中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2025)07-2499-15

论文引用格式:Yang T Y, Huo H T, Liu X W and Zheng B W. 2025. Information interaction linear Transformer network for infrared and visible image fusion. Journal of Image and Graphics, 30(7):2499-2513(杨天宇, 霍宏涛, 刘晓文, 郑博文. 2025. 用于红外与可见光图像融合的信息交互线性Transformer网络. 中国图象图形学报, 30(7):2499-2513)[DOI:10.11834/jig.240569]

用于红外与可见光图像融合的信息交互线性Transformer网络

杨天宇,霍宏涛*,刘晓文,郑博文

中国人民公安大学信息网络安全学院,北京　100038

摘　要:**目的**　基于Transformer的深度学习技术在红外与可见光图像融合任务中表现出优异性能。然而基于Transformer的图像融合网络架构计算效率较低,训练过程中需要消耗大量计算资源和储存空间。此外,现有大多数融合方法忽略了编码阶段的跨模态信息交互。**方法**　针对这两个问题,提出一种新型的红外与可见光图像融合算法,采用基于信息交互线性Transformer的网络结构,包括双分支编码器和解码器。双分支编码器首先利用卷积层提取源图像浅层特征,随后通过全局信息快速交互模块中信息交互快速Transformer实现跨模态信息的深度整合,解码器利用多层级联卷积模块得到高质量重建图像。此外,设计傅里叶变换损失引导模型训练,以保留源图像重要频域信息。**结果**　对比实验在MSRS(multi-spectral road scenarios)和TNO(the Netherlands organization)数据集上与13种传统以及深度学习融合方法进行比较。主观评价方面,本文方法在复杂场景下融合效果优势明显,能够产生人物目标轮廓清晰、高视觉质量的融合结果;客观指标方面,在MSRS数据集上本文算法在信息熵、视觉保真度、平均梯度、边缘强度和基于梯度的相似性度量5项指标上取得最优值,相比于对比方法最优值分别提升0.75%、0.15%、1.56%、1.52%和1.27%。在TNO数据集上本文算法在的信息熵、视觉保真度、平均梯度和基于梯度相似性度量4项指标上取得最优值,相比于对比方法最优值分别提升1.53%、3.79%、0.17%和6.94%。消融实验验证了本文融合网络各组件的有效性,并对本文方法计算复杂度进行分析。**结论**　本文提出基于信息交互线性Transformer的融合网络,在提升视觉效果、保留复杂场景下纹理细节信息以及提高计算效率等方面均具有优越性。

关键词:图像融合;信息交互;线性Transformer;傅里叶变换损失;深度学习

Information interaction linear Transformer network for infrared and visible image fusion

Yang Tianyu, Huo Hongtao*, Liu Xiaowen, Zheng Bowen

Department of Information and Cyber Security, People's Public Security University of China, Beijing　100038, China

Abstract: **Objective**　Transformer-based deep learning techniques have demonstrated excellent performance in infrared and visible image fusion. However, Transformer-based image fusion networks suffer from low computational efficiency and require substantial computational resources and storage space during training. In addition, most existing fusion methods overlook cross-modal information interaction during the encoding phase, leading to the loss of complementary information,

收稿日期:2024-09-27;**修回日期**:2024-12-09;**预印本日期**:2024-12-16

*通信作者:霍宏涛　huohongtao@ppsuc.edu.cn

基金项目:国家重点研发计划项目

Supported by:National Key R&D Program of China

particularly in complex scenarios, such as low-light conditions or smoke occlusion, where highlighting target features is challenging. To address these challenges, we propose a fusion network based on an information interaction linear Transformer (I2F Transformer). The proposed I2F Transformer reduces computational costs while enhancing cross-modal information integration, producing fused images with a clear background and prominent targets. **Method** In this study, we propose a novel infrared and visible image fusion algorithm that employs a network structure based on an information interaction linear Transformer, including a dual-branch encoder and decoder. First, the encoder adopts a dual-branch structure to extract features separately from infrared and visible light images. The encoder extracts shallow features from the source images by using a 3 × 3 convolutional layer with a stride of 1. Then, these shallow features are input into three cascaded global information fast interaction modules, where complementary features from infrared and visible feature maps are further extracted and integrated through the I2F Transformer to achieve deep integration of cross-modal features. The decoder consists of five cascaded convolutional modules, with each containing a convolutional layer and an activation function layer, ultimately reconstructing the fused features into a high-quality fused image. To enhance the model's training effectiveness, a Fourier transform loss function is designed to preserve critical frequency domain information from the source images, ensuring the detail and clarity of the fused images. **Result** Comparative experiments were conducted on the Multispectral Road Scenarios (MSRS) and TNO datasets, comparing 13 traditional and deep learning-based fusion methods. In the subjective evaluation, the proposed method demonstrated a clear advantage in complex scenarios. With regard to the objective evaluation, our method achieved optimal values in 5 objective metrics on the MSRS dataset: entropy (EN), visual information fidelity (VIF), average gradient (AG), edge intensity (EI), and gradient-based similarity measure ($Q^{AB/F}$). Compared with the best values in the 5 metrics of 13 existing fusion algorithms, an average increase of 0.75%, 0.15%, 1.56%, 1.52%, and 1.27% are reached. Our method achieved the best values in EN, VIF, AG, and $Q^{AB/F}$ on the TNO dataset. Compared with the best values in the 4 aforementioned metrics of 13 existing fusion algorithms, an average increase of 1.53%, 3.79%, 0.17%, and 6.94% are reached. In addition, we validated the effectiveness of each component in the network through ablation experiments. To further validate the computational efficiency of the fusion model, we analyzed the algorithm's computational complexity. **Conclusion** In this study, we proposed a fusion network based on an information interaction linear Transformer, innovatively introduced a cross-modal information interaction mechanism, and achieved network lightweighting through linear modules in the I2F Transformer. The experimental results show that compared with 13 existing fusion methods, the proposed method demonstrates superiority in enhancing visual effects, preserving texture details in complex scenes, and improving computational efficiency.

