DOI:10.11918/202410008

跨尺度交叉注意力的红外和可见光图像融合

邬开俊[1],张园园[1],张巧艳[2]

(1.兰州交通大学 电子与信息工程学院,兰州 730070;2.中国铁路信息科技集团有限公司 数据和装备处,北京 100080)

摘 要:为改善红外和可见光图像融合过程中信息丢失及特征融合局限性等问题,提出一种跨尺度交叉注意力的Transformer网络。首先,设计跨尺度视觉注意力模块以加强弱光照条件下特征提取能力,重点关注低对比度区域的信息;其次,构造交叉约束模块,通过嵌套连接进行全局特征的交互与融合;最后,提出一种基于Transformer的融合策略,进一步提升图像融合性能,增强细节信息保留。与其他9种先进融合网络的对比实验表明:在客观评价方面,跨尺度交叉注意力的Transformer网络在TNO数据集上的8项评价指标分别提升了10.5%~27.8%,在Road scene数据集上的8项评价指标分别提升了8.6%~30.3%;在主观评价方面,跨尺度交叉注意力的Transformer网络展现出更优的视觉表现,显著提升了融合图像的可解释性与实用性,能有效保留红外图像中的热目标信息与可见光图像中的细节纹理。所提网络能够充分提取源图像信息,生成纹理细节更清晰、对比度更佳的融合结果,为提升图像融合性能及支持下游任务提供有益参考。

关键词:红外和可见光;图像融合;交叉注意力机制;跨尺度交互式Transformer

中图分类号:TP391.41 **文献标志码**:A **文章编号**:0367-6234(2026)09-0135-09

Infrared and visible image fusion with cross-scale and cross-attention

Wu Kaijun[1], Zhang Yuanyuan[1], Zhang Qiaoyan[2]

(1. School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China;
2. Data and Equipment Department, China Railway Information Technology Group Co. Ltd., Beijing 100080, China)

Abstract: To address issues such as information loss and limitations in feature fusion during infrared and visible image fusion, a cross-scale cross-attention Transformer network is proposed. First, a cross-scale visual attention module is designed to enhance feature extraction capabilities under low-light conditions, focusing on information in low-contrast regions. Second, a cross-domain constrain module is constructed to facilitate interaction and fusion of global features through nested connections. Finally, a Transformer-based fusion strategy is proposed to further enhance image fusion performance and improve detail retention. Comparative experiments with nine other advanced fusion networks showed that in terms of objective evaluation, the cross-scale cross-attention Transformer network improved 8 evaluation indicators by 10.5% to 27.8% on the TNO dataset and by 8.6% to 30.3% on the Road scene dataset; In terms of subjective evaluation, the cross-scale cross-attention Transformer network exhibits better visual performance, significantly improving the interpretability and practicality of fused images, and effectively preserving thermal target information in infrared images and detailed textures in visible images. The proposed network can fully extract source image information, generate fusion results with clearer texture details and better contrast, providing useful references for improving image fusion performance and supporting downstream tasks.

Keywords: infrared and visible; image fusion; cross-attention mechanism; cross-scale interactive Transformer

图像融合的目的是将同一场景中不同模式、不同位置、不同时间的多幅图像进行融合,生成信息丰富的融合图像[1]。由于图像融合技术能够有效整合来自不同成像传感器的数据,利用不同数据之间的互补优势,生成信息完整的场景图像,因此,图像融合被广泛应用于军事、遥感、医疗诊断、安防、监控、摄影等[2-4]领域。

红外和可见光图像具有很强的信息互补性,红外和可见光图像融合因此成为一个重要的研究课题[5]。红外图像含有明显的目标信息,但缺乏纹理特征且分辨率较低。而可见光图像具有高空间分辨率和丰富的纹理细节特征。通过将红外和可见光图

收稿日期:2024-10-08;**录用日期**:2024-11-14;**网络首发日期**:2025-11-05
网络首发地址:https://link.cnki.net/urlid/23.1235.T.20251105.1502.002
基金项目:甘肃省自然科学基金(23JRRA913);内蒙古自治区重点研发与成果转化计划项目(2023YFSH0043,2023YFDZ0054);兰州交通大学重点研发项目(ZDYF2304)
作者简介:邬开俊(1978—),男,教授,博士生导师
通信作者:邬开俊,wkj@mail.lzjtu.cn

像进行融合,可以同时利用目标的辐射信息和纹理信息,生成目标突出、细节丰富的融合图像。

红外和可见光图像融合方法主要包括传统方法[6]和基于深度学习的方法[7]。传统图像融合大多采用基于多尺度变换方法[8]、基于稀疏表示方法[9]、基于显著性方法[10]和基于子空间方法[11]等对不同模态图像进行整合。由于传统图像融合方法依赖人工设计,对源图像特征提取不充分,且在分离不同传感器的图像时使用相同变换方式,得到的特征性能具有局限性。而基于深度学习方法由于其强大的自动学习能力,可以获得信息丰富的多层次深度特征,在图像融合任务中显示出了巨大的潜力。Prabhakar 等[12]提出了用于融合极端曝光差异图像对的一种基于深度学习的无监督网络,该网络利用深度卷积网络学习源图像亮度通道的多级特征,并通过通道间的平均操作实现特征融合,从而生成信息丰富的融合图像。Li 等[13]提出了一种密集嵌套连接的红外与可见光图像融合网络,该网络结合了残差和密集连接,通过编码器提取多尺度特征。Ma 等[14]提出了一种基于 Swin Transformer 的跨域远程学习通用图像融合方法(SwinFusion),该方法有效融合了来自不同领域和具有不同依赖关系的信息。李永萍等[15]提出了一种基于变换域 VGGNet19 的图像融合网络,该网络结合了变换域和深度学习网络的优势,有效防止了融合结果丢失细节边缘信息。刘光辉等[16]提出了一种并行混合注意力的渐进融合图像增强方法,该方法通过聚合不同感受野下学习到的多尺度特征,加强了原始图像细节信息的保留。Tang 等[17]提出了一种具备语义信息的多功能图像配准与融合网络(SuperFusion),该网络通过整合语义信息,为图像的对齐与融合提供了全面的解决方案。Li 等[18]提出了一种基于交叉注意机制的融合新方法(CrossFuse),该方法能够从源图像中提取深层特征,保持细节信息并增强互补信息。尽管基于深度学习的图像融合技术已取得了很多成果,但其在泛化能力、融合结果的视觉质量与信息保留方面仍面临诸多问题与挑战。现有的融合网络未考虑红外和可见光图像之间结构的特有性和相似性,导致特征提取不充分,造成融合图像的视觉效果欠佳。

