Journal of Image and Signal Processing 图像与信号处理, 2025, 14(4), 377-386
Published Online October 2025 in Hans. https://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2025.144035

Hans汉斯

基于Transformer-CNN的红外与可见光融合网络

李　玉，张志超，祁艳杰*

太原科技大学电子信息工程学院，山西　太原

收稿日期：2025年8月22日；录用日期：2025年9月12日；发布日期：2025年9月24日

摘　要

红外与可见光图像融合是多模态融合领域的重要研究方向之一，广泛应用于医药、工业、自动驾驶等领域。本文提出了一种基于Transformer-CNN网络的融合算法。首先，构建双分支特征提取网络提取不同模态图像的特征信息；然后，在融合阶段通过模态间特征差分计算提取互补特征，采用参数自适应的Swish激活函数动态生成通道权重，结合全局平均池化压缩与特征级联策略，实现红外与可见光模态的跨尺度特征融合。最后，提出一种掩码损失以提升融合质量。在公开数据集上的实验结果表明，该方法在进行红外和可见光图像融合后图像具有清晰的背景纹理细节。在主观和客观评价上均能取得较好或相近的结果。

关键词

图像融合，Transformer，参数自适应Swish函数，混合结构，红外与可见光图像

Infrared and Visible Image Fusion Network Based on Transformer-CNN

Yu Li, Zhichao Zhang, Yanjie Qi*

School of Electronic Information Engineering, Taiyuan University of Science and Technology, Taiyuan Shanxi

Received: August 22, 2025; accepted: September 12, 2025; published: September 24, 2025

Abstract

Infrared and visible image fusion is a significant research direction in multimodal fusion, with extensive applications in medicine, industry, autonomous driving, and other fields. This paper proposes a fusion algorithm based on a Transformer-CNN hybrid network. First, a dual-branch feature

*通讯作者。

文章引用: 李玉, 张志超, 祁艳杰. 基于 Transformer-CNN 的红外与可见光融合网络[J]. 图像与信号处理, 2025, 14(4): 377-386. DOI: 10.12677/jisp.2025.144035

extraction network is constructed to extract features from different modal images. Subsequently, during the fusion stage, complementary features are extracted through cross-modal feature difference computation. A parameter-adaptive Swish activation function is employed to dynamically generate channel weights, which are combined with global average pooling for feature compression and cascading strategies, enabling cross-scale feature fusion of infrared and visible modalities. Finally, a mask loss is introduced to enhance fusion quality. Experimental results on public datasets demonstrate that the fused images produced by this method exhibit clear background texture details. Both subjective evaluation and objective metrics indicate that our approach achieves superior or comparable results.

Keywords

Image Fusion, Transformer, Parameter-Adaptive Swish Function, Hybrid Architecture, Infrared and Visible Images

1. 引言

图像融合算法将两种或两种以上的互补图像融合成信息丰富的图像。新的融合图像具有更详细地描述对应场景的特点，更便于人类视觉捕捉或机器感知识别。利用不同模态传感器的特性，捕捉可见光和红外图像的光反射和热辐射。因此，红外和可见光图像的融合可以将不同传感器的互补信息结合起来，提供对图像场景的全面表征，提高高级视觉任务的性能。它们捕获不同的光谱范围并显示部分场景信息，每个都有自己的缺点和有限的信息。

可见图像具有丰富的纹理细节和明显的对比度，与人类的视觉感知一致。可见图像容易受到诸如障碍物、天气和光照等环境因素的影响。红外图像对热辐射信息具有鲁棒性，并且不容易受到光线和天气变化的影响。但是在纹理和其他细节上表现不佳。因此，红外与可见光图像的融合得到了广泛的关注，并应用于许多领域[1]，如军事场景中的目标分类和检测[2]、医学图像处理[3]等。然而，目前用于融合红外和可见光图像的方法主要考虑视觉性能，往往忽略了后续高级视觉任务的必要条件。

传统的图像融合算法最大的缺点是需要人工调整融合参数，而基于深度学习的图像融合算法可以避免人工调整参数[4]。基于深度学习的图像融合算法具有数据驱动、学习能力强、可移植性好等特点。基于深度学习的融合算法主要有无监督自编码器方法[5]、有监督 CNN 方法[6]和 GAN 方法[7]。自编码器方法通常在一些大型数据集上进行预训练，以实现特征提取和图像重建。自编码器方法的优点是框架不复杂，易于训练，但需要找到合适的融合策略。

在本文中，提出了一种基于 Transformer-CNN 的红外和可见光图像融合算法。利用红外图像掩码和可见光图像掩码两种掩码来引导网络更好地搜索显著目标。所使用的掩码不仅可以确定重要目标的位置，而且可以区分目标的重要性。这样，本文的网络可以更好地突出重要目标，同时保留其他重要信息的细节。本研究是通过特殊的编解码器结构和精心设计的融合策略，进一步提高图像融合算法的性能，使所提出的算法能够适应多领域或多场景。本文将 Transformer 与自编码器网络相结合，提出了一种通用的图像融合算法。与常见的图像融合方法相比，该方法不仅考虑视觉性能和统计指标，而且强调解决后续高级视觉任务的纹理和对比度需求。该方法通过改善目标区域的纹理和对比度，有效地促进了后续的高级视觉任务。

