max(P_i)\right] P_i^n\right)} \cdot \boldsymbol{\omega}_v \tag{2}$$

式中，p_i^j和Q_i^j分别表示第i级特征图中第j个像素的输入和输出。N_i表示像素的总数。$\boldsymbol{\omega}_{qk}$和$\boldsymbol{\omega}_v$是用于投影特征图的线性变换矩阵，其通过1×1 Conv卷积来简化。avg(•)和max(•)分别表示全局平均池化和全局最

大池化。在通道信息提取方面，SCAM模块首先通过一个1×1 Conv卷积生成特征映射**QK**，并将其变换为[C, $H\times W$]的形式。同时，另一条分支通过1×1 Conv卷积卷积得到特征V，该特征被压缩为单通道并经Softmax归一化，从而形成空间维度上的权重分布。随后将V与**QK**进行矩阵相乘，以实现对各通道响应的加权，从而评估不同通道的重要性即在关键空间位置具有更强响应的通道将获得更高的权重。在空间信息提取方面，模块利用全局平均池化和全局最大池化获取全局通道描述，并将其与V相结合，生成两张空间注意力图，用于刻画全局统计下更值得关注的空间位置。接着将两张注意力图进行拼接，并通过卷积操作得到最终的空间注意力表示。SCAM模块将前向过程划分为“通道路径”和“空间路径”两条支路，通过逐像素门控机制实现通道与空间特征的联合建模，并采用残差连接回输入特征。这一设计能够有效引导特征映射在通道维度上选择更具判别力的信息，使得SCAM模块能够学习到富含上下文的通道特征表征。

1.2 CMFE模块设计

为了更好进行待融合图像的特征提取，本文设计了可以在局部和大范围感受野内更充分提取特征的CMFE模块。该模块以VMamba模块[17]为基础，设计了一个由CNN与SiLu激活函数构成的额外分支，从而使CMFE模块能够更好地捕捉图像中的长短序列中重要特征区域。模块具体结构如图1所示。

设可见光图像和红外图像分别为I_1和I_2。首先，为了对图像中的噪声进行平滑处理并为后续处理提供充足的低级特征表示，本文应用1×1卷积模块使得输入信息通道数改变为64。接着使用InstanceNorm2d(IN2d)模块进行归一化操作，以保持每个样本的特征独立性，从而更好地捕捉和转换图像的风格。输出记作y_i^C，这个过程表示为

$$y_i^C = \mathrm{IN2d}\left(\mathrm{Conv}\left(I_i\right)\right) \tag{3}$$

式中，IN2d表示InstanceNorm2d ，Conv表示1×1卷积。i=1时代表可见光图像，i=2时表示红外光图像。接着y_i^C分别通过两个支路，第一条支路由3×3Conv卷积和LeakyReLU激活函数构成。卷积网络能够聚焦于邻域内像素关系，对边缘、角点和纹理等低级特征比较敏感，从而有效提取局部特征。第一条支路过程可以表示为

$$y_i^a = \mathrm{LeakyReLU}\left(\mathrm{Conv}\left(y_i^C\right)\right) \tag{4}$$

第二条支路由SS2D模块构成。SS2D源自于VMamba中选择性扫描模块，主要用于维度上捕获跨越长距离的依赖关系，以克服传统卷积难以覆大的感受野的缺点。这种分支设计使得模型能够以不同的方式处理局部特征和全局依赖。在面对复杂场景时，局部特征与全局依赖的共同作用能够更有效地分辨真实目标，从而提高模型的性能。这一过程可以表示为

$$y_i^b = \mathrm{Lin}\left(\mathrm{LN}\left(\mathrm{SS\,2D}\left(y_i^C\right)\right)\right) \tag{5}$$

式中，LN表示LayerNormal归一化层，Lin表示线性层。在经过SCAM模块得到模态特征增强后，CMFE模块进一步承担跨模态特征提取的作用。它一方面通过局部卷积分支强化可见光图像的细节表达，另一方面依托全局扫描分支确保红外与可见光目标和结构的连续性，从而实现细节保留与全局一致性的统一。最后，将分支一和分支二用Concat方式连接。这一步骤表示为

$$y_i^o = \mathrm{Concat}\left(y_i^a, y_i^b\right) \tag{6}$$

式中，y_i^o表示CMFE模块的输出结果。

1.3 自适应权重选择融合层

在红外与可见光图像融合任务中，红外模态能够捕获目标的热辐射分布，在黑暗、烟尘或逆光场景下呈现出高对比度目标轮廓，但缺乏纹理与颜色线索；可见光模态能够展现边缘、纹理和语义细节，却对光照条件敏感。逐像素相加、通道级联或固定权重加权等传统融合策略往往采用人工设定或端到端网络隐式学习单一融合规则，在面对场景复杂和模态间信息分布差异较大的情况时往往无法实现有效的信息融合，最终导致在复杂场景下出现过曝和伪影等问题。近年来，基于端到端的学习网络被广泛引入图像融合规则的设计中，尽管提升了融合效果，但同时也显著增加了计算负担和资源消耗[18]。基于以上原因，本文提出的基于归一化互信息的自适应选择融合层既保持了深度特征表达的丰富性，又兼顾了高计算效率和跨场景泛化能力。该融合层的设计思路是：在特征提取后分别计算两种模态的熵EN_i(i=1, 2, 当i=1时表示可见光图像

信息熵，当i=2时表示红外光图像信息熵），用于衡量各自的信息量。图3中，不同区域的两模态信息存在显著差异（蓝色框）或较高相似性（红色框）。针对这一特点，本文通过计算两模态间的归一化互信息来评估其局部相似性，并据此自适应分配两种模态各自融合权重，从而在融合结果中最大限度保留红外与可见光图像的有效信息。

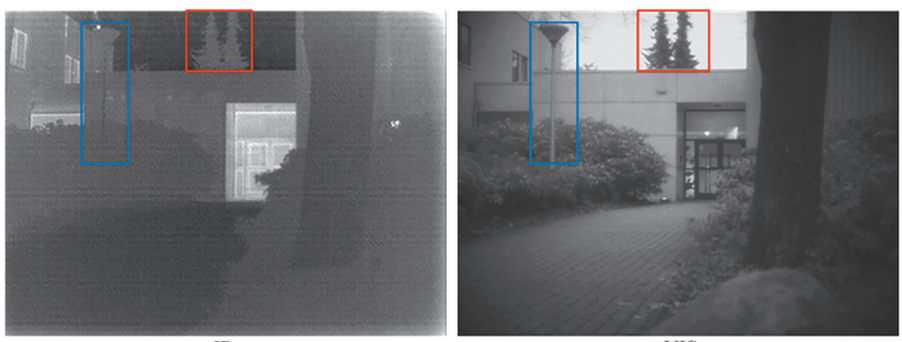


图3　红外与可见光图像直观比较

Fig. 3　Intuitive comparison of infrared and visible images

在概率论和信息论中，两个随机变量的互信息是变量间相互依赖性的量度[19]。互信息值越大表示两幅图像的相似程度越高。为了便于统计通常将互信息归一化，信息熵与互信息的计算公式为

$$\mathrm{EN}=-\sum_{i=0}^{L-1}p_l\log_2 p_l \tag{7}$$

$$\mathrm{MI}(I_1,I_2)=\mathrm{EN}_{I_1}+\mathrm{EN}_{I_2}-\mathrm{EN}(I_1,I_2) \tag{8}$$