针对以上问题,本文提出一种跨尺度交叉注意力的 Transformer 红外和可见光图像融合网络。利用 Transformer 的优势,增强融合图像的空间恢复能力。通过跨尺度视觉注意力模块增强弱光照条件下的特征提取,并通过交叉注意力模块实现全局特征的交互与融合,从而提高融合图像的空间恢复能力。

1 融合网络

本文提出跨尺度交叉注意力的红外和可见光图像融合网络,该网络基于 Transformer 架构设计,由编码器网络、基于 Transformer 的融合网络和解码器网络 3 部分构成,具体结构如图 1 所示。

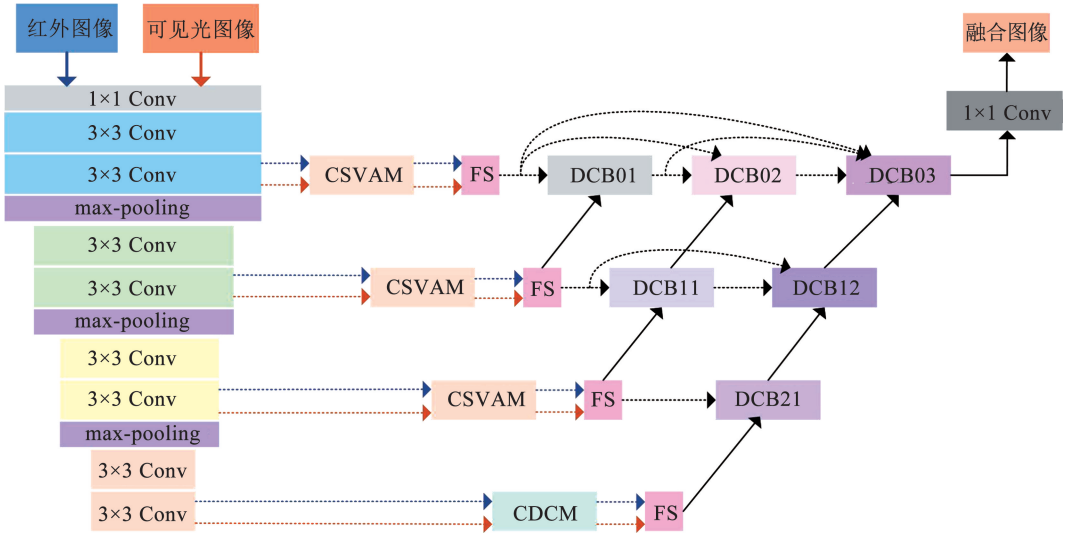


图 1 跨尺度交叉注意力的红外和可见光图像融合网络结构

Fig. 1 Structure of infrared and visible image fusion with cross-scale and cross-attention

编码器网络由 4 个卷积块构成,每个块包括3 × 3 卷积、ReLU 激活函数和最大池化(max-pooling)操作,从红外和可见光图像中提取多尺度的深度特征。通过多层次的特征提取机制,编码器能够捕捉图像的局部和全局信息,确保丰富的特征表达。之后,提取到的多尺度特征进入基于 Transformer 的融合网络进行深度融合。该融合网络的核心是多头自注意力机制,在多个尺度上处理图像特征,捕捉全局上下

文关系并处理长距离依赖特征信息，有效克服传统方法在特征融合中的局限性。最后，融合后的特征被输入至解码器网络中。解码器网络由解码器卷积（decoder conv-block，DCB）组成，并采用融合策略（fusion strategy，FS）融合特征，输出融合图像。该网络通过跨尺度视觉注意力模块（cross-scale visual attention module，CSVAM）和交叉约束模块（cross-domain constrain module，CDCM）进一步处理特征，确保远程依赖特征和空间特征的有效迁移与整合。CSVAM 用于增强弱光照条件下的特征提取，而 CDCM 通过交叉域的特征交互，减少局部信息损失，生成高质量融合图像。

1.1　跨尺度视觉注意力模块

CSVAM 由两阶段跨尺度交互式 Transformer 注意力模块（cross-scale interactive Transformer attention module，A_{CSITM}）组成，如图 2 所示。

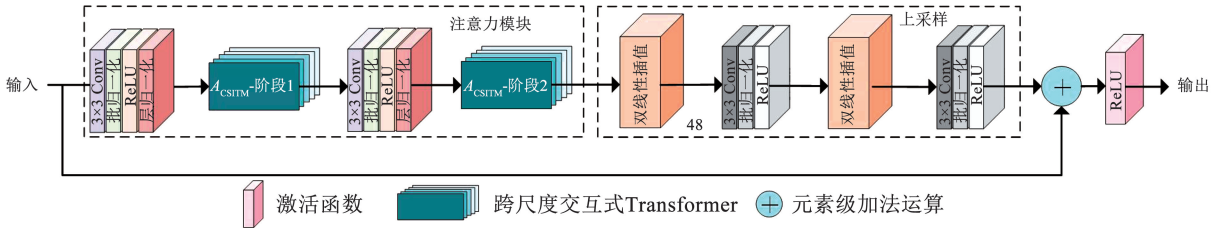


图 2　跨尺度视觉注意力模块结构

Fig. 2　Structure of cross-scale visual attention module

CSVAM 的设计旨在提高计算效率，增强抗噪声能力，特别是在弱光照条件下的特征提取能力。首先，通过步长大于 1 的卷积进行两级下采样，降低特征图空间分辨率，减少计算复杂度。弱光环境下，A_{CSITM}通过多尺度特征提取机制，捕捉不同尺度的特征以补充光照不足导致的细节缺失。接着，经过两阶段的 A_{CSITM}处理：第 1 阶段，利用自注意力机制提取关键特征，动态聚焦显著区域；随后进入第 2 阶段，进一步融合多尺度特征。为避免上采样中的“棋盘效应”，采用双线性插值与卷积组合进行上采样，确保特征图复原时不引入伪影或噪声。最后，利用逐元素加法整合不同路径特征，ReLU 激活进一步提升显著区域的提取能力，使网络在各种光照条件下能准确捕捉关键特征。CSVAM 通过结合跨尺度交互式 Transformer、双线性插值与卷积的上采样方式，增强网络泛化能力和对显著区域的准确预测能力。

在 A_{CSITM}中，输入特征首先通过动态通道分割（dynamic channel partitioning，C_{DP}）模块将其划分为多个不同尺度的特征。这些特征随后被输入至各个多头注意力（multi-head attention，A_{MH}）模块中，每个 A_{MH}分别处理不同尺度的特征（$\boldsymbol{S}_{\mathrm{cale1}}$、$\boldsymbol{S}_{\mathrm{cale2}}$和 $\boldsymbol{S}_{\mathrm{cale3}}$），确保其在不增加计算复杂度的情况下充分捕捉不同尺度之间的相关性。其具体计算式为

$$\boldsymbol{S}_{\mathrm{cale}m}=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_m\boldsymbol{K}_m^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_m \tag{1}$$