2. 网络架构

本文的核心结构包括编码器、融合层和解码器，具体算法结构如图 1 所示。待融合的红外与可见光图像分别为 I_1 和 I_2，通过本算法融合的图像是 I_f。获取 I_f 的步骤如下：r_1 和 r_2 是由使用输入为 I_1 和 I_2 的自编码器生成。采用跨模态差异模块的自编码器融合层使用输入 r_1 和 r_2 生成 r_f。r_f 为融合层的输出和解码器的输入。融合图像 I_f 由解码器和 r_f 生成并重构。

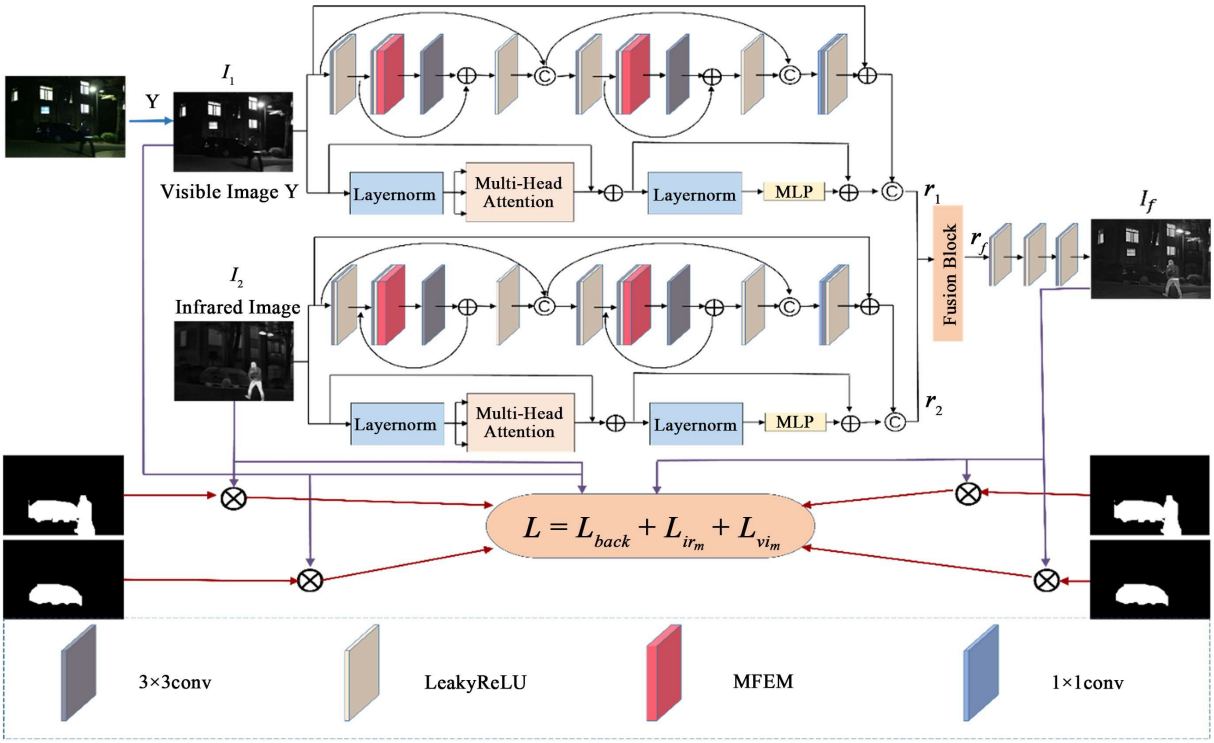

Figure 1. Network framework
图 1. 网络框架

2.1. 编码器结构

在本文提出的算法中，编码器主体部分采用 Transformer-CNN 结构的混合模型。这种混合结构的优点是可以使用 CNN 提取原始图像的局部特征，并且利用 Transformer 强大的全局提取能力，进一步提高了全局信息感知能力，提高图像融合的通用性。

混合编码器结构中的 CNN 分支，采用泛化能力较强的多尺度特征提取模块。CNN 分支由 3 × 3 卷积块、MFEM 模块组成，将 MFEM 块嵌套到密集连接的残差结构中。编码器中的 CNN 分支具有良好的特征提取能力[8]。MFEM 模块由空间注意分支(SAB)、通道注意分支(CAB)并行分支和全局平均池化、多尺度卷积层(MSCB)和增强分支(EB)组成，如图 2 所示。给定输入特征 $X \in \mathbb{R}^{C\times H\times W}$，MFEM 过程定义为：

$$C_o = \left(\mathrm{EB}\left(\mathrm{MSCB}\left(\left[\mathrm{SAB}(X), \mathrm{CAB}(X)\right]\right)\right)\odot \mathrm{EB}(X)\right) + X = E\left(\left(\mathrm{MSCB}\left(\left[SA, CA\right]\right)\right)\right) + X \tag{1}$$