则归一化互信息公式为

$$\mathrm{NMI}(I_1,I_2)=\frac{\mathrm{MI}(I_1,I_2)}{\sqrt{\mathrm{EN}_{I_1}\cdot\mathrm{EN}_{I_2}}} \tag{9}$$

式中，L表示灰度级数，p_l表示融合图像中对应灰度级的归一化直方图。I_1表示可见光模态信息，I_2表示红外光模态信息。根据文献[20]，本文设定当归一化互信息NMI(I_1, I_2)>0.8时即可认为两种模态信息高度相似，当归一化互信息NMI(I_1, I_2)<0.8时则认为两种模态信息差距较大。特征融合结果则表示为

$$F=\omega_1\cdot y_1^o+\omega_2\cdot y_2^o \tag{10}$$

式中，ω_1、ω_2分别表示可见光模态和红外光模态信息权重。由于红外模态主要包含低频信息，可见光模态主要包含细节和纹理信息。因此，本文设定融合规则如下：当NMI(I_1, I_2)>0.8时，则$\omega_1=\omega_2=0.5$。当NMI(I_1, I_2)<0.8时，为了确保更多具有高信息熵的模态信息得以保留，信息熵更高的模态将分配更大的权重。为了确保两种模态的信息均保留在融合结果中，权重分配系数限定为$0<\omega_i<1$。这一自适应权重选择过程可以表示为

$$\omega_i=\frac{\max(\mathrm{EN}_1,\mathrm{EN}_2)}{\sum_{i=1}^{2}\mathrm{EN}_i} \tag{11}$$

式中，ω_i表示高信息熵模态所对应权重，低信息熵模态所对应权重为1-ω_i。传统融合算法采用两种模态固定权重取值的方式，无法区分不同区域的重要性，因此在复杂纹理或高对比度场景下往往表现不佳。本文提出的自适应权重选择融合层通过实时计算两种模态的信息熵和归一化互信息，动态调整融合权重，实现对不同区域的显式强化与侧重。与基于深度学习的融合策略如残差融合[21]、密集融合[22]和残差密集融合等[23]相比，本方法无需大规模标注即可适应多场景，且仅在现有卷积特征后增加轻量级的熵与互信息计算，显著减少了计算开销，提升了融合网络的泛化能力和实时性。

1.3　损失函数

图像的强度、梯度以及结构相似性都会对融合结果产生显著影响。为了最大限度地提升融合结果的质量，本文所设计的损失函数L_f主要由强度损失L_in、梯度损失L_g和结构相似度损失L_SSIM三部分构成。总的损

失函数L_f表达式为

$$L_f=\beta_1 L_g+\beta_2 L_{in}+\beta_3 L_{SSIM} \tag{12}$$

式中，β_1、β_2、β_3分别表示超参数。

梯度损失L_g主要用于衡量融合图像的梯度与可见光图像和红外图像梯度之间的差异，进而指导模型生成具有更丰富边缘和纹理信息的图像。梯度损失L_g表达式为

$$L_{grad}=\frac{1}{CHW}\left\|\nabla I_{fuse}-\max\left(\nabla I_{vi},\nabla \mathrm{Bila}\left(I_{ir}\right)\right)\right\|_1 \tag{13}$$

式中，I_{fuse}表示融合图像结果，C表示通道数、H和W分别表示表示图像高度以及宽度，∇表示梯度算子，本文使用Laplace算子作为梯度算子，Bila为双边滤波算子。

强度损失L_{in}主要作用是确保输出图像在整体亮度和纹理细节上尽可能接近原始图像，从而不丢失关键的强度信息。强度损失L_{in}其表达式为

$$L_{in}=\frac{1}{CHW}\left\|I^c_{Fuse}(i,j)-\max\left(I^c_{vi}(i,j),I^c_{ir}(i,j)\right)\right\|_1 \tag{14}$$

结构相似度损失L_{SSIM}是图像的一种视觉感知度量，旨在改善传统方法无法充分反映人类视觉感知 特性的问题，该损失更加注重图像的结构和感知特性，特别符合人类视觉系统特性。结构相似度损失L_{SSIM}表达式为

$$\mathrm{SSIM}_{(x,y)}\left(I,I_f\right)=\frac{\left(2\mu_{I(x,y)}\mu_{I_f(x,y)}+c_1\right)\left(2\sigma_{I(x,y)}\sigma_{I_f(x,y)}+c_2\right)}{\left(\mu^2_{I(x,y)}+\mu^2_{I_f(x,y)}+c_1\right)\left(\sigma^2_{I(x,y)}+\sigma^2_{I_f(x,y)}+c_2\right)} \tag{15}$$

$$L_{SSIM}=1-\mathrm{SSIM}_{(x,y)}\left(I_I,I_f\right) \tag{16}$$

式中，$\mu_{I(x,y)}$和$\mu_{If(x,y)}$分别表示源图像与融合图像的平均值，$\sigma_{I(x,y)}$和$\sigma_{If(x,y)}$分别表示源图像与融合图像的方差，$\sigma_{I(x,y)}\sigma_{If(x,y)}$表示源图像与融合图像之间的协方差，$c_1$和$c_2$为起调节作用的超参数。

2 实验与分析

2.1 实验设置与数据集

为了充分展示本文所提出的基于CNN-Mamba特征提取结合信息选择的红外与可见光图像融合网络的优越性，主要在MSRS[24]、TNO[25]、Road-Scene[26]和M3FD[27]四种具有代表性的红外与可见光图像数据上进行消融实验和对比实验。为了验证融合网络的泛化能力，本文在MSRS数据集中选择1 083对图像进行训练和验证，训练集和验证集图像比例为4∶1，测试集数据图像为361对，其中Road-Scene、TNO和M3FD数据集全部用于测试。在训练前，本文将所有红外与可见光数据集图像调整为大小224×224的灰度图像，并在融合之后将结果图像还原。考虑到三种数据集数据量差异，本文在TNO中选择了5组结果、Road-Scene中选择了20组结果、M3FD中选择了20组结果以及MSRS中选择了50组结果用于最终的测试实验结果展示与分析。

本次实验的主要硬件平台为：NVIDIA GTX 4090 GPU 24 G显卡，Intel(R) Xeon(R) Platinum 8481C型号CPU，软件平台为VScode。网络训练的主要参数为：Batchsize设置为16、epoch为150、Adam优化策略以及学习率为$1e^{-4}$。此外，每经过50个训练轮次学习率衰减为原来的0.5倍。损失函数超参数分别为$\beta_1=1$，$\beta_2=10$，$\beta_3=1$。