式中：$m=1,2,3$；Softmax(·)为激活函数；$\boldsymbol{Q}_m$、$\boldsymbol{K}_m$ 和 $\boldsymbol{V}_m$ 分别为查询矩阵、键矩阵和值矩阵；d_k 为键的维数。

在每个 A_{MH}中，特征通过点积注意力机制进行处理。$\boldsymbol{Q}_m$ 和 $\boldsymbol{K}_m$ 的维度通过可学习的权重进行线性变换，并使用常数对 $\boldsymbol{K}_m$ 的维数进行归一化，以稳定训练过程中的梯度下降。在完成点积注意力计算后，不同尺度的特征在信道维度上进行连接，从而生成多尺度的注意力特征。跨尺度交互式 Transformer 注意力模块的结构如图 3 所示，其计算式为

$$\begin{aligned}A_{\mathrm{CSITM}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=C_{\mathrm{at}}[&A_{\mathrm{MH1}}(\boldsymbol{Q}_1,\boldsymbol{K}_2,\boldsymbol{V}_3),\\&A_{\mathrm{MH2}}(\boldsymbol{Q}_2,\boldsymbol{K}_3,\boldsymbol{V}_1),A_{\mathrm{MH3}}(\boldsymbol{Q}_3,\boldsymbol{K}_1,\boldsymbol{V}_2)]\boldsymbol{W}^{\mathrm{o}}\end{aligned} \tag{2}$$

式中：A_{MH1}、A_{MH2}、A_{MH3}为多头注意力模块，C_{at}为注意力特征拼接操作，$\boldsymbol{W}^{\mathrm{o}}$ 为可学习的权重矩阵。

在跨尺度注意力计算之后，模块还包括层归一化（layer normalization，N_{L}）和多层感知器（multilayer perceptrons，P_{ML}）处理步骤，进一步提升特征的表达能力，并确保网络的稳定性。最终输出融合 3 组不同尺度的注意力特征。

A_{CSITM}整体过程表达式为：

$$\langle\boldsymbol{Q}_{l-1},\boldsymbol{K}_{l-1},\boldsymbol{V}_{l-1}\rangle=C_{\mathrm{DP}}(\boldsymbol{F}_{l-1},k,d,s) \tag{3}$$

$$\boldsymbol{F}'_l=N_{\mathrm{L}}(A_{\mathrm{CSITM}}(\boldsymbol{Q}_{l-1},\boldsymbol{K}_{l-1},\boldsymbol{V}_{l-1}))+\boldsymbol{Q}_{l-1} \tag{4}$$

$$\boldsymbol{F}_l=N_{\mathrm{L}}(P_{\mathrm{ML}}(\boldsymbol{F}'_l))+\boldsymbol{F}'_l \tag{5}$$

式中：l 为提取特征所在的层数，$\boldsymbol{F}_{l-1}$、$\boldsymbol{Q}_{l-1}$、$\boldsymbol{K}_{l-1}$和 $\boldsymbol{V}_{l-1}$分别为 $l-1$ 层的特征、查询矩阵、键矩阵和值矩阵，k 为卷积核大小，d 为空洞卷积膨胀率，s 为卷积步长，$\boldsymbol{F}'_l$ 为运算过程中的中间特征，$\boldsymbol{F}_l$ 为包含多尺度融合的特征信息。

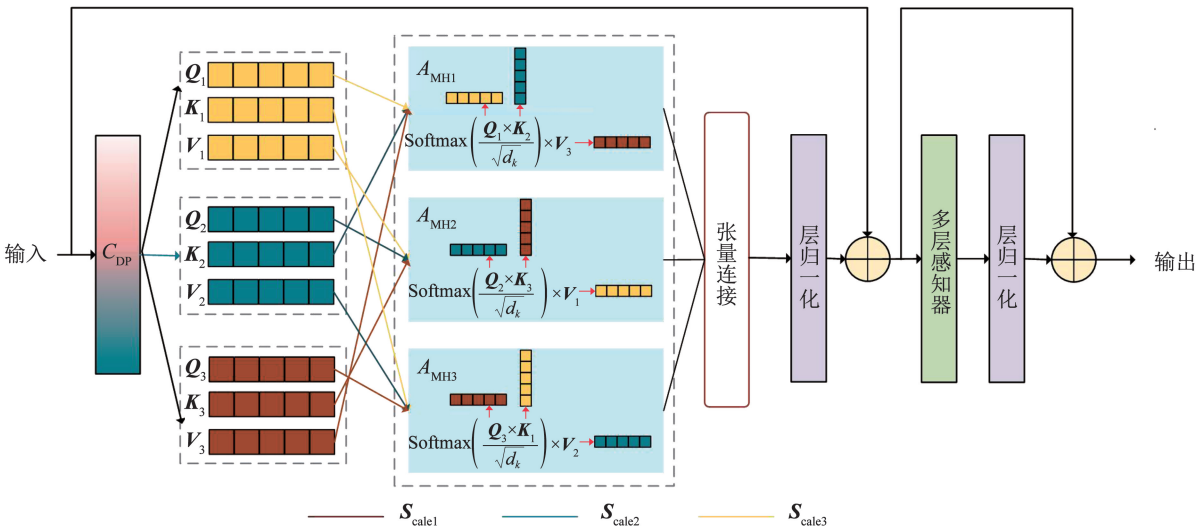


图 3　跨尺度交互式 Transformer 注意力模块结构

Fig. 3　Structure of cross-scale interactive Transformer attention module

1.2　交叉约束模块

为在不同图像之间促进共享特征建模，本文展示一种新型的交叉约束模块。该模块的核心是跨域交互式 Transformer（cross-domain interactive Transformer，CDIT），用于跨域传递互补信息。图 4 为交叉约束模块结构。首先，输入特征经过并行的卷积层和归一化处理；然后，通过 CDIT 进行跨域信息的交互与融合，从而捕捉和传递来自不同域的互补特征；随后，CDCM 利用下采样结构降低特征图的空间维度，减少计算复杂度，再通过卷积层与双线性插值恢复特征图的原始分辨率，确保细节信息的保留，残差连接确保输入信息的有效传递；最终，通过逐元素相加，将初始输入特征与处理后的特征进行融合，再经过 ReLU 激活，输出优化后的特征表示。在 A_{CSITM} 之后，提出跨域交互式 Transformer，其核心意义是利用多尺度特征交换的特性处理和共享提取后的输入特征。