Key words: image fusion; information interaction; linear Transformer; Fourier transform loss; deep learning

0 引 言

由于成像原理限制,单一的红外或可见光图像都难以全面描述完整的场景信息。可见光图像具有高空间分辨率,在良好的照明条件下能够清晰地展现场景的细节,但是可见光成像易受照明条件和恶劣天气影响。相比之下,红外图像对于场景温度变化的敏感性高,易于在夜间或恶劣天气条件下捕获场景中的隐蔽目标信息。然而,红外图像空间分辨率较低,缺乏纹理和背景信息。红外与可见光图像融合作为计算机视觉领域中一项重要技术,旨在分别提取红外与可见光图像中的互补特征信息并去除冗余信息,生成单幅目标显著、背景细节信息丰富、能够对场景进行全面表征的融合图像。目前,红外与可见光图像融合技术已在军事侦察(Paramanandham 和 Rajendiran, 2018)、安防监控(李国梁 等, 2022)和遥感监测(Amarsaikhan 等, 2012)等领域得到广泛应用。

现阶段红外与可见光图像融合算法可分为传统方法和基于深度学习的方法。传统融合方法通常采用相关数学方法或图像变换提取特征,并使用手工设计的融合方法将不同图像的特征进行融合。传统方法包括基于多尺度变换的方法(陈木生, 2016; Wu 等, 2020)、基于子空间的方法(Harsanyi 和 Chang, 1994)、基于稀疏表示的方法(Lu 等, 2014; 杨培 等, 2021)以及基于显著性的方法(Zhang 等, 2015)等。然而,传统融合方法对于不同的图像变换需要设计

特定的特征提取算法。由于人工设计的特征融合算法深度特征提取能力不足，在复杂场景中传统方法往往难以获得目标显著、细节清晰的融合图像。

随着人工神经网络的快速发展，图像融合领域出现了许多基于深度学习的方法。根据所采用网络架构的区别，基于深度学习的图像融合方法可以分为基于自编码器（auto-encoder，AE）的方法、基于生成对抗网络（generative adversarial network，GAN）的方法和基于Transformer的方法（唐霖峰 等，2023）。基于自编码器的方法通过编码器提取源图像层次化特征表示，并通过解码器对融合后的特征图进行图像重建。为了提取源图像中深度特征，Li和Wu（2019）在编码网络中引入密集连接。随后，Li等人（2021）提出一种端到端的残差融合网络，并采用新颖的两阶段训练策略，实现了更好的融合性能。除此之外，Zhang等人（2020）提出基于卷积神经网络（convolutional neural network，CNN）的通用图像融合框架，并首次在图像融合模型中引入感知损失。基于生成对抗网络的方法通过不断对抗训练优化生成器和判别器，以生成高质量融合图像。Ma等人（2019）首次采用生成对抗网络处理图像融合任务，提出端到端的生成式红外与可见光图像融合对抗网络。针对单判别器网络架构中存在的模态失衡的问题，Ma等人（2020）又提出可用于多分辨率图像融合的双判别器生成对抗网络。

Transformer最初由Vaswani等人（2017）针对自然语言处理（natural language processing，NLP）中的序列建模任务提出。Dosovitskiy等人（2021）将Transformer引入计算机视觉领域以来，出现了大量基于Transformer图像融合方法。视觉Transformer通过自注意力和位置编码的方式，建立图像长距离依赖关系，实现了对图像全局信息的有效感知，从而克服了卷积神经网络在全局特征建模方面的局限性。视觉Transformer模型将输入的图像划分为一系列16 × 16像素的图像块，然后将这些图像块展平并嵌入到高维特征向量中。Transformer编码器由多层自注意力机制和前馈神经网络组成。自注意力机制通过计算输入序列中每个块之间的相关性，生成新的表示。自注意力机制可表示为

$$f_{\text{Attention}}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=f_{\text{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\boldsymbol{V}\right) \tag{1}$$

式中，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$分别表示查询、键和值的向量，d_k表示键向量维度。前馈神经网络通过非线性变换进一步处理注意力机制的输出。

近年来，Wang等人（2022）提出基于Swin Transformer的红外与可见光图像融合算法，构建了一个完全基于注意力机制的特征编码骨干网络。Tang等人（2023）提出基于Y形动态Transformer的融合网络，通过两条分支分别提取红外和可见光图像信息，构建全局信息依赖关系。Zhao等人（2023）在红外与可见光图像融合中引入特征解耦思想，将源图像信息解耦为共有信息和特有信息，融合网络结合了卷积神经网络和Transformer的优点。

上述方法在图像融合任务中均取得了较好的融合结果，但依然存在一些易被忽略的问题。首先，Transformer模型通过自注意力机制捕捉全局依赖关系，由于自注意力的计算复杂度与输入序列呈二次方关系，导致基于Transformer的网络架构在训练过程中需要消耗大量计算资源和储存空间。其次，多模态图像融合任务中，不同模态之间的互补信息对于融合图像的视觉效果至关重要。现有方法（Qu等，2022）缺少跨模态交互机制，导致部分互补信息丢失，尤其是在复杂场景下（如光照不足、烟雾阻挡等）难以突出目标特征。此外，大多数现有方法在表征图像频域信息方面仍存在不足。

针对上述问题，本文提出一种基于信息交互线性Transformer的融合网络，将卷积神经网络和Transformer相结合，提取局部特征的同时建立全局信息依赖关系。为了解决资源消耗和计算效率问题，如图1所示，本文通过适当的近似映射函数（Han等，2023）来解耦视觉Transformer中的softmax函数，改变注意力机制中矩阵运算顺序，从而将网络架构中Transformer计算降为线性。Oppenheim和Lim（1981）研究表明图像频域变换后局部相位则能够表征目标的形状和边缘细节，因此本文在融合网络训练过程中，引入傅里叶变换相关损失函数，以保留重要频域信息。

本文主要贡献如下：1）在融合网络架构中引入线性Transformer，节省整体计算成本的同时关注图像整体信息，与视觉Transformer相比，融合效果相当甚至更优；2）为了解决现有大多数融合算法忽略跨模态互补信息的问题，提出信息交互快速Transformer（information interaction fast Transformer，I2F

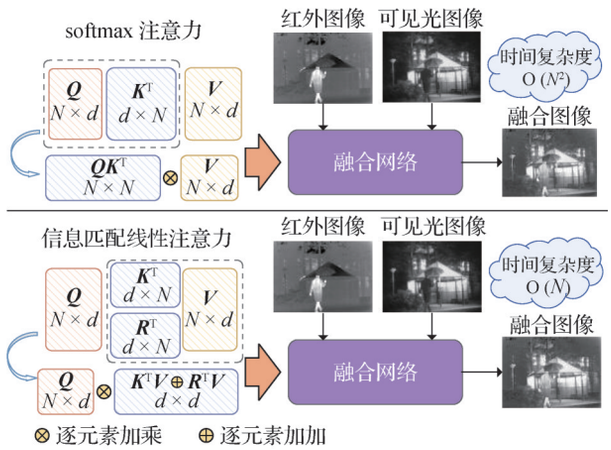


图1　本文方法信息匹配线性注意力与softmax注意力的计算方式对比

Fig. 1　Comparison of information matching linear attention of proposed network and softmax attention