图像融合通常属于无监督学习任务。因此，其性能评估主要依赖于主观视觉观察和客观指标对比[28]。主观观察直接反映图像对人的视觉效果，符合人类的视觉感知习惯。客观指标通过融合图像与源图像之间数值大小来衡量图像融合质量，能够提供具体评估数值使评价更加客观且准确。在客观指标评价上，本文主要选择了六种客观指标用于融合结果评价[29]。这六种指标分别为：互信息(Mutual Information, MI)、相关系数(Correlation Coefficient, CC))、视觉信息置信度(The Visual Information Fidelity For Fusion, VIFF)、峰值信噪比(Peak Signal-to-Noise Ratio, PSNR)、多尺度结构相似度(Multi-Scale Structural Similarity Index, MS-SSIM)以及信息熵(Entropy, EN)。其中，MI用于测量两幅图像之间的相似程度，即融合图像获取了原图像信息量的多少。互信息越大，表示融合图像保留更多源图像信息，质量越好。CC是

一种用于评估图像融合质量的指标，它用于度量两幅图像之间的相似程度。越相似相关系数约接近于1。VIFF是一种图像质量评估方法，用于比较两幅图像的相似性，数值越高说明所融合图像信息量越多。PSNR是一种用于衡量两幅图像之间差异指标。其数值越高，表示比较图像越相似，融合结果越优秀。MS-SSIM是一种用于评估图像质量的指标，考虑了图像在不同尺度上的结构相似性，数值越大，代表融合结果越好。EN是度量图像包含信息量多少的一个客观评价指标。信息熵越高，表示融合图像的信息量越丰富。

2.2 消融实验

为了验证所提出融合网络的有效性，本文首先在实验部分对融合网络进行消融实验，消融实验主要分为损失函数、网络结构和损失函数超参数三部分。

对于损失函数消融实验，本文分别将三项损失函数中的其中一项设置为0，其余项不变，以评估该损失函数项在网络中的有效性。第一组实验不考虑强度损失项L_{in}、第二组不考虑梯度损失项L_g、第三组不考虑结构相似度损失项L_{SSIM}。实验的客观评价结果如表1所示，主观结果如图4、图5、图6和图7所示。实验结果表明，不考虑强度损失L_{in}会导致融合网络无法优化边缘和纹理信息使得成像细节不够清晰。在不考虑梯度损失L_g之后图像高频信息与低频信息区别不够明显，难以看清细节纹理信息。而不考虑结构相似度损失L_{SSIM}之后，融合结果在整体结构和对比度上均与人类视觉系统的感知不一致，无法保持原图的结构连贯性。结构相似度损失旨在衡量融合图像与源图像在局部结构特征上的一致性，通过对比像素块间的均值、方差及协方差来增强融合结果的感知质量与视觉真实感。综合消融实验结果可知，这三种损失函数在边缘保留、纹理细节和整体结构重建方面各自发挥着不可或缺的作用。

表1 损失函数消融实验结果

Table 1 Ablation study results for the loss functions

Datasets	Loss functions	MI	VIFF	CC	PSNR	EN	MS-SSIM
MSRS	w/o L_{in}	0.327	1.963	0.613	13.600	0.629	0.686
	w/o L_g	0.309	1.803	0.686	13.184	0.609	0.677
	w/o L_{SSIM}	0.389	1.917	0.677	13.260	0.698	0.622
	Ours	0.540	2.291	0.691	13.984	0.702	0.679
Road-Scene	w/o L_{in}	0.385	1.048	0.792	15.958	0.755	0.748
	w/o L_g	0.605	0.958	0.834	17.976	0.934	0.743
	w/o L_{SSIM}	0.501	1.371	0.835	18.259	0.865	0.743
	Ours	0.483	1.539	0.837	17.214	0.963	0.676
TNO	w/o L_{in}	0.186	1.664	0.613	14.925	0.564	0.669
	w/o L_g	0.228	1.114	0.597	14.940	0.791	0.680
	w/o L_{SSIM}	0.227	1.285	0.543	14.549	0.681	0.585
	Ours	0.241	1.800	0.674	15.215	0.841	0.675
M3FD	w/o L_{in}	0.291	1.316	0.799	14.501	0.990	0.773
	w/o L_g	0.364	0.893	0.812	19.417	0.614	0.801
	w/o L_{SSIM}	0.427	1.316	0.812	18.757	0.740	0.804
	Ours	0.298	1.746	0.815	18.818	0.544	0.814



图4 MSRS数据集损失函数消融结果

Fig. 4 Ablation study results of the loss function on the MSRS dataset



图5 Road-Scene数据集损失函数消融结果

Fig. 5 Ablation study results of the loss function on the Road-Scene dataset

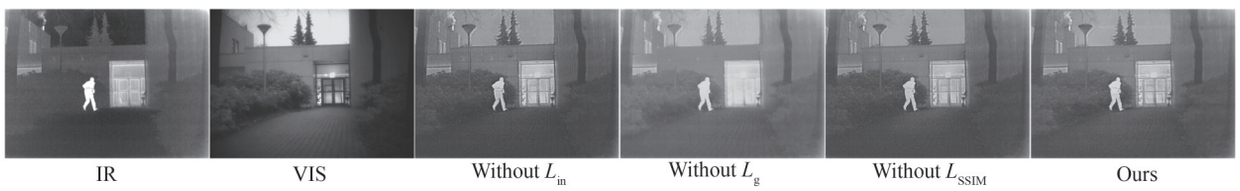


图6 TNO数据集损失函数消融结果

Fig. 6 Ablation study results of the loss function on the TNO dataset

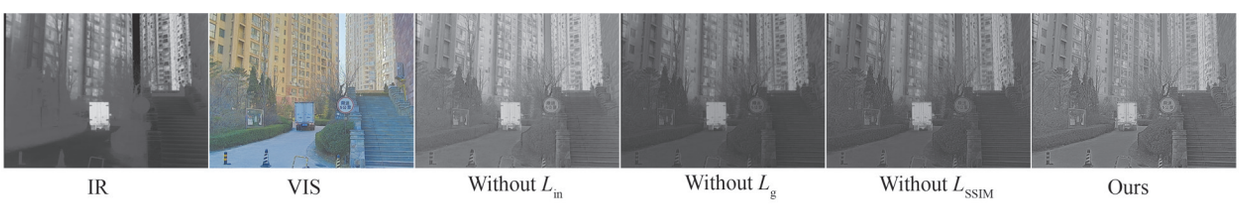


图7 M3FD数据集损失函数消融结果

Fig. 7 Ablation study results of the loss function on the M3FD dataset

网络结构消融实验主要针对本文引入或者设计的不同模块进行。第一组针对SCAM模块、第二组针对CMFE模块、第三组用NMI(I_1, I_2)>0.8,即红外与可见光权重各取0.5替代本文所提出的融合策略、第四组用NMI(I_1, I_2)<0.8,即两种模态权重按照信息熵大小取值替代本文所提出的融合策略、第五组用Concatenate方式替代本文所提出融合策略。网络结构消融实验的客观评价结果如表2和表3所示,主观结果如图8、图9、图10和图11所示。可以看出,各模块在保持图像细节、纹理表达和信息交互方面发挥着不可替代的作用。首先,在不考虑SCAM模块的情况下,融合网络失去了对多尺度特征通道重要性的自适应权重分配能力,导致重要纹理和边缘信息无法被有效放大与保留,融合图像的局部细节明显模糊。其次,在不考虑特征提取模块CMFE后,网络丧失了对深层语义和大范围空间依赖的建模能力,无法从源图像中提取到高级语义特征,使得融合图像在复杂场景下的纹理细节缺失严重,从表2和表3中可以看到各指标均下降