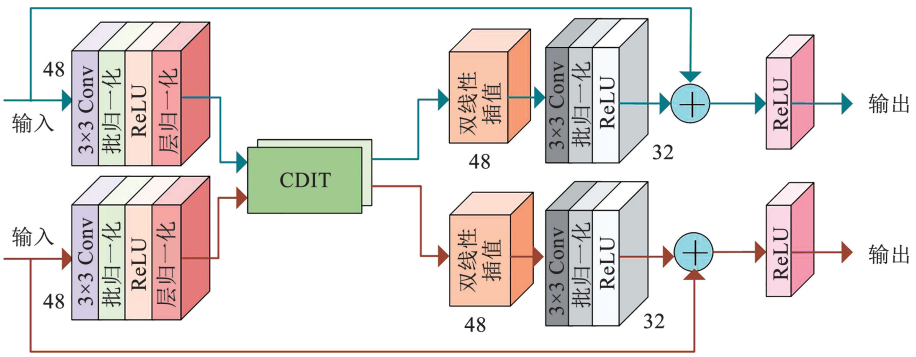


图 4　交叉约束模块结构

Fig. 4　Structure of cross-domain constrain module

在 CDIT 中，输入特征首先经过 C_{DP} 进行预处理；然后分别通过 A_{MH1} 和 A_{MH2} 在各自域内和跨域之间进行交互注意力计算，确保不同任务域的信息有效融合；接着，特征经过 N_L 和 P_{ML} 处理，进一步增强特征的非线性表示能力；最终，来自不同图像的特征通过逐元素相加进行融合，生成各自的输出。CDIT 结构如图 5 所示。

CDIT 模块整体过程表达式为：

$$\langle \boldsymbol{Q}_{l-1}^{G},\boldsymbol{K}_{l-1}^{G},\boldsymbol{V}_{l-1}^{G}\rangle = C_{DP}(\boldsymbol{F}_{l-1}^{G},k) \tag{6}$$

$$\langle \boldsymbol{Q}_{l-1}^{D},\boldsymbol{K}_{l-1}^{D},\boldsymbol{V}_{l-1}^{D}\rangle = C_{DP}(\boldsymbol{F}_{l-1}^{D},k) \tag{7}$$

$$\boldsymbol{F}_{l}^{G'} = N_L(A_{MH1}(\boldsymbol{Q}_{l-1}^{G},\boldsymbol{K}_{l-1}^{G},\boldsymbol{V}_{l-1}^{G})) + \boldsymbol{F}_{l-1}^{G} \tag{8}$$

$$\boldsymbol{F}_{l}^{D'} = N_L(A_{MH1}(\boldsymbol{Q}_{l-1}^{D},\boldsymbol{K}_{l-1}^{D},\boldsymbol{V}_{l-1}^{D})) + \boldsymbol{F}_{l-1}^{D} \tag{9}$$

$$\boldsymbol{F}_{l}^{G} = N_L(P_{ML}(\boldsymbol{F}_{l}^{G'})) + \boldsymbol{F}_{l}^{G'} \tag{10}$$

$$\boldsymbol{F}_{l}^{D} = N_L(P_{ML}(\boldsymbol{F}_{l}^{D'})) + \boldsymbol{F}_{l}^{D'} \tag{11}$$

式中：$\boldsymbol{Q}_{l-1}^{G}$、$\boldsymbol{K}_{l-1}^{G}$、$\boldsymbol{V}_{l-1}^{G}$ 和 $\boldsymbol{Q}_{l-1}^{D}$、$\boldsymbol{K}_{l-1}^{D}$、$\boldsymbol{V}_{l-1}^{D}$ 分别为生成性任务和区别性任务的查询矩阵、键矩阵和值矩阵，$\boldsymbol{F}_{l-1}^{G}$、$\boldsymbol{F}_{l-1}^{D}$ 为 CDIT 模块的输入特征，$\boldsymbol{F}_{l}^{G'}$、$\boldsymbol{F}_{l}^{D'}$ 为中间特征。

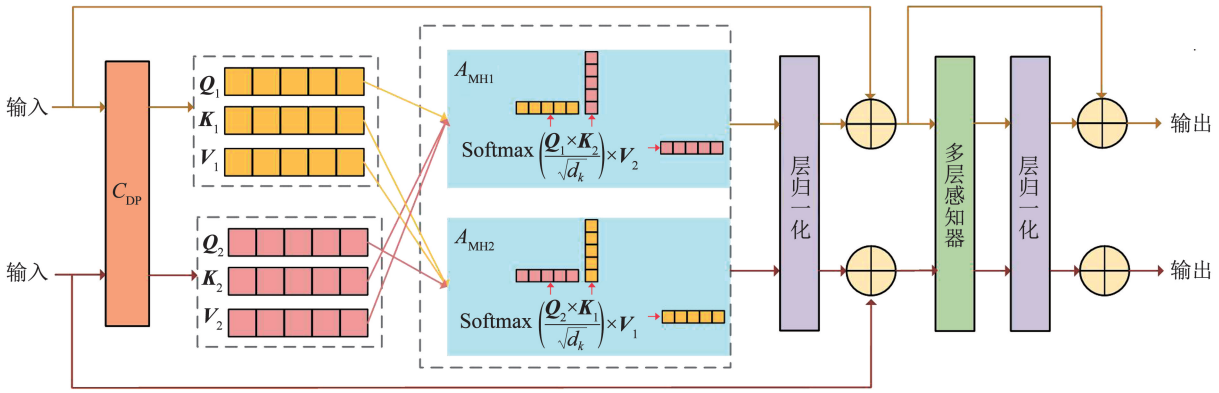


图 5　跨域交互式 Transformer 模块结构

Fig. 5　Structure of cross-domain interactive Transformer module

1.3　基于 Transformer 的融合策略

基于 Transformer 的融合策略如图 6 所示，其中 φ_{vis}^m 和 φ_{ir}^m 表示编码网络提取 $m\in\{1,2,3,4\}$ 尺度可见光和红外图像的深层特征，φ_f^m 为通过变压器融合的深层特征。借助多头自注意力模块构建长距离依赖关系并学习全局上下文特征。将红外和可见光特征通过基于 Transformer 的融合模块进行融合，并利用 C_{DP} 建立特征图各元素之间的联系。本文将基于 Transformer 的融合策略应用于多个尺度上，经过特征解码后得到融合图像。