Transformer)，即使在复杂场景下也能够产生背景清晰、目标显著的融合结果；3)网络优化过程中，引入快速傅里叶变换损失函数，显著提高了融合图像边缘和纹理清晰度；4)与现有融合方法对比实验表明，本文所提出的融合算法在视觉效果和定量指标上都具有优越性。

1　融合方法

1.1　网络结构

本文提出一种基于信息交互线性Transformer的红外与可见光图像融合算法，算法结构如图2所示，整体网络结构包含编码器和解码器。由于红外图像与可见光图像的模态信息差异较大，编码器采用双分支结构独立提取源图像特征。

如图2编码器结构所示，输入图像分别通过一个步长为1的3×3卷积层，以提取图像浅层特征。随后将源图像的浅层特征输入3个级联的全局信息快速交互模块(global information fast interaction module, GIFM)，通过信息交互机制，进一步提取深度互补特征，有效整合红外与可见光图像的模态信息。

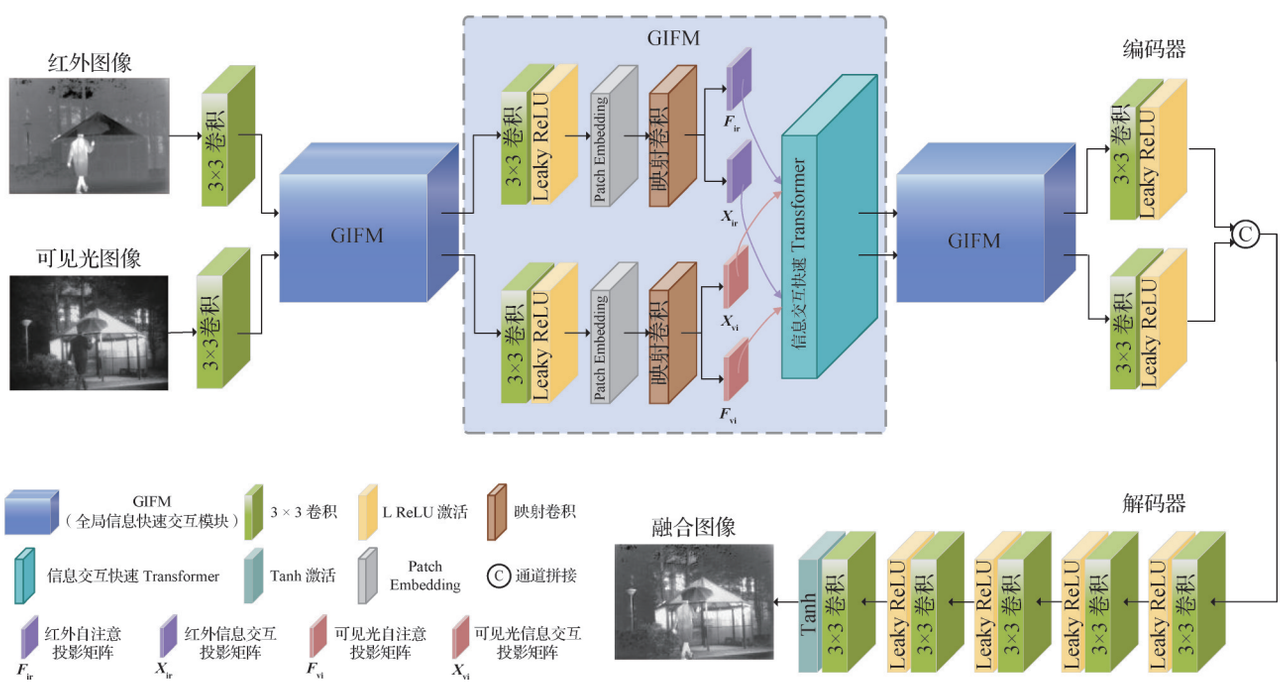


图2　融合网络结构

Fig. 2　Structure of fusion network

首先，特征图输入GIFM后分别通过卷积层和Leaky ReLU(leaky rectified linear unit)层进一步提取局部特征。局部特征通过Patch Embedding层将特征图划分为一系列非重叠的小块，通常称为patches，每个patch都包含了局部的特征信息。映射卷积层对来自Patch Embedding层的特征序列进行处理，分别生成红外自注意投影矩阵($\boldsymbol{F}_{\mathrm{ir}}$)、红外信息交互投影矩阵($\boldsymbol{X}_{\mathrm{ir}}$)、可见光自注意投影矩阵($\boldsymbol{F}_{\mathrm{vi}}$)以及可见光信息交互投影矩阵($\boldsymbol{X}_{\mathrm{vi}}$)。自注意投影矩阵($\boldsymbol{F}$)主要用于对序列中所有patches的全局依赖关系

的建模，从而实现对局部特征的整合。信息交互投影矩阵（$\boldsymbol{X}$）则专注于捕捉不同模态特征之间的相互关系，即红外图像和可见光图像之间的互补信息。随后，将上述投影矩阵交叉分组作为 I2F Transformer 的输入，I2F Transformer 的结构将在 2.2 节详细介绍。通过这种交叉分组策略，实现 Transformer 模块中跨模态特征的深度整合。

将编码器输出特征在通道维度上拼接后送入解码器。解码器部分由 5 个级联的卷积模块构成，每个卷积模块包括一个卷积层和一个激活函数层，旨在将融合特征重建为高质量的融合图像。为了解决 ReLU 激活函数训练过程中神经元死亡问题，除解码器最后一层外，本文网络架构中所有卷积层均使用 LeakyReLU 激活函数。解码器最后一个卷积模块使用 Tanh 激活函数。

本文提出的网络中，所有卷积层中卷积核尺寸均为 3 × 3，且步长和填充均设置为 1。此外，为避免编码过程中图像信息丢失，融合网络中不引入任何下采样操作，深度特征和融合图像的大小与源图像保持一致。

1.2 I2F Transformer

由于 Transformer 的全局信息建模优势，基于 Transformer 的融合方法被提出。然而，现有的基于 Transformer 的融合方法大多忽略了跨模态信息的引入。为此，本文提出 I2F Transformer，通过引入跨模态域外信息，对互补信息进行自适应编码。