表2 MSRS、Road-Scene数据集网络结构消融实验结果

Table 2 Ablation study results on the MSRS and Road-Scene network structures

Datasets	Network	MI	VIFF	CC	PSNR	EN	MS-SSIM
MSRS	w/o SCAM	0.381	1.817	0.664	13.263	0.667	0.680
	w/o CMFE	0.221	1.405	0.566	12.720	0.559	0.594
	Weight =0.5	0.344	2.223	0.690	13.639	0.656	0.680
	Weight=max(EN_1,EN_2)	0.289	1.874	0.709	13.769	0.604	0.686
	Concatenate	0.348	1.994	0.685	13.416	0.661	0.685
	Ours	0.540	2.291	0.691	13.984	0.702	0.679
Road-Scene	w/o SCAM	0.509	1.450	0.820	15.601	0.893	0.747
	w/o CMFE	0.217	1.135	0.733	14.297	0.759	0.651
	Weight =0.5	0.540	1.349	0.838	16.018	0.916	0.747
	Weight=max(EN_1,EN_2)	0.394	1.468	0.834	16.891	0.774	0.739
	Concatenate	0.434	1.367	0.828	17.319	0.825	0.732
	Ours	0.483	1.539	0.837	17.214	0.963	0.676

表3　TNO、M3FD数据集网络结构消融实验结果

Table 3　Ablation study results on the TNO and M3FD network structures

Datasets	Network	MI	VIFF	CC	PSNR	EN	MS-SSIM
TNO	w/o SCAM	0.233	1.437	0.595	15.093	0.703	0.671
	w/o CMFE	0.154	0.858	0.593	14.209	0.627	0.509
	Weight =0.5	0.230	1.591	0.598	15.177	0.727	0.662
	Weight=max(EN_1,EN_2)	0.136	1.654	0.610	14.823	0.559	0.641
	Concatenate	0.192	1.709	0.642	14.275	0.753	0.673
	Ours	0.241	1.800	0.674	15.215	0.841	0.675
M3FD	w/o SCAM	0.435	1.344	0.811	18.452	0.818	0.795
	w/o CMFE	0.445	0.943	0.810	18.078	0.821	0.801
	Weight =0.5	0.439	1.631	0.808	18.192	0.826	0.787
	Weight =max(EN1,EN2)	0.459	1.720	0.807	18.398	0.779	0.796
	Concatenate	0.425	1.805	0.804	18.285	0.790	0.787
	Ours	0.298	1.746	0.815	18.818	0.544	0.814



图8　MSRS数据集网络结构消融实验结果

Fig. 8　Ablation results of the MSRS dataset network structure

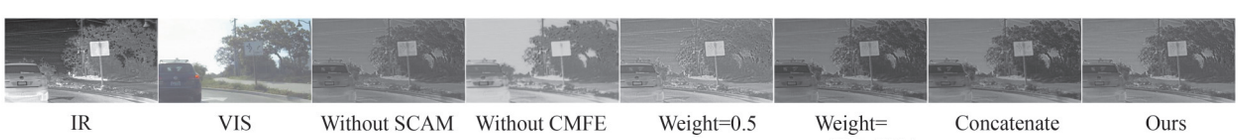


图9　Road-Scene数据集网络结构消融实验结果

Fig. 9　Ablation results of the Road-Scene dataset network structure



图10　TNO数据集网络结构消融实验结果

Fig. 10　Ablation results of the TNO dataset network structure

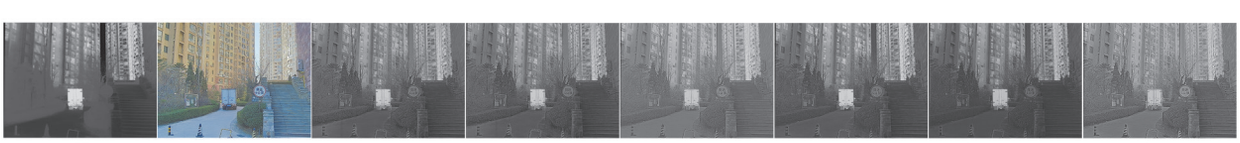


图11　M3FD数据集网络结构消融实验结果

Fig. 11　Ablation results of the M3FD dataset network structure

严重。最后，如果不考虑本文设计的自适应权重选择融合层，两条分支的特征信息仅通过简单加进行融合会导致深度交互缺乏，因而在光照和对比度突变等复杂环境下无法实现充分的特征互补与融合致使整体成像效果欠佳。按照信息熵大小代替本文融合策略时，尽管在部分场景下能够一定程度反映模态贡献度，但这种权重分配方式无法针对不同空间区域的异质性特征进行动态调整，导致某些区域的细节信息无法得到充分保留。例如在复杂光照条件下，红外模态在暗部区域本应发挥主导作用，但熵值全局分配会使得亮部区域的权重配置过高，最终导致融合结果出现对比度不足与细节缺失的情况。采用简单的 Concatenate 方式虽然能保留模态的原始信息，但由于缺乏特征交互与筛选机制，导致融合特征存在冗余和噪声，无法有效突出关键区域的互补优势。由此可见，SCAM 模块主要通过通道注意力突出重要特征，CMFE 特征提取模块可以通过上下文建模捕捉全局语义，而信息选择融合层则通过计算两种模态归一化互信息分别分配权重使得双分支信息的充分交互。并且三者协同作用才能同时兼顾细节清晰度、纹理丰富度和全局一致性。

为验证本文所选取的损失函数超参数，本文分别对三个损失函数项设定不同权值。从表 4 的客观评价指标可以看出，本文方法在多个数据集上均表现出较强的综合优势。在 MSRS 数据集上，本文方法在 MI 与 VIFF 指标上均取得了最优结果，说明所提融合策略能够更充分地保留源图像信息并提升视觉保真度。在 Road-Scene 数据集中，本文方法在 VIFF 与 PSNR 指标上均处于最优水平，表明其在复杂道路场景下能够有效保持结构清晰度和细节完整性。在 TNO 数据集中，本文方法在所有主要指标上均显著优于对照组，尤其在 CC、PSNR 与 EN 上提升明显，证明其在夜间及红外主导场景下能够有效增强对比度并保留更多信息。在 M3FD 数据集中，本文方法在 PSNR 与 MS-SSIM 两项指标上取得最优结果，表明融合图像在质量和结构一致性方面表现突出，整体性能依然保持了较好的平衡。整体来看，本文在损失函数设计中对强度项赋予了更高权重，从而有效保证了融合图像的清晰度和细节最佳。同时，梯度项与结构项的协同作用进一步提升了边缘与纹理的表达能力，使得融合结果在不同场景下均具备较强的鲁棒性与普适性。