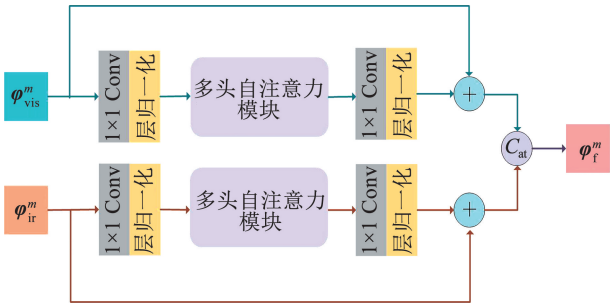


图 6　基于 Transformer 的融合策略

Fig. 6　Fusion strategy based on Transformer

1.4　损失函数

在图像融合网络中，损失函数决定融合图像中关键信息的比例与质量。为此，本文设计了总体损失函数 L_{total}、像素强度损失函数 L_{pixel} 和细节损失函数 L_{detail}。各损失函数的表达式分别为：

$$L_{pixel}=\frac{1}{HW}\sum_{i}^{H}\sum_{j}^{W}\left|I_f^{(i,j)}-\max\left(I_{ir}^{(i,j)},I_{vis}^{(i,j)}\right)\right| \tag{12}$$

$$L_{detail}=\frac{1}{HW}\sum_{i}^{H}\sum_{j}^{W}\left|\nabla I_f^{(i,j)}-\max\left(\nabla I_{ir}^{(i,j)},\nabla I_{vis}^{(i,j)}\right)\right| \tag{13}$$

$$L_{total}=\alpha L_{pixel}+\beta L_{detail} \tag{14}$$

式中：α 和 β 分别为平衡像素强度损失和细节损失的系数，i 和 j 为输入图像的像素坐标，H 和 W 分别为图像的高度和宽度，∇ 为 Sobel 梯度算子。

通过采用最大值选择策略，损失函数在计算过程中能够有效增强融合图像中红外目标的显著性，使红外信息在结果中更为突出。

2　实验结果与分析

2.1　实验数据集和实验设置

在 TNO[19] 数据集的 4×10^4 对红外与可见光图像上训练编码器和解码器网络，并在 TNO 数据集的 42 对图像及 Road scene[20] 数据集的 40 对图像上测试网络性能。

实验在 Windows 11 操作系统上进行，软件环境为 Python 3.7，硬件配置为 13 代英特尔酷睿处理器、2.50 GHz，搭配 GPU。训练与测试均在 PyCharm 编译器中进行。所有图像大小均调整为 256×256 像素，超参数 $\alpha=1$、$\beta=10$，所有实验的学习率均为 1×10^{-4}，epoch 为 4，batch size 为 2。

2.2　对比网络和评价指标

为验证本文网络的优越性，将所提网络与 9 种目前最先进的基于深度学习的图像融合网络（SwinFusion[14]、CrossFuse[18]、DenseFuse[21]、DIDFuse[22]、SDNet[23]、IFCNN[24]、PIAFusion[25]、TarDAL[26] 和 MTDFusion[27]）进行对比。

为客观比较不同网络的融合性能，采用信息熵（entropy，E_N）、标准差（standard deviation，S_D）、空间频率（spatial frequency，S_F）、视觉保真度（visual fidelity，V_{IF}）、平均梯度（average gradient，A_G）、差异相关系数（sum of correlations of differences，S_{CD}）、多层级结构相似性（multi-scale structure similarity，$M_{I\text{-}SSIM}$）和边缘信息保留质量（quality metric based on edge preservation，$Q^{AB/F}$）8 种客观评价指标对融合图像进行评价。E_N 表示图像信息量的丰富程度，E_N 越高，图像的信息量越大。S_D 反映图像灰度值的离散程度，S_D 越大，图像的对比度越高。S_F 衡量图像

细节变化的快慢，S_F 越高，图像的清晰度越高。V_{IF} 评估融合图像与原始图像在感知上的相似度，其值越高，视觉质量越好。A_G 表示图像灰度变化的速度，A_G 越大，图像的边缘和细节越清晰。S_{CD} 衡量融合图像与源图像之间的相关性，较高的 S_{CD} 表示融合图像更好地保留了源图像的特征。$M_{I\text{-}SSIM}$ 衡量融合图像与参考图像在不同尺度上的结构相似性，是衡量图像结构保真度的综合指标。$Q^{AB/F}$ 专门用于评价图像融合质量，尤其关注边缘和纹理的保留情况。

2.3 融合策略实验分析

在融合模块中，比较了最大值（max）、平均值（avg）、相加（add）、基于核范数（nuclear-norm）、视觉显著性注意力机制融合策略（saliency concatenation attention，SCA）与本文选取的融合策略对融合结果的影响，不同融合策略下 TNO 数据集 Marne_04 图组中的两对图像主观比较如图 7 所示。

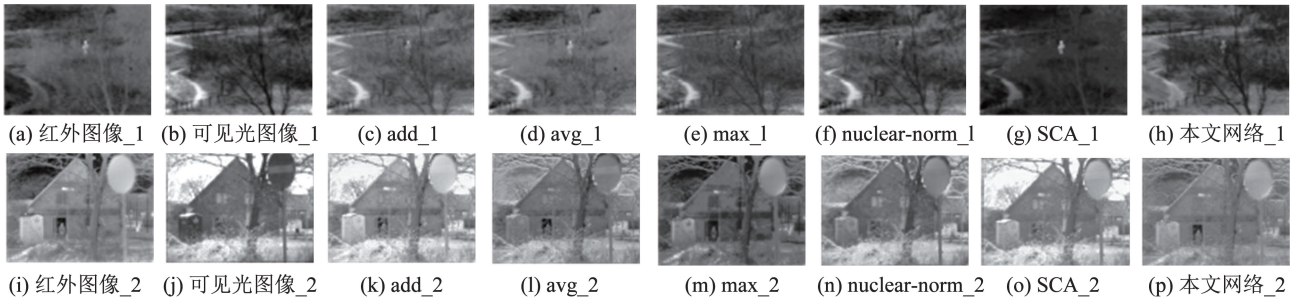


图 7 不同融合策略下 TNO 数据集 Marne_04 图组的主观比较

Fig. 7 Subjective comparison of TNO dataset Marne_04 under different fusion strategies

由图 7 可知，对比其他融合策略，本文提出的基于 Transformer 的融合策略能生成更清晰、视觉效果更佳的图像，体现出了所提策略在提升融合性能方面的优势。

2.4 对比实验

2.4.1 TNO 数据集对比实验

为验证所提网络在图像融合任务中的融合性能，在 TNO 数据集 42 对红外与可见光图像对上与上述 9 种红外和可见光图像融合网络开展定性和定量对比实验，结果如图 8 和表 1 所示。