如图 3 所示，I2F Transformer 整体为双通路结构，其中可见光自注意投影矩阵（$\boldsymbol{F}_{\mathrm{vi}}$）与红外信息交互投影矩阵（$\boldsymbol{X}_{\mathrm{ir}}$）配对，作为可见光特征提取部分的输入；红外自注意投影矩阵（$\boldsymbol{F}_{\mathrm{ir}}$）与可见光信息交互投影矩阵（$\boldsymbol{X}_{\mathrm{vi}}$）配对，作为红外特征提取部分的输入。本文利用信息匹配线性聚焦注意力（information matches linear focused attention，MILF attention）分别提取不同模态间的互补信息，计算过程为

$$\boldsymbol{Z}_{\mathrm{vi}} = MIFL\,Att\left(\boldsymbol{F}_{\mathrm{vi}}, \boldsymbol{X}_{\mathrm{ir}}\right) \tag{2}$$

$$\boldsymbol{Z}_{\mathrm{ir}} = MIFL\,Att\left(\boldsymbol{F}_{\mathrm{ir}}, \boldsymbol{X}_{\mathrm{vi}}\right) \tag{3}$$

式中，$MIFL\,Att(\cdot)$ 表示信息匹配线性聚焦注意力计算，$\boldsymbol{Z}_{\mathrm{vi}}$ 和 $\boldsymbol{Z}_{\mathrm{ir}}$ 分别表示可见光和红外通路信息匹配线性聚焦注意力的输出。

输入的投影矩阵在通过注意力层后，继续经过一个前馈网络进行进线性变换，并进行归一化处理

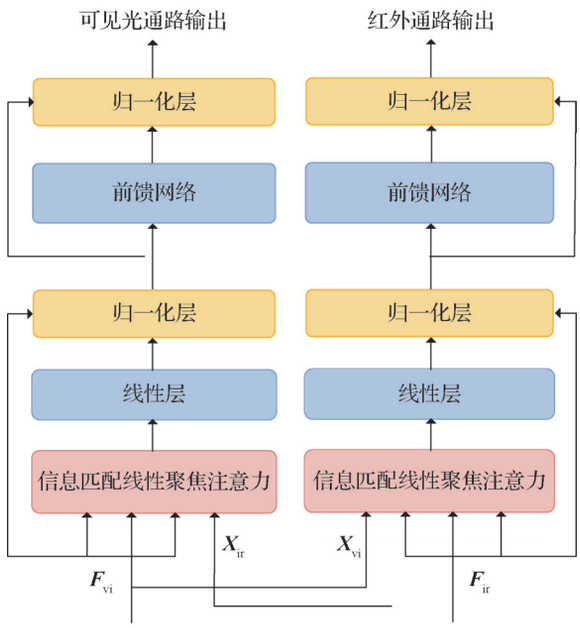


图 3　I2F Transformer 结构

Fig. 3　Structure of I2F Transformer

和残差连接，整个 Transformer 模块内的计算过程表达为

$$\boldsymbol{Z}'_{\mathrm{vi}} = LN\left(f_{\mathrm{Linear}}\left(\boldsymbol{Z}_{\mathrm{vi}}\right)\right) + \boldsymbol{F}_{\mathrm{vi}} \tag{4}$$

$$\boldsymbol{O}_{\mathrm{vi}} = LN\left(FFN\left(\boldsymbol{Z}'_{\mathrm{vi}}\right)\right) + \boldsymbol{Z}'_{\mathrm{vi}} \tag{5}$$

$$\boldsymbol{Z}'_{\mathrm{ir}} = LN\left(f_{\mathrm{Linear}}\left(\boldsymbol{Z}_{\mathrm{ir}}\right)\right) + \boldsymbol{F}_{\mathrm{ir}} \tag{6}$$

$$\boldsymbol{O}_{\mathrm{ir}} = LN\left(FFN\left(\boldsymbol{Z}'_{\mathrm{ir}}\right)\right) + \boldsymbol{Z}'_{\mathrm{ir}} \tag{7}$$

式中，LN 表示 layer norm 操作，f_{Linear} 表示线性层，FFN 表示前馈网络，由多个线性层和激活函数组成，$\boldsymbol{O}_{\mathrm{vi}}$ 和 $\boldsymbol{O}_{\mathrm{ir}}$ 分别表示 I2F Transformer 模块可见光和红外通路的输出。

1.3 信息匹配线性聚焦注意力

本文在 I2F Transformer 模块中设计了全新的信息匹配线性聚焦注意力，通过信息交互投影矩阵（$\boldsymbol{X}$）引入跨模态信息，其计算过程为

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{F} \times \boldsymbol{W}^{Q} \\ \boldsymbol{K} = \boldsymbol{F} \times \boldsymbol{W}^{K} \\ \boldsymbol{V} = \boldsymbol{F} \times \boldsymbol{W}^{V} \\ \boldsymbol{R} = \boldsymbol{X} \times \boldsymbol{W}^{R} \end{cases} \tag{8}$$

$$Att\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \boldsymbol{R}\right) = f_{\mathrm{softmax}}\left(\frac{\left(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}} + \boldsymbol{Q}\boldsymbol{R}^{\mathrm{T}}\right)}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{9}$$

式中，$\boldsymbol{F}$ 和 $\boldsymbol{X}$ 分别表示自注意投影矩阵和信息交互投影矩阵，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别表示查询向量、键向量和值向量，$\boldsymbol{R}$ 表示跨模态信息向量，其作用与键向量相似，d_k 表示键向量维度。

本文通过引入聚焦函数f(Han等，2023)解耦Transformer中softmax函数，改变注意力中矩阵运算顺序，将I2F Transformer时间复杂度降为线性。聚焦函数通过调整查询和键特征的方向，使相似的查询键对更接近，解耦过程为

$$f_{\text{softmax}}\left(\boldsymbol{QK}^{\mathrm{T}}\right)=\phi\left(\boldsymbol{Q}\right)\phi\left(\boldsymbol{K}\right)^{\mathrm{T}} \tag{10}$$

$$\phi(x)=f\left(f_{\text{ReLU}}(x)\right),\ f(x)=\frac{\|x\|}{\left\|x^{p}\right\|}x^{p} \tag{11}$$

式中，$f(\cdot)$表示聚焦函数，x^p表示x的p次幂，ReLU函数确保输入的非负性。聚焦函数只调整查询和键的特征方向，而映射后特征范数保持不变，即$\|x\|=\|f(x)\|$。

因此，引入聚焦函数后，I2F Transformer中注意力机制的计算过程可表示为

$$Att\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V},\boldsymbol{R}\right)=\left(\phi\left(\boldsymbol{Q}\right)\phi\left(\boldsymbol{K}\right)^{\mathrm{T}}+\phi\left(\boldsymbol{Q}\right)\phi\left(\boldsymbol{R}\right)^{\mathrm{T}}\right)\boldsymbol{V} \tag{12}$$