表 4　损失函数超参数消融实验客观结果

Table 4　Objective ablation results on loss function hyperparameters

Datasets	Loss functions	MI	VIFF	CC	PSNR	EN	MS-SSIM
MSRS	$\beta_1=10, \beta_2=1, \beta_3=1$	0.453	2.132	0.609	11.668	0.849	0.607
	$\beta_1=1, \beta_2=1, \beta_3=10$	0.115	1.317	0.724	15.122	0.276	0.673
	$\beta_1=1, \beta_2=1, \beta_3=1$	0.390	1.608	0.682	12.348	0.730	0.671
	Ours	0.540	2.291	0. 691	13.984	0.702	0.679
Road-Scene	$\beta_1=10, \beta_2=1, \beta_3=1$	0.583	1.218	0.800	16.587	0.954	0.682
	$\beta_1=1, \beta_2=1, \beta_3=10$	0.237	1.054	0.820	15.114	0.472	0.746
	$\beta_1=1, \beta_2=1, \beta_3=1$	0.619	1.248	0.835	18.599	0.941	0.739
	Ours	0.483	1.539	0.837	17.214	0.963	0.676
TNO	$\beta_1=10, \beta_2=1, \beta_3=1$	0.182	1.577	0.537	13.219	0.795	0.601
	$\beta_1=1, \beta_2=1, \beta_3=10$	0.081	1.052	0.616	14.401	0.207	0.698
	$\beta_1=1, \beta_2=1, \beta_3=1$	0.218	1.430	0.605	14.910	0.829	0.675
	Ours	0.241	1.800	0.674	15.215	0.841	0.675
M3FD	$\beta_1=10, \beta_2=1, \beta_3=1$	0.341	1.306	0.774	14.163	0.960	0.693
	$\beta_1=1, \beta_2=1, \beta_3=10$	0.270	1.246	0.815	19.176	0.474	0.804
	$\beta_1=1, \beta_2=1, \beta_3=1$	0.332	1.467	0.802	15.677	0.980	0.773
	Ours	0.298	1.746	0.815	18.818	0.544	0.814

2.3　对比实验

为了更加客观全面评价本文所提出融合网络的性能，本文引入七种近年来先进典型的红外与可见光图像融合方法进行对比，其中包括：DIDFuse[30]、DDcGAN[31]、DCSFuse[32]、FECFusion[33]、PFCFuse[34]、DATFuse[15]、ITFuse[35]。这八种融合算法的主观结果对比如图 12、图 13、图 14 和图 15 所示。表 5 和表 6 分别表示对应数据集的客观评价结果。由图 12 局部放大结果可见，DATFuse 与 DCSFuse 等算法生成的融合影像存在显著过曝现象，致使建筑物边缘与纹理细节严重丢失，其根源在于未能自适应地权衡红外与可见

光模态的权重。图13表明DIDFuse与DDcGAN在红色通道的细粒度结构保持方面表现欠佳，主要归因于纹理特征提取能力不足。图14显示FECFusion与IFuse方法在复杂干扰场景下出现信息残缺，说明了它们对复杂场景输入的缺乏鲁棒性。图15说明PFCFuse和DDCGAN部分区域存在曝光严重，难以看清细节信息。综合分析四个数据集的定性与定量结果，可以发现本文方法在保持源图像细节的同时，呈现的视觉特性更贴近真实场景。

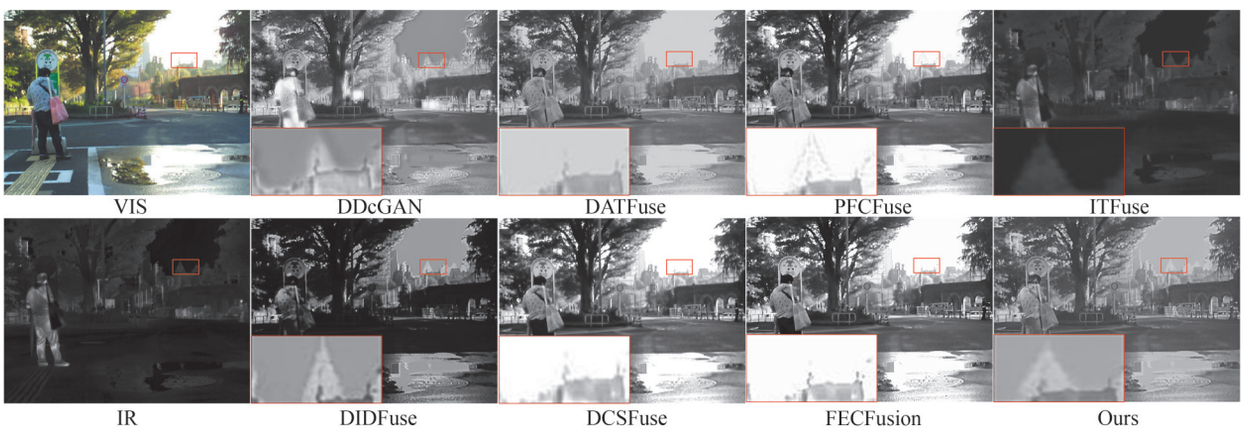


图12　MSRS数据集对比实验

Fig. 12　Comparison experiment of MSRS dataset



图13　Road-Scene数据集对比实验

Fig. 13　Comparison experiment of Road-Scene dataset

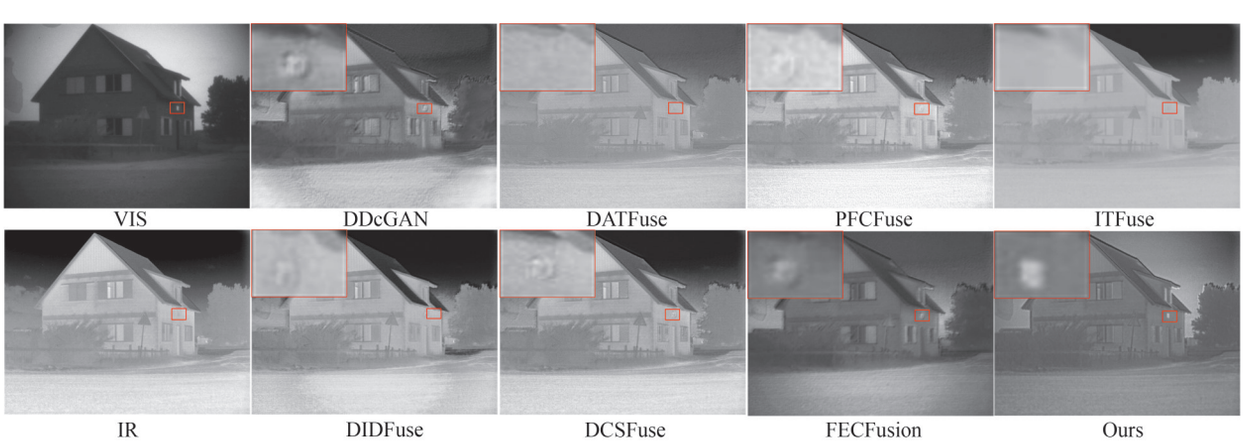


图14　TNO数据集对比实验

Fig. 14　Comparison experiment of TNO dataset

表5和表6的统计结果也证实，本文方法在每个数据集上至少有三项指标位列第一，且CC值在全部测试中均达最高。具体来说在MSRS数据集上，本文方法在MI、CC与EN等指标上表现优异，说明其能够兼