由图 8 可以看出：DenseFuse 和 MTDFusion 未能有效平衡红外和可见光图像的信息，丢失了较多信息，如树木的边缘轮廓等；DIDFuse 结果对比度较低；SwinFusion 融合图像的红外显著信息清晰，但轮廓边缘模糊；SDNet 和 TarDAL 部分区域出现了伪影情况，不利于观察；IFCNN 的人物边缘轮廓比较模糊；PIAFusion 和 CrossFuse 均能在充分保留可见光丰富纹理细节的同时，突出红外图像中的显著目标，但融合图像出现了不同程度的伪边缘现象；本文网络能够全面集成源图像的边缘细节、纹理及强度信息，同时确保关键特征在融合图像中得到充分保留，从而生成高质量的图像。

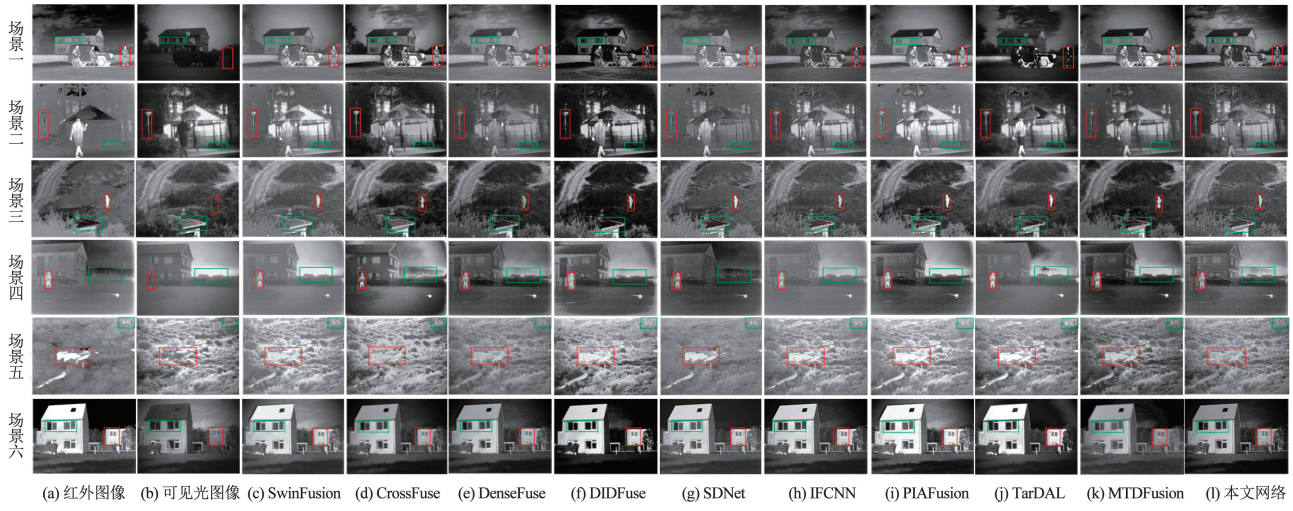


图 8 TNO 数据集对比实验主观比较

Fig. 8 Subjective comparison of TNO dataset comparison experiment

由表 1 可以看出，本文网络的 E_N、S_D、V_{IF}、A_G、S_{CD}、$M_{I\text{-}SSIM}$ 和 $Q^{AB/F}$ 均为最优，S_F 为次优，本文网络在整体视觉质量、信息保留和结构相似性方面进行了优化，导致在特定的纹理和细节增强即 S_F 上略微有

所折中。然而，这样的次优表现并不影响本文网络在其他关键指标上的突出表现，并且仍然达到了综合性能的最佳效果。同时，与其他 9 种网络的 E_N、S_D、S_F、V_{IF}、A_G、S_{CD}、$M_{I\text{-}SSIM}$ 和 $Q^{AB/F}$ 的平均值相比，本文网络分别提升了 10.5%、27.8%、19.5%、17.1%、24.1%、12.5%、14.7% 和 13.3%，说明本文网络能在特征层面还原源图像，更完整地保留了源图像信息，进而产生了更符合人眼偏好的融合效果。

表 1　TNO 数据集对比实验客观评价结果

Tab. 1　Objective evaluation results of TNO dataset comparison experiment

网络	E_N	S_D	S_F	V_{IF}	A_G	S_{CD}	$M_{I\text{-}SSIM}$	$Q^{AB/F}$	t/s
SwinFusion	6.817	39.364	13.185	0.677	5.305	1.714	0.685	0.524	**0.549**
CrossFuse	7.118	45.926	12.512	0.657	5.289	1.737	0.669	0.498	3.155
DenseFuse	6.801	35.332	9.648	0.641	3.789	1.726	0.625	0.502	0.913
DIDFuse	6.431	27.544	6.555	0.413	2.535	1.275	0.525	0.484	2.395
SDNet	6.582	30.985	12.755	0.553	5.008	1.473	0.599	0.426	0.759
IFCNN	6.452	40.450	**14.551**	0.622	5.747	1.753	0.581	0.448	1.294
PIAFusion	6.908	40.682	11.082	0.727	4.422	1.578	0.703	0.437	0.704
TarDAL	6.944	49.437	8.518	0.541	3.298	1.508	0.573	0.482	3.077
MTDFusion	6.926	40.450	11.253	0.633	4.299	1.613	0.603	0.473	2.941
本文网络	**7.585**	**53.903**	13.807	**0.733**	**5.811**	**1.826**	**0.725**	**0.548**	0.783

2.4.2　Road scene 数据集对比实验

为进一步验证本文网络的合理性，采用 Road scene 数据集中的 40 对源图像进行测试，并从中选取了 3 组典型图像进行主客观对比。主观对比结果如图 9 所示，客观对比结果见表 2。



图 9　Road scene 数据集对比实验主观比较

Fig. 9　Subjective comparison of Road scene dataset comparison experiment

表 2　Road scene 数据集对比实验客观评价结果

Tab. 2　Objective evaluation results of Road scene dataset comparison experiment

网络	E_N	S_D	S_F	V_{IF}	A_G	S_{CD}	$M_{I\text{-}SSIM}$	$Q^{AB/F}$	t/s
SwinFusion	6.670	41.794	12.577	0.609	4.526	1.603	0.471	0.453	0.621
CrossFuse	6.800	39.816	17.325	0.585	6.143	1.709	0.713	0.489	3.268
DenseFuse	7.226	45.059	13.564	0.578	5.222	**1.854**	0.719	0.486	0.912
DIDFuse	6.705	29.650	10.710	0.519	4.102	1.509	0.616	0.654	2.644
SDNet	7.391	49.010	16.597	0.566	6.596	1.638	**0.879**	0.486	0.741
IFCNN	7.093	37.554	16.402	0.566	6.399	1.648	0.610	0.524	1.346
PIAFusion	6.732	41.993	13.155	0.714	4.729	1.527	0.381	0.477	0.733
TarDAL	7.638	45.273	17.917	0.402	4.425	1.516	0.721	0.291	3.597
MTDFusion	7.022	37.623	11.570	0.557	6.630	1.594	0.719	0.444	2.224
本文网络	**7.769**	**49.156**	**19.183**	**0.772**	**7.421**	1.800	0.843	**0.721**	0.767