如图4所示，根据矩阵乘法相关性质，将计算顺序从$\left(\boldsymbol{QK}^{\mathrm{T}}\right)\boldsymbol{V}$改为$\boldsymbol{Q}\left(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{V}\right)$，使得信息匹配线性聚焦注意力的计算复杂度降至$O(n)$，计算过程可表示为

$$MIFL\,Att\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V},\boldsymbol{R}\right)=\phi\left(\boldsymbol{Q}\right)\left(\phi\left(\boldsymbol{K}\right)^{\mathrm{T}}\boldsymbol{V}+\phi\left(\boldsymbol{R}\right)^{\mathrm{T}}\boldsymbol{V}\right) \tag{13}$$

I2F Transformer通过引入信息匹配线性聚焦注意力，在保持线性计算复杂度的同时，引入跨模态互补信息。即使在复杂场景下也能够产生目标显著、背景清晰的融合结果。

1.4 损失函数

红外图像和可见光图像的模态信息由于成像原理不同存在较大差异。因此，在设计损失函数时应综合考虑源图像固有特征。本文所提出的融合网络通过结构相似度损失(L_{SSIM})、纹理损失(L_{texture})和傅里叶变换损失(L_{fft})进行监督训练。

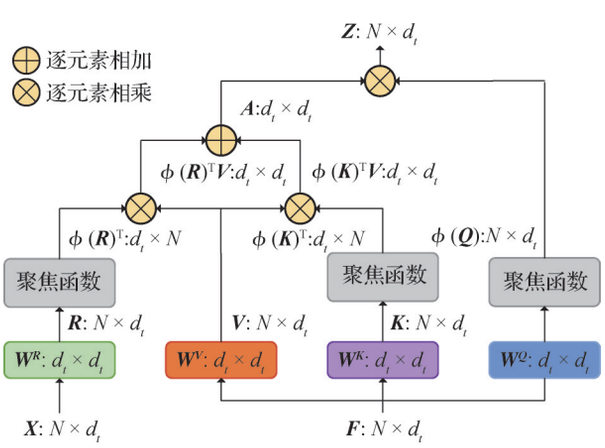


图4 信息匹配线性聚焦注意力机制计算图

Fig. 4 The computation graph of MILF attention mechanism

在图像融合任务中，源图像纹理细节的保留尤为重要。由于融合图像的最优纹理可以表示为红外和可见光图像纹理的最大集合，引入纹理损失L_{texture}来强制融合图像包含更多的纹理信息，L_{texture}计算为

$$L_{\text{texture}}=\frac{1}{HW}\left\|\left|\nabla\boldsymbol{I}_{\text{f}}\right|-\max\left(\left|\nabla\boldsymbol{I}_{\text{ir}}\right|,\left|\nabla\boldsymbol{I}_{\text{vi}}\right|\right)\right\|_{1} \tag{14}$$

式中，$\boldsymbol{I}_{\text{f}}$、$\boldsymbol{I}_{\text{ir}}$和$\boldsymbol{I}_{\text{vi}}$分别表示融合图像、红外图像和可见光图像，$H$、$W$分别表示图像的长和宽，$\|\cdot\|_1$表示L1范数，$\nabla$表示Sobel梯度算子，$\max(\cdot)$表示逐元素取最大值。

结构相似度损失函数L_{SSIM}分别通过亮度、对比度和结构信息来衡量融合图像和源图像之间的相似性，L_{SSIM}计算为

$$L_{\text{SSIM}}=\frac{1}{2}\left(1-S\left(\boldsymbol{I}_{\text{f}},\boldsymbol{I}_{\text{ir}}\right)\right)+\frac{1}{2}\left(1-S\left(\boldsymbol{I}_{\text{f}},\boldsymbol{I}_{\text{vi}}\right)\right) \tag{15}$$

$$S\left(\boldsymbol{x},\boldsymbol{y}\right)=\frac{\left(2\mu_{x}\mu_{y}+c_{1}\right)\left(2\delta_{xy}+c_{2}\right)}{\left(\mu_{x}^{2}+\mu_{y}^{2}+c_{1}\right)\left(\delta_{x}^{2}+\delta_{y}^{2}+c_{2}\right)} \tag{16}$$

式中，$S\left(\boldsymbol{x},\boldsymbol{y}\right)$表示$\boldsymbol{x}$和$\boldsymbol{y}$两幅图像之间的SSIM值。$\mu_x$和$\mu_y$分别表示$\boldsymbol{x}$和$\boldsymbol{y}$的均值，$\delta_x$和$\delta_y$分别表示$\boldsymbol{x}$和$\boldsymbol{y}$的方差，$\delta_{xy}$表示$\boldsymbol{x}$和$\boldsymbol{y}$的协方差。

Jiang等人(2021)研究发现，图像生成模型通过缩小频域的间隙可以有效改善图像的重建和合成质量。由于现有融合方法通常使用空间域损失函数训练网络，而很少考虑频域优化，本文在网络优化过程中引入傅里叶变换损失L_{fft}，以保留源图像中重要频域信息，L_{fft}计算为

$$L_{\text{fft}}=\frac{1}{HW}\left\|\mathcal{F}\left(\boldsymbol{I}_{f}\right)-\max\left(\mathcal{F}\left(\boldsymbol{I}_{ir}\right),\mathcal{F}\left(\boldsymbol{I}_{vi}\right)\right)\right\|_{1} \tag{17}$$

式中，$\mathcal{F}(\cdot)$表示快速傅里叶变换，$\|\cdot\|_1$表示L1范数。

综上，本文融合网络训练过程中总损失函数为$L_{\text{total}}=\lambda_1 L_{\text{texture}}+\lambda_2 L_{\text{SSIM}}+\lambda_3 L_{\text{fft}}$，具体超参数设置在实验部分给出。

2 实 验

2.1 实验设置

为验证所提出融合算法的有效性和先进性，本文采用MSRS(multi-spectral road scenarios)数据集对

融合网络进行训练。MSRS数据集为多光谱道路场景数据集，包含715对白天图像对和729对夜晚图像对，其空间分辨率为480 × 640像素。在训练阶段，在MSRS数据集中选取376对白天图像对和376对夜晚图像对，并将选取的图像对裁剪为64 × 64像素图像块进行训练，即26 320对白天图像块和26 320对夜晚图像块。在测试阶段，采用MSRS数据集中测试集进行网络模型测试，其中包含361对红外与可见光图像。同时，为验证融合网络泛化性能，选取TNO（The Netherlands Organization）数据集（Toet，2017）中的20对图像对作为TNO测试集进行测试，该数据集图像来源广泛，均为真实场景图像。

模型的训练和测试阶段均在NVIDIA GeForce RTX 3090 GPU和Intel Xeon Silver 4214R CPU上完成，使用的深度学习框架为PyTorch。训练过程中，所有图像均转为灰度图像。采用Adam优化器对模型参数进行更新，学习率初始值设置为1×10^{-4}，模型训练周期为60个epoch，batch_size设置为16，超参数λ_1、λ_2、λ_3分别设置为40、13、7。