图 15 M3FD 数据集对比实验

Fig. 15 Comparison experiment of M3FD dataset

表 5 MSRS 和 Road-Scene 数据集客观指标对比

Table 5 Comparison of objective indicators of the MSRS and Road-Scene datasets

Datasets	Comparison methods	MI	VIFF	CC	PSNR	EN	MS-SSIM
MSRS	DIDFuse	0.092	1.576	0.567	16.069	0.185	0.648
	DDcGAN	0.120	2.958	0.540	9.160	0.969	0.521
	DCSFuse	0.005	3.760	0.056	13.120	0.474	0.386
	DATFuse	0.341	2.111	0.599	15.107	0.521	0.714
	PFCFuse	0.410	2.888	0.603	16.240	0.465	0.720
	FECFusion	0.383	3.530	0.514	14.912	0.512	0.663
	ITFuse	0.044	0.758	0.517	16.095	0.057	0.684
	Ours	0.540	2.291	0.691	13.984	0.702	0.679
Road-Scene	DIDFuse	0.445	1.685	0.639	14.232	0.765	0.638
	DDcGAN	0.370	1.275	0.611	13.798	0.961	0.586
	DCSFuse	0.453	1.819	0.588	14.391	0.788	0.619
	DATFuse	0.386	0.781	0.606	14.364	0.521	0.632
	PFCFuse	0.414	2.659	0.732	16.342	0.828	0.657
	FECFusion	0.642	1.546	0.484	13.355	0.783	0.592
	ITFuse	0.731	0.805	0.471	14.384	0.948	0.576
	Ours	0.483	1.539	0.837	17.214	0.963	0.676

表 6 TNO 和 M3FD 数据集客观指标对比

Table 6 Comparison of objective indicators of the TNO and M3FD datasets

Datasets	Comparison methods	MI	VIFF	CC	PSNR	EN	MS-SSIM
TNO	DIDFuse	0.236	0.851	0.512	15.755	0.705	0.647
	DDcGAN	0.307	2.852	0.666	14.446	0.752	0.658
	DCSFuse	0.456	2.318	0.562	18.467	0.693	0.628
	DATFuse	0.555	1.293	0.545	16.411	0.755	0.519
	PFCFuse	0.360	1.944	0.457	13.327	0.717	0.613
	FECFusion	0.477	1.927	0.669	16.724	0.562	0.656
	ITFuse	0.532	0.866	0.319	13.286	0.687	0.573
	Ours	0.241	1.800	0.674	15.215	0.841	0.675
M3FD	DIDFuse	0.317	1.471	0.798	14.375	0.999	0.643
	DDcGAN	0.323	1.729	0.759	14.787	0.967	0.707
	DCSFuse	0.336	1.341	0.744	18.429	0.494	0.774
	DATFuse	0.515	0.765	0.776	16.663	0.790	0.782
	PFCFuse	0.484	1.742	0.800	18.167	0.782	0.797
	FECFusion	0.363	1.436	0.788	19.071	0.610	0.793
	ITFuse	0.359	0.388	0.806	18.621	0.682	0.805
	Ours	0.298	1.746	0.815	18.818	0.544	0.814

顾信息量与相关性。在Road-Scene数据集中本文方法在CC与PSNR上显著领先，体现了对细节保真度和结构一致性的增强。从TNO数据集来看，指标CC与EN等指标上保持优势，说明其在复杂背景下依然能够稳健工作。此外，M3FD数据集结果表明，本文方法不仅在CC与MS-SSIM上表现最佳，同时在PSNR与EN等指标上也具有较强竞争力，进一步验证了其在多场景下的泛化性和鲁棒性。整体来看本文提出的方法能够实现充分的跨模态信息交互，在最大程度上保留原始图像信息的同时，有效提升了融合结果的相关性和一致性。表7给出了本文所提出的方法与另外7种方法的模型参数量（Parameters）、浮点运算次数（Flops）和平均运行时间（Average time）的对比，可以本文方法在保持优越客观指标的同时，展现出更高的计算效率和更低的模型复杂度。能够在保证融合质量的前提下有效降低计算开销。这主要归功于本文引入的轻量化特征提取与自适应融合设计。

表7　融合方法模型参数量与浮点运算次数对比

Table 7　Comparison of flops and parameter counts across fusion methods

Comparison methods	Flops	Parameters	Average time
DIDFuse	17.142 G	260.935 K	1.33 s
DDcGAN	3.628 G	169.885 K	1.59 s
DCSFuse	27.608 G	316.787 K	0.76 s
DATFuse	8.519 G	13.112 K	0.77 s
PFCFuse	51.986 G	525.664 K	4.78 s
FECFusion	18.772 G	144.453 K	1.26 s
ITFuse	5.681 G	18.333 K	0.34 s
Ours	14.601 G	91.017 K	1.05 s

2.4　目标检测实验

图像融合在实际生活中具有广泛的应用场景，如工业检测与质量控制[36]、无人机边缘计算[37]、机器人视觉与自动驾驶[38]和遥感与卫星成像[39]等。为验证所提方法的后续任务性能，本文基于YOLOv5[40]对融合图像中的行人和车辆目标进行了检测，并对不同融合算法的检测结果进行了对比分析。本文在Road-Scene数据集中划分200张训练集、10张验证集和11张测试集。为了客观评价图像融合算法对目标检测性能比较，本文选取准确率（Precision）、召回率（Recall）、mAP50（Average Precision）和mAP@50:5:95分别用于车辆与行人客观评价，在评价指标中，Precision越大，表示将其他类别预测为本类别的个数也就越少，即预测出的正确率越高。mAP代表对所有类别平均精度的平均值，数值越高说明预测精度越高。Recall用于衡量正确预测的正例数量占实际所有正例数量的比例。其数值越高说明分类越精确。mAP@50:5:95表示多个IOU阈值下的mAP，会在区间[0.5,0.95]内，以0.05为步长，取10个IOU阈值，分别计算这10个IOU阈值下的mAP，再取平均值。这四个指标越高，代表检测质量越高。最终检测结果指标数值如表8所示，主观检测结果对比如图16所示。

表8　目标检测客观评价

Table 8　Objective evaluation of object detection

	The detection of person				The detection of car			
	Precision	Recall	mAP50	mAP@50:5:95	Precision	Recall	mAP50	mAP@50:5:95
IR	0.75	0.872	0.782	0.471	0.889	0.826	0.866	0.393
VI	0.932	0.467	0.541	0.313	1.000	0.435	0.487	0.141
DIDFuse	0.751	0.333	0.432	0.205	0.461	0.24	0.287	0.089
DDcGAN	0.866	0.533	0.652	0.311	0.846	0.44	0.586	0.252
DCSFuse	0.754	0.733	0.861	0.462	0.864	0.759	0.867	0.391
DATFuse	0.917	0.733	0.741	0.318	0.946	0.760	0.862	0.427
PFCFuse	0.751	0.800	0.811	0.418	0.712	0.912	0.819	0.396
FECFusion	0.796	0.467	0.500	0.264	0.749	0.360	0.459	0.126
ITFuse	0.619	0.867	0.788	0.481	0.842	0.640	0.749	0.354
Ours	0.997	0.756	0.878	0.476	0.917	0.769	0.874	0.434