由图 9 可以看出：DIDfuse 和 SwinFusion 网络较好地保存了辐射目标信息，但丢失了部分纹理细节信息，如场景二中的车牌号以及道路标牌信息；IFCNN 和 PIAFusion 虽能有效突出红外目标，却在细节保留上有所欠缺，如场景三中的灌木丛轮廓纹理信息；SDNet 和 DenseFuse 网络能够有效保存纹理

细节信息,但缺乏红外显著目标,如场景三中的道路指示牌纹理细节模糊;TarDAL 融合图像出现失真现象,尤其是场景三中的天空区域偏暗;MTDFusion 和 CrossFuse 高效地融合了红外和可见光信息,但路面细节信息不够突出;本文网络的融合结果兼具完整显著的热目标与丰富的可见光细节,其综合性能优于对比网络。

由表 2 可以看出:本文网络的 E_N、S_D、S_F、V_{IF}、A_G 和 $Q^{AB/F}$ 均为最优,证明融合图像信息较为全面、纹理细节清晰、对比度较好且与主观视觉一致;S_{CD} 和 $M_{I\text{-}SSIM}$ 为次优,两个指标略逊于其他网络。与其他 9 种网络的 E_N、S_D、S_F、V_{IF}、A_G、S_{CD}、$M_{I\text{-}SSIM}$ 和 $Q^{AB/F}$ 的平均值相比,本文网络分别提升了 8.6%、15.2%、22.3%、24.0%、24.3%、8.9%、20.8% 和 30.3%,说明融合后的图像信息量大、细节丰富、对比度高,验证了本文网络的客观有效性。

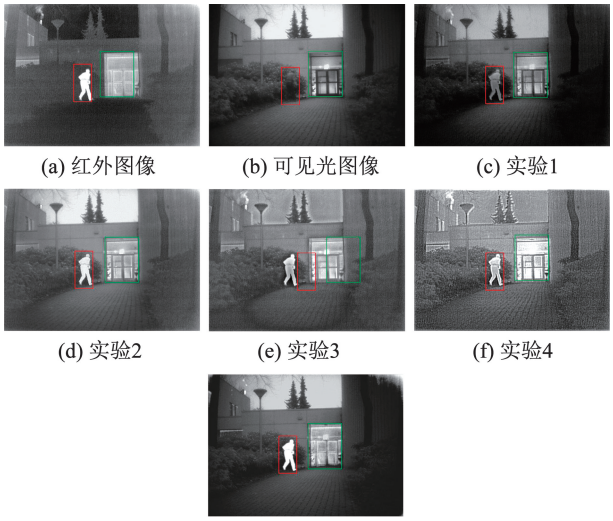
2.5 运行效率对比实验

融合网络的执行时间同样是衡量模型优越性的关键指标。从表 1 和表 2 的运行时间(t)对比可知,本文网络的计算开销略高于 SDNet、PIAFusion 和 SwinFusion,但这 3 种网络在融合图像的细节质量与对比度方面均存在明显不足,反映出其性能与效率间的权衡局限。综合融合效果与运行效率,本文网络整体性能优越。

2.6 消融实验

为进一步说明本文所提网络的有效性,通过消融实验分析跨尺度空间注意力和交叉注意力对网络性能的影响。设计 5 种消融实验:实验 1(去掉跨尺度视觉注意力中的跨尺度交互式 Transformer)、实验 2(去掉跨尺度视觉注意力模块)、实验 3(去掉交叉注意力中的跨域交互式 Transformer)、实验 4(去掉交叉注意力模块)和本文所提完整网络。在 TNO 测试集对选取的 20 组图像进行消融实验,分别得出 E_N、A_G、S_F、V_{IF} 和 S_D 评价指标数据,计算平均值进行客观分析。主观对比结果如图 10 所示,客观对比结果见表 3。

由图 10 可以看出,本文网络所得融合图像红外特征显著、目标清晰可辨,可见光纹理细节丰富、边缘轮廓清晰,整体视觉效果更佳。由表 3 可知,在 5 个评价指标中,本文网络均为最优,说明本文网络能有效提取源图像的信息,融合图像清晰。总体而言,本文网络无论是主观视觉还是客观评价指标均优于其他 4 个降级的模型框架,证明了本文网络中跨尺度视觉和交叉注意力的有效性。



(a) 红外图像 (b) 可见光图像 (c) 实验1

(d) 实验2 (e) 实验3 (f) 实验4

(g) 本文网络

图 10 消融实验主观对比结果

Fig. 10 Subjective comparison results of ablation experiment

表 3 消融实验的客观对比结果

Tab. 3 Objective comparison results of ablation experiment

实验	E_N	A_G	S_F	V_{IF}	S_D
实验 1	6.654	4.852	12.257	0.678	42.554
实验 2	7.358	5.248	10.254	0.654	40.232
实验 3	7.025	4.895	11.277	0.695	43.268
实验 4	6.985	4.698	12.021	0.667	41.257
本文网络	**7.498**	**5.687**	**13.245**	**0.729**	**47.529**

3 结 论

1)设计了跨尺度视觉注意力模块,在弱光照条件下强化了特征提取,引导注意力权重向低对比度区域倾斜,优化了融合图像的视觉效果。

2)设计了交叉约束模块,通过全局特征的交互与融合,精细地保留并强化边缘、纹理等关键信息,有效减少了局部细节的损失。

3)采用基于 Transformer 的融合策略,通过多头自注意力机制构建可见光与红外深层特征间的长程依赖,在各尺度上建立特征图元素间的广泛联系,提升了融合性能。

4)在 TNO 数据集上,与其他 9 种网络的 E_N、S_D、S_F、V_{IF}、A_G、S_{CD}、$M_{I\text{-}SSIM}$ 和 $Q^{AB/F}$ 的平均值相比,跨尺度交叉注意力的 Transformer 网络分别提升了 10.5%、27.8%、19.5%、17.1%、24.1%、12.5%、14.7%、13.3%;在 Road scene 数据集上,与其他 9 种网络的 E_N、S_D、S_F、V_{IF}、A_G、S_{CD}、$M_{I\text{-}SSIM}$ 和 $Q^{AB/F}$ 的平均值相比,跨尺度交叉注意力的 Transformer 网络分别提升了 8.6%、15.2%、22.3%、24.0%、24.3%、8.9%、20.8%、30.3%,说明本文网络生成的融合图像目标显著、纹理细节丰富、对比度高。