式中，$C_o \in \mathbb{R}^{C\times H\times W}$ 为 MFEM 的输出；SA、CA、E 分别是 SAB、CAB、EB 模块的输出。

1) 空间注意力分支：如图 2 所示，给定输入特征 $X \in \mathbb{R}^{C\times H\times W}$，沿着通道维度应用平均池化和最大池化来捕获不同的语义响应。然后将这些响应进行串接和聚合，得到空间权重，再与输入特征 X 相乘，得

到输出特征 $SA \in \mathbb{R}^{C\times H\times W}$ 。

$$SA = f_{3\times 3}^{s}\left(\left[f_{spatial}^{Avg}(X), f_{spatial}^{Max}(X)\right]\right) \odot X \tag{2}$$

式中，$f_{3\times 3}^{s}$ 表示 3 × 3 卷积和 sigmoid 激活函数；⊙ 是逐元素相乘；$f_{spatial}^{Avg}(\cdot)$ 和 $f_{spatial}^{Max}(\cdot)$ 表示通道级的平均池化和最大池化。

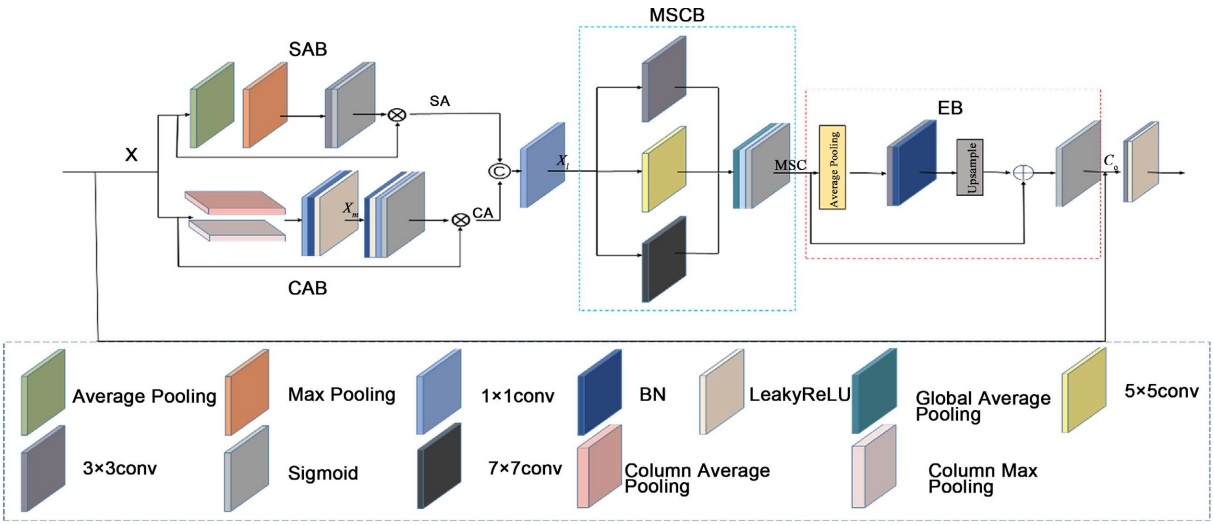

Figure 2. MFEM module
图 2. MFEM 模块

2) 通道注意力分支：如图 2 所示，对输入特征 $X \in \mathbb{R}^{C\times H\times W}$ 的每一列应用列平均池化函数 $f_{column}^{Avg}(\cdot)$ 和列最大池化函数 $f_{column}^{Max}(\cdot)$，其池化核尺寸为(H, 1)。基于列池化的通道注意力机制可以保持空间结构信息：列池化沿高度维度进行池化，保留宽度方向上的空间结构，适用于具有明显水平结构的场景(如地平线、建筑边缘等)。与空间注意力可以互补，空间注意力强调“哪里重要”，通道注意力强调“什么特征重要”，二者并行可实现对特征图在空间与通道维度的协同增强。由于不同池化响应对应相同的列，沿通道维度进行拼接以捕获这种共享特性，拼接后的特征通过共享机制实现信息交互，具体操作如下：

$$X_m = \mathcal{CBL}\left(\left[f_{column}^{Avg}(X), f_{column}^{Max}(X)\right]\right) \tag{3}$$