图 16 目标检测结果可视化

Fig. 16 Visualization of object detection results

从表 8 可以看出，本文方法在车辆检测与行人检测中分别取得两个最优的结果，这表明该方法预测框能够精准识别融合后的目标物体，并且具有较高的正确率。图 16 可以观察到，图中行人目标难以被有效检测，造成该结果的主要原因是在大多数被检测的图片中，行人所处位置的背景曝光较为严重，导致算法难以准确识别是否存在行人。而本文算法采用红外信息与可见光信息交互选择，能够有效地提取行人的轮廓并在融合图像中体现出来。同时还能够分别提取并保存红外与可见光图像的互补信息，有效降低在黑暗环境下曝光严重对检测性能的影响。

3 结论

本文提出了一种基于 CNN-Mamba 特征提取结合信息选择的红外与可见光图像融合方法。首先在编码阶段，利用空间上下文感知模块对图像信息进行全局上下文特征整合与线性变换，旨在增强高频与低频信息之间的关联并自适应分配权重。随后，设计同时具有局部和大范围感受野的 CMFE 模块进行深度特征提取。在融合阶段，本文提出了自适应选择融合策略的融合层，最大限度保留所需信息。最后，融合后的信息通过多个卷积网络逐步解码，重构出融合图像。为了验证所提出方法的合理性，本文分别进行了损失函数、网络结构和损失函数超参数的消融实验。在对比试验部分，将本文所提出的算法与七种近年来典型且先进的融合算法进行比较，实验结果表明，本文所提出的方法在多项客观指标上取得最优性能，综合表现也具有先进性。此外，在目标检测后续任务对比实验中，本文方法结果也取得了多项最优指标，充分验证了其实用价值。综合来看，本文所提出的方法在融合质量和后续应用均展现出显著优势。然而，本文方法与其它基于深度学习网络融合方法一样在损失函数超参数调试过程中消耗了大量计算资源。因此，如何高效自适应进行损失函数超参数调整将是未来研究的重点。

参考文献

[1] FENG Xin, YANG Jieming, ZHANG Hongde, et al. Infrared and visible image fusion based on dual channel residual dense network[J]. Acta Photonica Sinica, 2023, 52(11): 1110003.
冯鑫, 杨杰铭, 张鸿德, 等. 基于双通道残差密集网络的红外与可见光图像融合[J]. 光子学报, 2023, 52(11): 1110003.

[2] LUO Tao, CHEN Ning, ZHU Xianyou, et al. An information-driven modular visible-infrared image fusion algorithm[J]. Journal of Zhejiang University of Science and Technology, 2025, 37(3): 270-279.
罗涛, 陈宁, 朱先有, 等. 信息驱动的模块化可见光-红外图像融合算法[J]. 浙江科技大学学报, 2025, 37(3): 270-279.

[3] ZHOU Xinyan, YAO Jian. A deep feature fusion image retrieval algorithm based on swin transformer[J]. Surveying and Geographic Information, 2025, 50(3): 44-49.
周心妍, 姚剑. 基于 Swin Transformer 的深度特征融合图像检索算法[J]. 测绘地理信息, 2025, 50(3): 44-49.

[4] SHNE Yu, QIANG Zhenkai, WEI Ziyi, et al. A comprehensive review on deep learning-enabled infrared and visible light image fusion: development and prospects[J]. Control & Decision, 2025, 40(6): 1793-1806.
沈瑜, 强振凯, 魏子易, 等. 基于深度学习的红外与可见光图像融合综述: 发展与展望[J]. 控制与决策, 2025, 40(6): 1793-1806.

[5] LI Yunhang, PAN Qing, TIAN Nili. Structure similarity optimized hybrid multi-scale medical image fusion[J]. Computer Engineering, 2024, 50(7):264-270.
李云航, 潘晴, 田妮莉. 结构相似度优化的混合多尺度医学图像融合[J]. 计算机工程, 2024, 50(7):264-270.
[6] ZHANG Yajia, HEI Hongting, HAI Mei, et al. Brain image fusion based on improved residual neural network and sparse representation[J]. Journal of Shaanxi University of Technology (Natural Science Edition), 2025, 41(2):91-100.
张亚加, 黑荣婷, 海梅, 等. 基于改进残差神经网络和稀疏表示的脑部图像融合[J]. 陕西理工大学学报(自然科学版), 2025, 41(2):91-100.
[7] ZHANG Ying, NIE Rencan, MA Chaozhen, et al. MRI and PET/SPECT image fusion based on subspace feature interlearning[J]. Computer Science, 2022, 49(2): 382-387.
张瑛, 聂仁灿, 马朝振, 等. 基于子空间特征相互学习的MRI与PET/SPECT图像融合[J]. 计算机科学, 2022, 49(2): 382-387.
[8] LI Dongyang, NIE Rencan, PAN Lina, et al. Unsupervised significance mask guidance for infrared and visible light image fusion network[J]. Computers & Science, 2024, 51(1): 368-372.
李东阳, 聂仁灿, 潘琳娜, 等. 基于无监督显著性掩码引导的红外与可见光图像融合网络[J]. 计算机科学, 2024, 51(1): 368-372.
[9] SHEN Yu, CHEN Xiaopeng, LIU Cheng, et al. Fusing infrared and visible images based on hybrid model driver[J]. Control and Decision, 2021, 36(9): 2143-2151.
沈瑜, 陈小朋, 刘成, 等. 基于混合模型驱动的红外与可见光图像融合[J]. 控制与决策, 2021, 36(9):2143-2151.
[10] ZHANG Hongde, FENG Xin, YANG Jieming, et al. A dual branch edge convolution fusion network for infrared and visible images[J]. Acta Photonica Sinica, 2024, 53(8): 0810004.
张鸿德, 冯鑫, 杨杰铭, 等. 基于双分支边缘卷积融合网络的红外与可见光图像融合方法[J]. 光子学报, 2024, 53(8): 0810004.
[11] LI H, WU X J. DenseFuse: a fusion approach to infrared and visible images[J]. IEEE Transactions on Image Processing, 2019, 28(5):2614-2623.
[12] ZHANG Y, LIU Y, SUN P, et al. IFCNN: a general image fusion framework based on convolutional neural network[J]. Information Fusion, 2020, 54:99-118.
[13] MA J, YU W, LIANG P, et al. FusionGAN: a generative adversarial network for infrared and visible image fusion[J]. Information Fusion, 2019, 48:11-26.
[14] WANG Z, CHEN Y, SHAO W, et al. SwinFuse: a residual swin transformer fusion network for infrared and visible images[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71:1-12.
[15] 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.
[16] ZHANG Y, YE M, ZHU G, et al. FFCA-YOLO for small object detection in remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62:1-15.
[17] LIU Y, MA J, CHEN X, et al. Vmamba: visual state space model[J/OL]. arXiv, 2024, arXiv:2401.10166.
[18] 段朝伟. 红外与可见光图像像素级融合方法研究[D]. 南京:南京航空航天大学, 2023.
[19] GE Jianjun, TANG Siqi, LI Mingqiang, et al. Complexity perception system information theory and construction methods [J]. Journal of Radar Science (in Chinese & English), 2025, 14(3): 651-663.
葛建军, 唐思琦, 李明强, 等. 复杂感知系统信息理论与构建方法[J]. 雷达学报(中英文), 2025, 14(3):651-663.
[20] ROMANO S, BAILEY J, VINH N, et al. Standardized mutual information for clustering comparisons: one step further in adjustment for chance[C]. In Proceedings of the 31st International Conference on International Conference on Machine Learning, 2014.
[21] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 770-778.
[22] HUANG G, LIU Z, Van DER MAATEN L, et al. Densely connected convolutional networks[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017: 2261-2269.
[23] LONG Y, JIA H, ZHONG Y, et al. RXDNFuse: an aggregated residual dense network for infrared and visible image fusion[J]. Information Fusion, 2021, 69: 128-141.
[24] TANG L, YUAN J, MA J, et al. Image fusion in the loop of high-level vision tasks: a semantic-aware real-time infrared and visible image fusion network[J]. Information Fusion, 2022, 82:28-42.
[25] TOET A. The TNO multiband image data collection[J]. Data in Brief, 2017, 15: 249-251.
[26] 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.
[27] 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 (CVPR), 2022: 5802-5811.
[28] SUN Jing, WANG Zhishe, YANG Fan, et al. Multi-layer perceptron interactive fusion method for infrared and visible images[J]. Infrared Technology, 2025, 47(5): 619-627.
孙婧, 王志社, 杨帆, 等.红外与可见光图像多层感知机交互融合方法[J].红外技术,2025,47(5):619-627.
[29] WU X, FENG X, LU X, et al. W-EICMFusion: a fusion network for infrared and visible images utilising WOA hyperparameter optimisation[J]. Pattern Recognition, 2026, 172: 112531.
[30] ZHAO Z, LIANG P, LI M, et al. DIDFuse: deep image decomposition for infrared and visible image fusion[C]. Proceedings of the International Joint Conference on Artificial Intelligence, 2020:970-976.
[31] MA J, XU H, JIANG J, et al. DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion[J]. IEEE Transactions on Image Processing, 2020, 29:4980-4995.
[32] CAI Z, MA Y, HUANG J, et al. Correlation-guided discriminative cross-modality features network for infrared and visible image fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73:1-18.
[33] CHEN Z, FAN H, MA M, et al. FECFusion: infrared and visible image fusion network based on fast edge convolution[J]. Mathematical Biosciences and Engineering, 2023, 20(9):16060-16082.
[34] HU X, LIU Y, YANG F. PFCFuse: a poolformer and CNN fusion network for infrared-visible image fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73:1-14.
[35] TANG W, HE F, LIU Y. ITFuse: an interactive transformer for infrared and visible image fusion[J]. Pattern Recognition, 2024, 156:110822.
[36] LI J, GAO B, WOO W L, et al. A novel multispectral fusion defect detection framework with coarse-to-fine multispectral registration[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73:1-13.
[37] YAO Z, ZHU Q, ZHANG Y, et al. Minimizing long-term energy consumption in RIS-assisted AAV-enabled MEC network[J]. IEEE Internet of Things Journal, 2025, 12(12):20942-20958.
[38] QU L, LIU S, WANG M, et al. Trans2Fuse: empowering image fusion through self-supervised learning and multi-modal transformations via transformer networks[J]. Expert Systems with Applications, 2024, 236:121363.
[39] CHEN Yi, ZHANG Shuai, WANG Guipin. A vehicle detection algorithm based on laser radar and camera information fusion[J]. Machinery & Electronics, 2020, 38(1): 52-56.
陈毅, 张帅, 汪贵平.基于激光雷达和摄像头信息融合的车辆检测算法[J].机械与电子,2020,38(1):52-56.
[40] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: unified, real-time object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016:779-788.