未来的研究将进一步优化网络结构，提升图像融合质量，并将该网络应用于更多的实际场景和数据集，以验证其适用性和鲁棒性。

参考文献

[1] Liu Yu, Wang Lei, Cheng Juan, et al. Multi-focus image fusion: a survey of the state of the art[J]. Information Fusion, 2020, 64: 71. DOI: 10.1016/J.INFFUS.2020.06.013

[2] He Guiqing, Ji Jiaqi, Dong Dandan, et al. Infrared and visible image fusion method by using hybrid representation learning[J]. IEEE Geoscience and Remote Sensing Letters, 2019, 16(11): 1796. DOI: 10.1109/LGRS.2019.2907721

[3] Chen Lei, Han Jun, Tian Feng. Infrared and visible image fusion using two-layer generative adversarial network [J]. Journal of Intelligent & Fuzzy Systems, 2021, 40(6): 11897. DOI: 10.3233/JIFS-210041

[4] Ma Jiayi, Ma Yong, Li Chang. Infrared and visible image fusion methods and applications: a survey[J]. Information Fusion, 2019, 45: 153. DOI: 10.1016/J.INFFUS.2018.02.004

[5] Zhang Yihan, Fang Yichu, Zhang Qian. Focus fusion network for visible and infrared image fusion [C]//2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul: IEEE, 2024: 3850. DOI: 10.1109/ICASSP48485.2024.10445881

[6] Xu Xing, Shen Yilan, Han Shizhong. Dense-FG: a fusion GAN model by using densely connected blocks to fuse infrared and visible images[J]. Applied Sciences, 2023, 13(8): 4684. DOI: 10.3390/app13084684

[7] Gao Mingliang, Zhou Yinan, Zhai Wenzhe, et al. Saregan: a salient regional generative adversarial network for visible and infrared image fusion[J]. Multimedia Tools and Applications, 2024, 83(22): 61659. DOI: 10.1007/S11042-023-14393-2

[8] Dogra A, Goyal B, Agrawal S. From multi-scale decomposition to non-multi-scale decomposition methods: a comprehensive survey of image fusion techniques and its applications[J]. IEEE Access, 2017, 5: 16040. DOI: 10.1109/ACCESS.2017.2735865

[9] 张洲宇，曹云峰，丁萌，等. 采用多层卷积稀疏表示的红外与可见光图像融合[J]. 哈尔滨工业大学学报，2021，53(12)：51
Zhang Zhouyu, Cao Yunfeng, Ding Meng, et al. Infrared and visible image fusion using multilayer convolutional sparse representation[J]. Journal of Harbin Institute of Technology, 2021, 53(12): 51. DOI: 10.11918/202005038

[10] Yin Wenxia, He Kangjian, Xu Dan, et al. Significant target analysis and detail preserving based infrared and visible image fusion[J]. Infrared Physics & Technology, 2022, 121: 104041. DOI: 10.1016/J.INFRARED.2022.104041

[11] Liu Jianjun, Shen Dunbin, Wu Zebin, et al. Patch-aware deep hyperspectral and multispectral image fusion by unfolding subspace-based optimization model[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 1024. DOI: 10.1109/JSTARS.2022.3140211

[12] Prabhakar R K, Srikar S V, Venkatesh B R. DeepFuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs[C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017: 4714. DOI: 10.1109/ICCV.2017.505

[13] Li Hui, Wu Xiaojun, Kittler J. RFN-Nest: an end-to-end residual fusion network for infrared and visible images [J]. Information Fusion, 2021, 73: 72. DOI: 10.1016/J.INFFUS.2021.02.023

[14] Ma Jiayi, Tang Linfeng, Fan Fan, 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. DOI: 10.1109/JAS.2022.105686

[15] 李永萍，杨艳春，党建武，等. 基于变换域 VGGNet19 的红外与可见光图像融合[J]. 红外技术，2022，44(12)：1293
Li Yongping, Yang Yanchun, Dang Jianwu, et al. Infrared and visible image fusion based on transform domain VGGNet19 [J]. Infrared Technology, 2022, 44(12): 1293. DOI: 10.3969/j.issn.1001-8891.2022.12.009

[16] 刘光辉，杨琦，孟月波，等. 一种并行混合注意力的渐进融合图像增强方法[J]. 光电工程，2023，50(4)：52
Liu Guanghui, Yang Qi, Meng Yuebo, et al. A progressive fusion image enhancement method with parallel hybrid attention[J]. Opto-Electronic Engineering, 2023, 50(4): 52. DOI: 10.12086/oee.2023.220231

[17] Tang Linfeng, Deng Yuxin, Ma Yong, et al. SuperFusion: a versatile image registration and fusion network with semantic awareness[J]. IEEE/CAA Journal of Automatica Sinica, 2022, 9(12): 2121. DOI: 10.1109/JAS.2022.106082

[18] Li Hui, Wu Xiaojun. CrossFuse: a novel cross attention mechanism based infrared and visible image fusion approach[J]. Information Fusion, 2024, 103: 102147. DOI: 10.1016/J.INFFUS.2023.102147

[19] Toet A. The TNO multiband image data collection[J]. Data in Brief, 2017, 15: 249. DOI: 10.1016/J.DIB.2017.09.038

[20] Alvarez J M, Gevers T, Lecun Y, et al. Road scene segmentation from a single image[C]//2012 European Conference on Computer Vision. Florence: ECCV, 2012: 376. DOI: 10.1007/978-3-642-33786-4_28

[21] Li Hui, Wu Xiaojun. DenseFuse: a fusion approach to infrared and visible images[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2614. DOI: 10.1109/TIP.2018.2887342

[22] Zhao Zixiang, Xu Shuang, Zhang Chunxia, et al. DIDFuse: deep image decomposition for infrared and visible image fusion[PP/OL]. V1. arXiv (2020-03-20) [2024-06-20]. https://doi.org/10.48550/arXiv.2003.09210

[23] 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. DOI: 10.1007/S11263-021-01501-8

[24] Zhang Yu, Liu Yu, Sun Peng, et al. IFCNN: a general image fusion framework based on convolutional neural network [J]. Information Fusion, 2020, 54: 99. DOI: 10.1016/J.INFFUS.2019.07.011

[25] 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. DOI: 10.1016/J.INFFUS.2022.03.007

[26] 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]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022: 5802. DOI: 10.1109/CVPR52688.2022.00571

[27] Karim S, Tong Geng, Li Jinyang, et al. MTDFusion: a multilayer triple dense network for infrared and visible image fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 73: 1. DOI: 10.1109/TIM.2023.3329148