为验证所提融合方法的优越性，本文与13种现有红外与可见光融合算法进行对比，包括四阶偏微分方程（fourth-order partial differential equation，FPDE）（高雪琴 等，2020）、多分辨率奇异值分解（multi-resolution singular value decomposition，MSVD）（Naidu，2011）以及11种深度学习的方法：SeAFusion（semantic-aware real-time infrared and visible image fusion）（Tang 等，2022）、SegMiF（multi-interactive feature learning architecture for image fusion and segmentation）（Liu 等，2023）、CrossFuse（cross attention mechanism based infrared and visible image fusion）（Li 和 Wu，2024）、TarDAL（target-aware dual adversarial learning ）（Liu 等，2022）、RFN-Nest（residual fusion network）（Li 等，2021）、TUFusion（transformer-based universal fusion）（Zhao 等，2024）、SwinFusion（general image fusion via swin transformer）（Ma 等，2022）、SEDRFuse（symmetric encoder-decoder with residual block network for infrared and visible image fusion）（Jian 等，2021）、GANMcC（generative adversarial network with multiclassification constraints）（Ma 等，2021）、U2Fusion（unified unsupervised image fusion network）（Xu 等，2022）和SwinFuse（swin transformer fusion network）（Wang 等，2022）。

对于融合图像的客观评价，本文选取了6种广泛使用的评价指标进行定量分析，分别是信息熵（entropy，EN）（Roberts 等，2008）、标准差（standard deviation，SD）（Rao，1997）、视觉保真度（visual information fidelity，VIF）（Han 等，2013）、平均梯度（average gradient，AG）（Cui 等，2015）、边缘强度（edge intensity，EI）（Xydeas 和 Petrović，2000）和基于梯度的相似性度量（gradient-based similarity measurement，$Q^{AB/F}$）（Xydeas 和 Petrović，2000）。

EN衡量图像中信息的丰富程度，较高的信息熵通常表示融合图像内容的信息量较大。SD反映图像像素值的离散程度，用于衡量图像对比度的变化。VIF用于量化图像与参考图像之间的保真度，能够有效评估图像的视觉质量。AG用于评估图像的锐度和细节清晰度，而EI则反映图像的轮廓和结构特征的强度。$Q^{AB/F}$通过比较图像梯度信息的相似性来反映两幅图像之间的结构相似性。

2.2 对比实验

2.2.1 基于MSRS数据集的融合性能对比

模型在MSRS上的融合结果如图5和图6所示，本文选取了“00537D”和“01352N”两个场景，其中“01352N”为夜晚复杂场景。

如图5所示，U2Fusion、SwinFuse、FPDE和MSVD方法融合结果亮度较低，导致暗处纹理信息大量丢失。CrossFuse方法过度锐化，树木和高楼轮廓处产生伪影，视觉效果较差。TarDAL方法红外信息过于突出，人物目标部分纹理信息丢失。GANMcC、RFN-Nest和TUFusion方法整体对比度偏低，红色矩形区域人物轮廓模糊。SEDRFuse方法的融合结果人物目标轮廓清晰，但整体对比度较差，天空颜色灰暗。SeAFusion和本文方法均取得了目标突出、背景清晰的融合结果，但本文方法在MSRS数据集上定量指标（如表1所示）均优于SeAFusion。综上所述，其他基于Transformer的方法（如TUFusion、SwinFusion和SwinFuse）存在轮廓模糊、部分纹理信息丢失的问题，而本文方法由于引入信息交互机制，融合图像暗处的纹理细节仍清晰可见。

从图6可以看出，在复杂场景（光照不足）中，除SeAFusion、SegMiF、TarDAL、SEDRFuse和本文方法以外，其余方法均难以在黑暗的背景中同时突出红色矩形区域中的自行车和蓝色矩形区域中的空调外

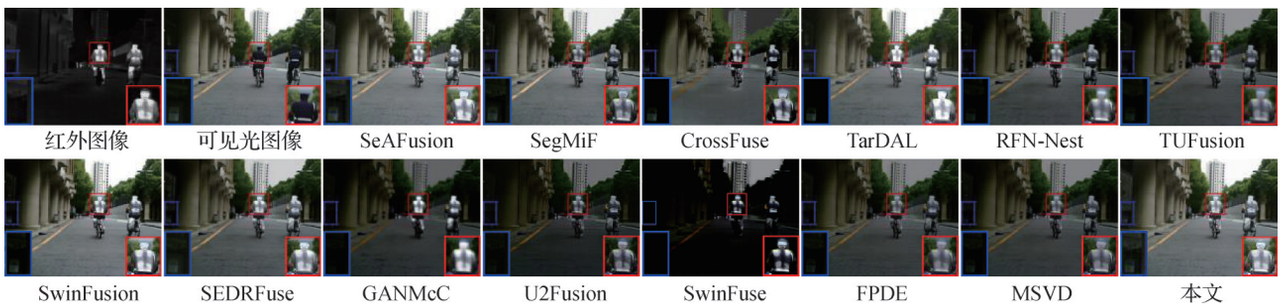


图5　MSRS数据集“00537D”主观评价对比

Fig. 5　Comparison of subjective evaluation of MSRS dataset “00537D”

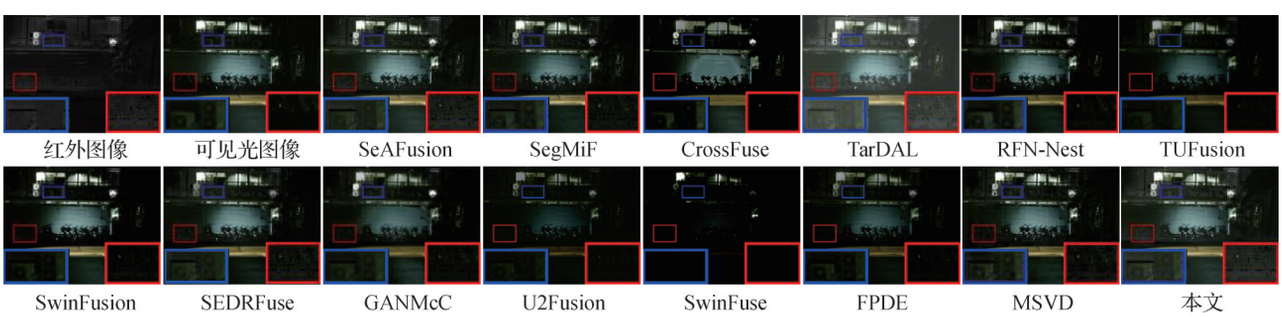


图6　MSRS数据集“01352N”主观评价对比

Fig. 6　Comparison of subjective evaluation of MSRS dataset “01352N”