A Method for Infrared and Visible Image Fusion Based on CNN-Mamba Feature Extraction Combined with Information Selection

WU Xiaocong[1], FENG Xin[1], HUANG Meina[1], ZHAO Bin[1], HU Kaiqun[1], SHEN Yu[2]
(1 *School of Mechanical Engineering, Chongqing Key Laboratory of Green Design and Manufacturing of Intelligent Equipment, Chongqing Technology and Business University, Chongqing* 400067, *China*)
(2 *School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou* 730070, *China*)

Abstract: Infrared sensors capture the thermal radiation emitted from the surface of objects to create images. However, infrared images are prone to noise interference and are limited in capturing fine details and texture information. In contrast, visible sensors capture reflected light from the surface of objects, providing rich details and texture information that align with human visual perception. However, the quality of visible images can be unstable due to factors such as lighting conditions. Therefore, effectively fusing infrared and visible images to generate high-quality images is an important research direction.

Currently, image fusion methods for infrared and visible images can be broadly categorised into traditional fusion strategies and deep learning-based fusion strategies. Traditional methods struggle to effectively integrate the feature information from the two modalities, especially in complex fusion scenarios. Deep learning-based fusion methods, such as ResNet and DenseNet, can better handle multi-modal information, but they involve complex training processes and high computational overhead. To address these challenges, this paper proposes an infrared and visible image fusion method that combines CNN-Mamba feature extraction with an information selection mechanism. The aim is to overcome the shortcomings of traditional methods in complex scenarios while simplifying the training process of deep

learning-based fusion strategies. In the image encoding phase, we introduce the Spatial Context Aware Module (SCAM) for the first time in the field of image fusion, which integrates contextual feature information and allocates attention weights adaptively. Then, we design the CNN-Mamba Feature Extract (CMFE) module, which integrates CNN and Mamba networks to perform deep feature extraction on both modalities of the image. In the fusion phase, considering the information differences between infrared and visible images in the same scene, we propose a novel adaptive weight selection fusion layer to retain more source information. Specifically, we calculate the normalised mutual information between the two modalities and set a threshold. When the normalised mutual information exceeds the threshold, the fusion weights are set to be equal to each other. When the mutual information is below the threshold, we assign higher weights to the modality with higher information entropy to preserve more source information. Finally, in the decoding phase, we utilise a decoder consisting of four CNN modules to progressively reduce the number of channels and reconstruct the fused image.

During the experimental phase, ablation studies validate the rationality of the proposed network structure and loss function design. When compared with seven advanced fusion methods, our method demonstrates excellent performance on the MSRS, Road-Scene, TNO, and M3FD public datasets, ranking first in at least three objective metrics. Moreover, the EN and CC values achieve the highest in all tests. This indicates that the proposed method effectively facilitates cross-modal information interaction and preserves the original information as much as possible. Additionally, pedestrian and vehicle detection experiments were conducted on the Road-Scene dataset, and the results show that our method achieved the optimal performance in both pedestrian and vehicle detection, further validating its practical application advantages.

In conclusion, the proposed method demonstrates significant advantages in both image fusion quality and subsequent applications.

Key words: Image fusion; Normalized mutual information; CMFE Module; Information Entropy; Adaptive weight selection fusion layer

OCIS Codes: 100.4996; 150.1135; 100.2980; 110.3000

CSTR: 32255.14.gzxb20265501.0110001

Foundation item: National Natural Science Foundation of China (62241106, 61861025), the Key R&D Program of Gansu Province (24YFGA037), National Science Fund for Distinguished Young Scholars of China (42325502)