式中，$\mathcal{CBL}(\cdot)$ 由 1 × 1 的卷积块、批量归一化和 LeakyReLU 激活函数组成。随后，通过注意力模型计算列权重系数，最终生成的通道注意力矩阵 $CA \in \mathbb{R}^{C\times H\times W}$ 可表示为：

$$CA = X \odot \mathcal{F}_{column}^{Avg}(X_m) \tag{4}$$

式中，$\mathcal{F}_{column}^{Avg}(\cdot)$ 表示由一系列卷积核 1 × 1、批量归一化、ReLU 激活函数、Sigmoid 激活函数构成的通道注意力机制模块。

3) 多尺度卷积模块：该模块采用并行多分支结构实现多尺度特征提取，具体表达式为：

$$MSC = \text{sigmoid}\left(FC\left(G\left(Concat\left(Conv_{k\times k}\right)(X_l)\right)\right)\right) \tag{5}$$

式中，每个分支采用不同尺寸的卷积核(3 × 3, 5 × 5, 7 × 7)，各分支特征通过元素相加方式融合。

4) 增强分支：如图 2 所示，利用此分支实现强大的上下文信息建模，并在每个空间位置建立远程依赖关系。给定输入特征 $X \in \mathbb{R}^{C\times H\times W}$，EB 的计算可描述如下：

$$E=\delta_s\left(X+\mathcal{B}_2\left(conv\left(\mathcal{A}_2\left(MSC\right)\right)\right)\right) \tag{6}$$

式中，$\mathcal{B}_2(\cdot)$表示经过 2 倍上采样的双线性插值(Bilinear Interpolation, BI)；$\mathcal{A}_2(\cdot)$是滤波器尺寸为 2 × 2、步长为 2 的平均池化。

编码器中的 Transformer 支路，其结构主要包括两层归一化、多头注意机制和多层感知器(MLP)层。Transformer 模块的计算如式(7)~(10)所示。

$$z_0=\left[X_{class};X_p^1E;X_p^2E;\cdots;X_p^NE\right],\ E\in\mathbb{R}^{\left(p^2\cdot C\right)\times D} \tag{7}$$

$$z'_\ell=\mathrm{MSA}\left(LN\left(Z_{\ell-1}\right)\right)+Z_{\ell-1},\ \ell\in\left[1,2,\cdots,\ell\right] \tag{8}$$

$$Z_\ell=\mathrm{MLP}\left(LN\left(z'_\ell\right)\right)+z'_\ell \tag{9}$$

$$y=LN\left(Z_L^0\right) \tag{10}$$

式中，MSA 为多头注意机制的计算过程；MLP 为多层感知器的计算过程。对于 MSA 的计算，明确给出了单头注意机制的计算过程，如式(11)所示。

$$Attention\left(Q,K,V\right)=\mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{11}$$

式中，Q 为查询向量；K 为键向量；V 为值向量。单头注意力机制通过扩展形成多头注意力机制。该机制对 QKV 向量矩阵进行多次映射，并将结果组合为最终输出。具体计算过程如公式(12)和(13)所示：

$$MultiHead\left(Q,K,V\right)=Cat\left(head_1,head_2,\cdots,head_h\right) \tag{12}$$

$$head_i=Attention\left(Q\cdot W_i^Q,K\cdot W_i^K,V\cdot W_i^V\right) \tag{13}$$

2.2. 融合层

为了更好地整合互补和共有特征，本文提出了跨模态差异融合模块(Cross-Modal Difference Fusion Module, CMDFM)，以进一步提高红外与可见光图像融合模型的融合精度。具体的流程图如图 3 所示。具体来说，该模块以通过 Transformer-CNN 的混合特征提取模块从红外和可见光图像中提取的深度特征作为输入。在 CMDFM 中，通过跨模态差异计算获取互补特征，并采用通道加权过程整合互补信息。其计算过程定义如下：

$$F_{ir}=F_{ir}\odot\zeta\left(G\left(F_{vi}-F_{ir}\right)\right)\otimes\left(F_{vi}-F_{ir}\right) \tag{14}$$

$$F_{vi}=F_{vi}\odot\zeta\left(G\left(F_{ir}-F_{vi}\right)\right)\otimes\left(F_{ir}-F_{vi}\right) \tag{15}$$

式中，$\odot$ 表示通道级联；$\otimes$ 表示通道乘法；ζ 代表激活函数；G 表示全局平均池化，用于压缩特征为向量。该模块特别采用了参数可学习的 Swish 激活函数，能够根据输入数据的分布动态调整，自动适应不同模态的特征分布，并平衡不同模态间的信息差异。

2.3. 解码器

在得到融合特征后，通过解码器重构融合图像的 Y 通道亮度分量。具体来说，解码器包含四个 3 × 3 卷积层，前三层卷积层后采用批量归一化和 Leaky Relu 激活函数，最后一层采用 Tanh 激活函数。Tanh 激活函数的使用保证了变化范围在[−1, 1]内。每经过一层卷积，通道数会减半，同时通过填充操作保持特征图尺寸不变。