机。SegMiF和SEDRFuse方法整体亮度不足，红色矩形区域中部分细节纹理丢失。TarDAL方法较好地保留了目标的细节信息，但该方法的融合结果存在对比度低的问题。相比于SeAFusion，本文方法的融合结果在红色矩形区域中细节更加清晰。

相比于其他13种对比算法，本文方法在图5和图6的场景中均取得了最优的融合结果。由于本文方法在融合网络中建立了跨模态信息交互机制，并在网络优化过程中引入快速傅里叶变换损失，即使在复杂场景下，本文方法也能够产生对比度高、目标突出以及背景细节清晰的融合图像。

为了量化对比方法和本文方法的融合结果，保证对比实验的统一性和可靠性，本文选取6项评价指标对13种对比方法和本文方法进行定量比较。在MSRS数据集上的定量比较结果如表1所示，表中评价指标数值为MSRS数据集中361对全部测试数据的平均值。

从表1可以看出，本文算法在EN、VIF、AG、EI和$Q^{AB/F}$ 5项指标上取得最优值，在SD上取得次优值。对比其他方法，本文方法在EI和VIF上具有明显优势。这表明通过本文算法中跨模态信息交互机制所产生的融合图像在保持源图像细节纹理的同时，具有最优的视觉效果。此外，EN上取得的最优值表明本文方法可以在最大程度上保持源图像的信息。AG和$Q^{AB/F}$上取得的最优值表明本文的融合图像边缘信息丰富、纹理清晰。实验从主观分析和客观评价都验证了本文方法相较于其他主流方法的优越性。

2.2.2　基于TNO数据集的融合性能对比

为验证本文提出融合模型的泛化能力，模型在TNO上的融合结果如图7和图8所示，本文分别选取“Kaptein_1654”典型场景和“soldier_behind_smoke”复杂场景（浓烟遮挡）进行定性实验。

从图7可以看出，RFN-Nest、GANMcC和FPDE方法红色矩形区域中人物目标轮廓模糊，部分边缘信息丢失。SegMiF和SEDRFuse方法的融合图像背景亮度不足。CrossFuse、TarDAL和SwinFuse方法能够清晰地突出红色矩形区域中人物目标，但帐篷顶部产生了不同程度的伪影。U2Fusion方法融合结果整体对比度较低，背景偏暗。TUFusion和MSVD方法融合结果背景颜色灰暗，蓝色矩形区域中灌木丛部分纹理丢失。SeAFusion和SwinFusion方法整体对比度较高，目标突出，但与本文方法相比，部分区域（如灌木丛、长椅等）边缘锐化程度欠佳。本文方

法在突出红外图像热辐射信息的同时，更好地保留了源图像的细节纹理信息。

从图 8 可以看出，在复杂场景（浓烟遮挡）下，SeAFusion、TarDAL 和 SwinFuse 方法融合图像红色矩形区域中士兵轮廓较为清晰，但蓝色矩形区域中树干的纹理信息完全丢失。CrossFuse 方法融合结

表 1　不同融合方法在 MSRS 数据集上的定量结果比较

Table 1　Comparison of quantitative results of different fusion methods on MSRS dataset

方法	EN	SD	VIF	AG	EI	$Q^{AB/F}$
SeAFusion	6.525 7	39.600 4	0.689 7	3.503 5	37.478 0	0.671 8
SegMiF	6.000 0	33.764 1	0.602 4	3.076 9	32.614 0	0.565 5
CrossFuse	5.048 0	28.167 7	0.368 8	2.937 3	31.066 1	0.417 3
TarDAL	6.346 2	35.492 6	0.544 4	3.116 0	32.244 7	0.430 7
RFN-Nest	6.195 8	29.077 6	0.516 5	2.115 1	23.094 0	0.395 3
TUFusion	6.085 5	26.406 1	0.401 4	1.719 3	18.760 9	0.218 4
SwinFusion	6.517 9	**43.004 3**	0.723 3	3.541 4	37.850 7	0.671 5
SEDRFuse	6.418 7	35.087 0	0.613 7	3.096 0	32.777 5	0.544 1
GANMcC	6.118 6	26.339 4	0.429 8	1.997 0	21.332 6	0.307 7
U2Fusion	5.042 3	20.342 0	0.333 4	2.263 6	24.130 0	0.344 2
SwinFuse	4.236 4	29.719 6	0.304 6	1.938 3	20.463 7	0.180 6
FPDE	5.991 0	23.986 0	0.393 7	2.688 4	28.048 8	0.491 1
MSVD	5.954 6	23.777 2	0.378 1	2.245 5	22.818 6	0.352 1
本文	**6.574 5**	42.544 6	**0.724 4**	**3.596 7**	**38.425 2**	**0.680 3**

注：加粗、下划线字体分别表示各列最优、次优结果。

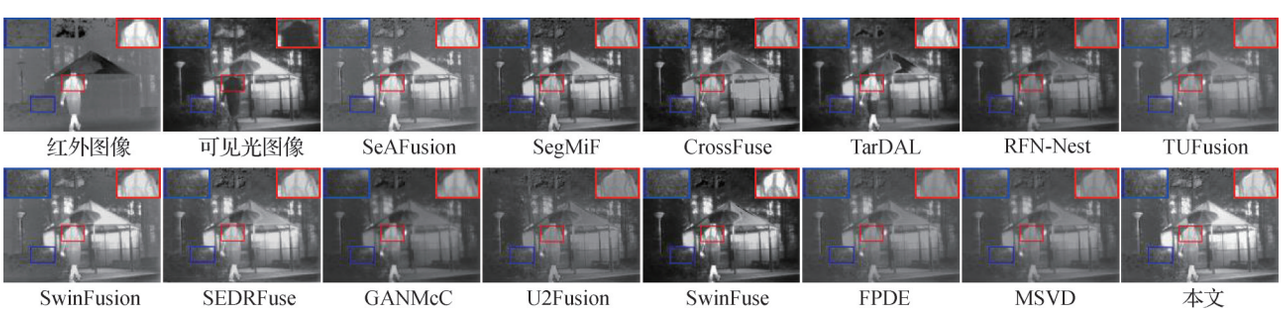


图 7　TNO 数据集“Kaptein_1654”主观评价对比

Fig. 7　Comparison of subjective evaluation of TNO dataset “Kaptein_1654”