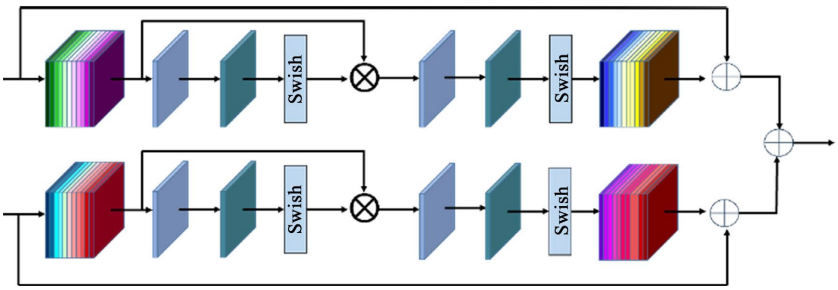

Figure 3. Fusion module
图 3. 融合模块

2.4. 损失函数

本文引入的损失函数指导网络在不同层级融合显著性目标，由三个子损失构成：像素损失、梯度损失和结构相似性损失[9]。像素损失确保融合图像与源图像在像素强度分布上的一致性。具体来说，红外掩模像素损失($L_{pixel}^{ir_m}$)、可见光掩模像素损失($L_{pixel}^{vi_m}$)以及背景像素损失(L_{pixel}^{back})的表达式如下：

$$L_{pixel}^{ir_m}=\frac{1}{HW}\left\|I_{ir_m}\odot\left(I_f-I_{ir}\right)\right\|_1 \tag{16}$$

$$L_{pixel}^{vi_m}=\frac{1}{HW}\left\|I_{vi_m}\odot\left(I_f-I_{vi}\right)\right\|_1 \tag{17}$$

$$L_{pixel}^{back}=\frac{1}{HW}\left\|I_f-I_{ir}\right\|_1+\frac{1}{HW}\left\|I_f-I_{vi}\right\|_1 \tag{18}$$

式中，H 与 W 分别表示图像的高度与宽度；$\|\cdot\|_1$ 为 L1 范数；I_{ir_m} 为红外图像对应的掩模；I_{vi_m} 为可见光图像对应的掩模；运算符 $\odot$ 表示逐元素相乘。

梯度损失的目的是促进源图像梯度信息向融合图像的整合。图像融合的核心目标之一是在融合结果中保留纹理细节，理想情况下融合图像的纹理应为红外与可见光图像纹理的最大化组合。红外掩模梯度损失($L_{gard}^{ir_m}$)、可见光掩模梯度($L_{gard}^{vi_m}$)损失及背景梯度损失(L_{gard}^{back})表达式如下：

$$L_{gard}^{ir_m}=\frac{1}{HW}\left\|I_{ir_m}\odot\left(\nabla I_f-\nabla I_{ir}\right)\right\|_1 \tag{19}$$

$$L_{gard}^{vi_m}=\frac{1}{HW}\left\|I_{vi_m}\odot\left(\nabla I_f-\nabla I_{vi}\right)\right\|_1 \tag{20}$$

$$L_{gard}^{back}=\frac{1}{HW}\left\|\nabla I_f\right|-\max\left(\left|\nabla I_{ir}\right|,\left|\nabla I_{vi}\right|\right)\right\|_1 \tag{21}$$

式中，∇ 表示梯度算子，本文采用 Sobel 算子计算图像梯度。

结构相似性损失用来确保融合图像与源图像保持较高的结构相似性。红外掩模结构相似性损失($L_{ssim}^{ir_m}$)、可见光掩模结构相似性损失($L_{ssim}^{vi_m}$)及背景结构相似性损失(L_{ssim}^{back})表达式如下：

$$L_{ssim}^{ir_m}=1-\mathrm{SSIM}\left(I_{ir_m}\odot I_f,I_{ir_m}\odot I_{ir}\right) \tag{22}$$

$$L_{ssim}^{vi_m}=1-\mathrm{SSIM}\left(I_{vi_m}\odot I_f,I_{vi_m}\odot I_{vi}\right) \tag{23}$$

$$L_{ssim}^{back}=2-\left(\mathrm{SSIM}\left(I_f,I_{ir}\right)+\mathrm{SSIM}\left(I_f,I_{vi}\right)\right) \tag{24}$$

式中，$\mathrm{SSIM}(\cdot)$表示结构相似性指数，通过亮度、对比度和结构三个维度量化图像失真程度。

3. 实验结果和分析

3.1. 实验设置

该算法在最终融合之前需要对模型进行训练，训练过程中的网络结构如图 4 所示。训练模型主要包括编码器和解码器两部分。训练时的网络结构与图 1 中编码器和解码器的网络结构完全相同。在模型训练结束后，训练阶段学习到的编码器与解码器参数将迁移至后期红外与可见光图像融合模型中，作为其基础特征提取与重建模块。

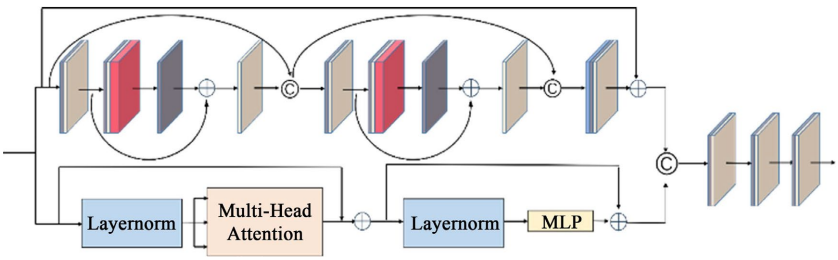

Figure 4. Training network architecture
图 4. 训练网络结构

本实验使用的硬件配置如下：NVIDIA GeForce RTX 3090 24GB、12th Gen Intel(R) Core(TM) i7-12700。编程语言是 python 3.6，深度学习框架使用 pytorch。模型训练阶段使用的数据集为 MS-COCO 数据集，使用的是 2014 年版本的数据集。算法初始参数设为 epoch = 50，batch size = 1，图像初始化大小 256 × 256，学习率为 0.001。选择 Adam 作为神经网络的优化器，样本量为 20,000。

3.2. 主观评价

为了验证本文算法的性能，将本方法与 7 种近几年提出的方法先后在 MSRS [10]和 RoadScene [11]数据集上进行对比实验，包括 U2Fusion [12]、GANMcC [13]、STDFusion [14]、ITFuse [15]、DSFusion [16]、BTSFusion [17]、RFN-Nest [18]，在 MSRS 数据集和 RoadScene 数据集中分别选择三张图片进行对比。图 5 分别显示了不同方法在 MSRS 和 RoadScene 数据集上的定性实验结果。

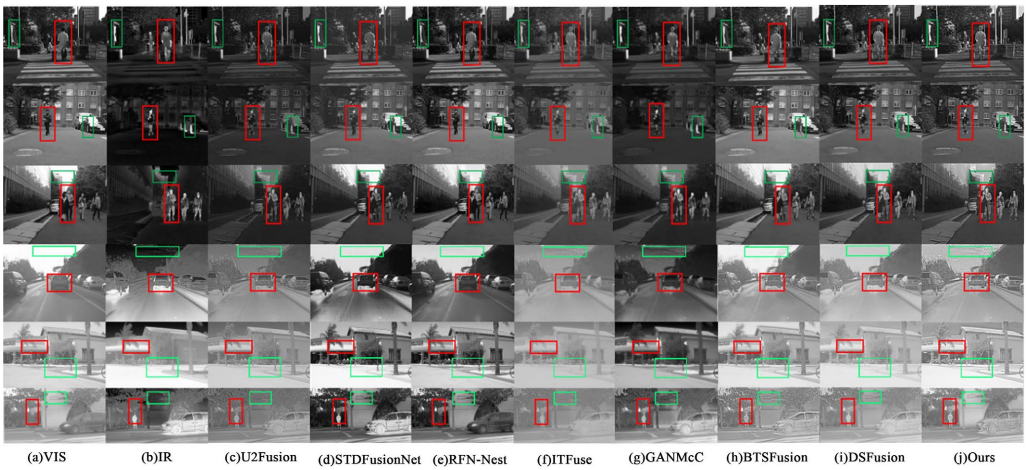

Figure 5. Qualitative comparison of different methods on MSRS and RoadScene datasets
图 5. 不同方法在 MSRS 和 RoadScene 数据集上的定性比较

从图 5 可以看出，STDFusion、RFN 和 BTSFusion 方法保留了更多的可见光细节，但缺乏对红外信息的感知。GANMcC 方法虽然保留了较多的红外信息，但缺乏对可见光特征的感知。其余算法对于人脸或者车轮的纹理细节均未能清晰保留。本算法和 DSFusion 方法可以均衡保留红外信息和可见光信息，但本算法在图像的目标区域表现最好。我们的图像非常完整地保留了人和衣服的纹理信息，清晰地呈现了人的面部特征，保留了衣服的细节。对于天空中的云、房檐和墙上的字来说，也保留了更多的细节，字体也更加清晰可见。结果表明，利用我们设计的编码器结构和融合策略，可以充分融合和提取全局和局部信息，实现红外可见光细节的协同保留，并保持更均匀的纹理特征。此外，本方法对边缘信息的保留效果优异，在定量评估指标中展现出独特优势。

3.3. 客观评价

定量实验结果如表 1 和表 2 所示，分别是 MSRS 数据集和 RoadScene 数据集。在实验数据中，最优值用加粗黑体表示。重要的是，本文的方法在所有七个指标(EN, SSIM, SCD, AG, MI, SF, VIF)上都表现得非常好。通过这些指标强调了我们在解决图像融合任务方面的卓越性。在两个数据集中的 EN、SSIM、SCD、AG、SF、VIF 指标上取得最优值，从而进一步表明所提算法相较于其它 7 种对比算法具有更好的图像融合效果。SSIM 用于衡量融合图像与源图像在结构信息上的一致性，CMDFM 模块通过跨模态特征差分计算，有效捕捉红外与可见光图像之间的互补结构信息，避免冗余特征的重复融合，从而更好地保留源图像的结构一致性。MFEM 模块中的空间与通道注意力机制进一步强化了对重要结构区域的关注，避免细节丢失。SF 指标反映图像的空间细节和边缘清晰度，CMDFM 模块引入的可学习 Swish 激活函数能够动态调整通道权重，增强对高频细节的响应，避免融合过程中的细节平滑。MFEM 模块中的多尺度卷积结构有效捕获不同尺度的边缘和纹理特征，并通过增强分支进一步强化上下文信息，提升细节复原能力。VIF 和 AG 的提升进一步验证了本文方法在视觉感知质量和细节清晰度方面的优势，这与 MFEM 和 CMDFM 的协同作用密切相关。

Table 1. Objective evaluation metrics for comparative experiments on the MSRS dataset
表 1. MSRS 数据集对比实验客观评价指标

Algorithms	EN	SSIM	SCD	AG	MI	SF	VIF
DSFusion	6.218	0.8033	1.2138	3.9525	1.7552	0.0554	0.8529
U2Fusion	5.637	0.8835	1.1755	2.4098	2.453	0.0239	0.8742
GANMcC	6.0848	0.8797	1.3501	2.2943	1.1157	0.014	0.7729
STDFusion	5.5092	0.9085	1.1832	3.3945	1.0352	0.0466	0.8634
ITFuse	5.7776	0.835	1.2407	1.8028	0.8372	0.0252	0.8523
BTSFusion	6.4996	0.8664	1.3737	4.133	**2.546**	0.049	0.8698
RFN-nest	6.4127	0.7386	1.291	3.6921	1.1021	0.0554	0.8974
Ours	**6.5903**	**0.9359**	**1.5821**	**5.012**	1.9559	**0.0562**	**0.9684**

Table 2. Objective evaluation metrics for comparative experiments on the RoadScene dataset
表 2. RoadScene 数据集对比实验客观评价指标

Algorithms	EN	SSIM	SCD	AG	MI	SF	VIF
DSFusion	6.6509	0.8986	1.3454	4.6539	1.8563	0.0277	0.8119
U2Fusion	6.7412	0.8967	1.3425	5.6638	2.7425	0.033	0.8553

续表

GANMcC	7.3532	0.7113	1.462	4.5523	2.7927	0.0362	0.7844
STDFusion	7.3343	0.6052	1.3731	5.6416	1.8832	0.0469	0.7764
ITFuse	6.252	0.865	1.2613	2.3287	2.6548	0.0438	0.8661
BTSFusion	6.9397	0.8361	1.4205	5.4403	2.2326	0.0508	0.8397
RFN-nest	7.3514	0.8928	1.4319	4.991	**4.2928**	0.0323	0.7832
Ours	**7.95**	**0.961**	**1.5205**	**6.1717**	3.3774	**0.0518**	**0.8914**

图 6 为消融结果，依次为红外图像、可见光图像、融合结果、移除多尺度特征提取模块(MFEM)的融合结果，以及移除跨模态融合模块(CMDFM)的融合结果。

Figure 6. Visual comparison of ablation study results
图 6. 消融实验结果视觉对比

为验证多尺度特征提取机制的有效性，通过移除该模块进一步对比分析该模块在特征提取方面的性能差异。如图 6(d)所示，当移除该模块，仅采用单尺度特征时，融合图像出现了明显的局部纹理模糊(例如建筑物轮廓不清)。当移除融合模块，仅采用加法融合时。其效果如图 6(e)所示，关键区域的对比敏感度下降明显(表现为车尾灯周围出现光晕扩散)。改进后的融合网络生成的融合图像在信息提取方面实现了较好平衡，且复原后细节损失更小。

另外，表 3 是实验融合图像的客观评价指标结果，可以看出无论是多尺度特征提取模块还是融合模块都能明显提升各项指标，这说明融合图像有显著的视觉优势，并且信息整合能力强，实用价值突出。

Table 3. Objective metrics for ablation studies
表 3. 消融实验客观指标

Experiment	EN	SSIM	SCD	AG	MI	SF	VIF
No MFEM	6.1493	0.4076	1.1003	3.9086	1.7531	0.0231	0.7087
No CMDFM	6.1756	0.4041	1.1096	4.2908	1.6877	0.0119	0.7892

4. 结论

本文提出一种基于 Transformer-CNN 的混合模型，该方法主要采用双分支混合架构的编码器，通过 CNN 分支的多尺度 MFEM 模块捕捉局部纹理特征，结合 Transformer 分支的全局注意力机制，实现了局部细节与全局上下文的高效协同。消融实验表明，该混合结构显著优于单一架构模型。结合精心设计的跨模态差异融合模块，通过可学习的 Swish 激活函数动态平衡模态间特征差异，本算法能够在全局与局部特征整合过程中保留更多源图像细节信息，并且实现多模态图像融合任务中强泛化性与高精度融合的平衡。在主观视觉评价与客观量化评估中，本算法均展现出显著优势。

参考文献

[1] Tang, L., Zhang, H., Xu, H. and Ma, J. (2023) Deep Learning-Based Image Fusion: A Survey. *Journal of Image and Graphics*, **28**, 3-36. https://doi.org/10.11834/jig.220422

[2] 常天庆, 张杰, 赵立阳, 等. 基于可见光与红外图像融合的装甲目标检测算法[J]. 兵工学报, 2024, 45(7): 2085-2096.

[3] 黄渝萍, 李伟生. 医学图像融合方法综述[J]. 中国图象图形学报, 2023, 28(1): 118-143.

[4] 张宏钢, 杨海涛, 郑逢杰, 等. 特征级红外与可见光图像融合方法综述[J]. 计算机工程与应用, 2024, 60(18): 17-31.

[5] Liu, J., Fan, X., Jiang, J., Liu, R. and Luo, Z. (2022) Learning a Deep Multi-Scale Feature Ensemble and an Edge-Attention Guidance for Image Fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, **32**, 105-119. https://doi.org/10.1109/tcsvt.2021.3056725

[6] Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X. and Zhang, L. (2020) IFCNN: A General Image Fusion Framework Based on Convolutional Neural Network. *Information Fusion*, **54**, 99-118. https://doi.org/10.1016/j.inffus.2019.07.011

[7] Zhang, H., Le, Z., Shao, Z., Xu, H. and Ma, J. (2021) MFF-GAN: An Unsupervised Generative Adversarial Network with Adaptive and Gradient Joint Constraints for Multi-Focus Image Fusion. *Information Fusion*, **66**, 40-53. https://doi.org/10.1016/j.inffus.2020.08.022

[8] Zheng, Q., Zhao, Y., Zhang, X., Zhu, P. and Ma, W. (2022) A Multi-View Image Fusion Algorithm for Industrial Weld. *IET Image Processing*, **17**, 193-203. https://doi.org/10.1049/ipr2.12627

[9] Zhou, W., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P. (2004) Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, **13**, 600-612. https://doi.org/10.1109/tip.2003.819861

[10] Tang, L., Yuan, J., Zhang, H., Jiang, X. and Ma, J. (2022) PIAFusion: A Progressive Infrared and Visible Image Fusion Network Based on Illumination Aware. *Information Fusion*, **83**, 79-92. https://doi.org/10.1016/j.inffus.2022.03.007

[11] Xu, H. (2020) RoadScene Database. https://github.com/hanna-xu/RoadScene

[12] Xu, H., Ma, J., Jiang, J., Guo, X. and Ling, H. (2022) U2Fusion: A Unified Unsupervised Image Fusion Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 502-518. https://doi.org/10.1109/tpami.2020.3012548

[13] Ma, J., Zhang, H., Shao, Z., Liang, P. and Xu, H. (2021) GANMcC: A Generative Adversarial Network with Multiclassification Constraints for Infrared and Visible Image Fusion. *IEEE Transactions on Instrumentation and Measurement*, **70**, 1-14. https://doi.org/10.1109/tim.2020.3038013

[14] Ma, J., Tang, L., Xu, M., Zhang, H. and Xiao, G. (2021) STDFusionNet: An Infrared and Visible Image Fusion Network Based on Salient Target Detection. *IEEE Transactions on Instrumentation and Measurement*, **70**, 1-13. https://doi.org/10.1109/tim.2021.3075747

[15] Tang, W., He, F. and Liu, Y. (2024) ITFuse: An Interactive Transformer for Infrared and Visible Image Fusion. *Pattern Recognition*, **156**, Article ID: 110822. https://doi.org/10.1016/j.patcog.2024.110822

[16] Liu, K., Li, M., Chen, C., Rao, C., Zuo, E., Wang, Y., *et al*. (2024) Dsfusion: Infrared and Visible Image Fusion Method Combining Detail and Scene Information. *Pattern Recognition*, **154**, Article ID: 110633. https://doi.org/10.1016/j.patcog.2024.110633

[17] Qian, Y., Liu, G., Tang, H., Xing, M. and Chang, R. (2024) BTSFusion: Fusion of Infrared and Visible Image via a Mechanism of Balancing Texture and Salience. *Optics and Lasers in Engineering*, **173**, Article ID: 107925. https://doi.org/10.1016/j.optlaseng.2023.107925

[18] Li, H., Wu, X. and Kittler, J. (2021) RFN-Nest: An End-to-End Residual Fusion Network for Infrared and Visible Images. *Information Fusion*, **73**, 72-86. https://doi.org/10.1016/j.inffus.2021.02.023