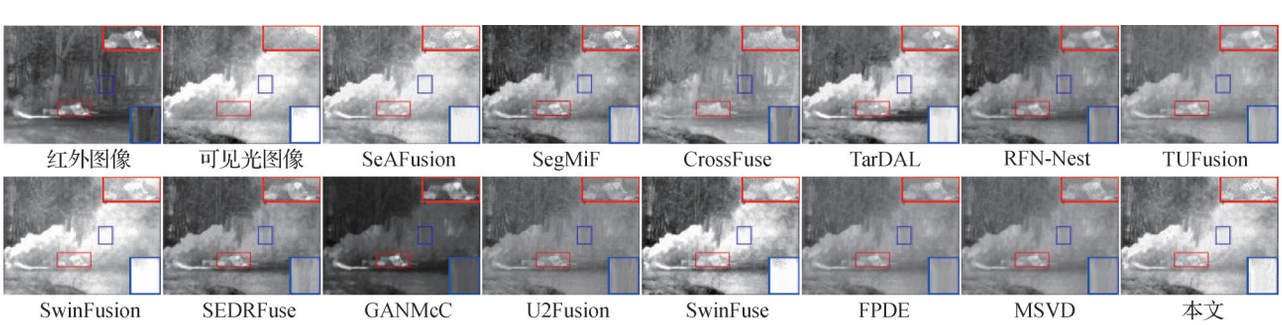


图 8　TNO 数据集“soldier_behind_smoke”主观评价对比

Fig. 8　Comparison of subjective evaluation of TNO dataset “soldier_behind_smoke”

果将红色矩形区域中士兵与浓烟混淆。GANMcC方法融合图像亮度不足，部分背景的纹理细节信息丢失。TUFusion和MSVD方法融合图像背景对比度和边缘锐化方面表现欠佳。相比于其他基于Transformer的方法（如SwinFusion和SwinFuse）均存在红色矩形区域中士兵轮廓模糊、梯度信息丢失的问题，本文方法即使在浓烟遮挡的复杂场景下，仍能够产生人物目标轮廓清晰、高视觉质量的融合图像。

为保证对比实验的一致性和有效性，本文选取相同的6项评价指标，在TNO数据集中所选取的20对图像上进行融合图像质量的客观评价，对比方法和本文方法各项指标的数值平均值如表2所示。可以看出，本文方法在EN、VIF、AG和$Q^{AB/F}$ 4项指标上仍然取得最优值，在SD和EI上取得次优值。由于CrossFuse融合图像中的轮廓噪声，使得SD指标数值偏高。实验结果表明，本文方法能够产生目标显著、轮廓清晰的融合图像，具有良好的泛化能力。

两个数据集上的定性和定量对比实验都验证了本文方法的优越性能和泛化能力。一方面，本文算法模型将卷积神经网络与线性Transformer级联，提高模型计算效率的同时，聚合局部特征和全局特征，使模型轻量化的同时具有优越的融合效果；另一方面，通过将跨模态信息交互机制引入Transformer，充分考虑图像融合过程中的跨模态互补信息。

表2　不同融合方法在TNO数据集上的定量结果比较

Table 2　Comparison of quantitative results of different fusion methods on TNO dataset

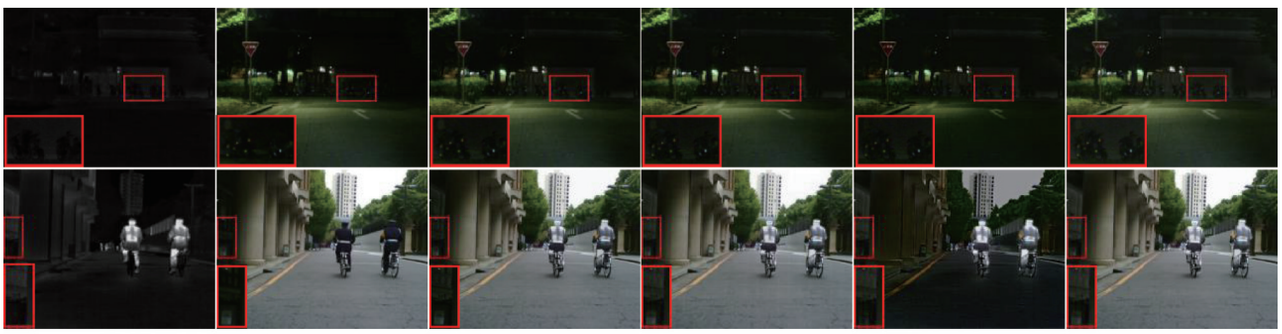
方法	EN	SD	VIF	AG	EI	$Q^{AB/F}$
SeAFusion	6.714 7	41.111 4	0.408 1	<u>4.650 6</u>	**47.065 9**	0.508 5
SegMiF	6.996 2	44.434 6	<u>0.551 7</u>	4.431 3	44.009 7	<u>0.590 8</u>
CrossFuse	<u>7.175 6</u>	**49.756 4**	0.429 8	4.489 3	45.053 2	0.538 5
TarDAL	6.847 5	44.207 3	0.450 6	2.763 9	29.272 3	0.309 0
RFN-Nest	6.544 0	31.461 4	0.306 3	2.161 0	23.308 0	0.201 2
TUFusion	6.461 7	26.558 0	0.291 3	1.791 6	19.295 3	0.225 9
SwinFusion	6.865 0	37.828 8	0.376 5	3.849 1	38.208 4	0.404 1
SEDRFuse	6.953 5	45.204 8	0.388 5	4.133 2	42.021 6	0.250 9
GANMcC	6.254 5	27.712 2	0.225 0	2.080 6	21.530 5	0.182 1
U2Fusion	6.579 8	28.704 7	0.352 8	3.731 1	38.205 3	0.411 1
SwinFuse	6.767 0	42.404 8	0.390 5	4.301 9	43.661 3	0.322 3
FPDE	6.331 8	27.408 4	0.258 3	3.165 1	31.228 1	0.249 2
MSVD	6.319 1	27.178 5	0.261 0	2.621 5	25.583 1	0.218 6
本文	**7.285 6**	<u>47.841 5</u>	**0.572 6**	**4.658 6**	<u>46.848 0</u>	**0.631 8**

注：加粗、下划线字体分别表示各列最优、次优结果。

2.3　消融实验

为验证本文方法的有效性，本节对融合网络架构Transformer中跨模态信息交互分支（information interaction branch，I2 branch）和线性聚焦（focused linear）模块以及快速傅里叶变换损失函数（fast Fourier transform loss function，FFT Loss）进行消融实验。消融实验选取MSRS数据集中所有测试集，所有实验均使用相同的数据集和参数设置，定性和定量结果如图9和表3所示。

为了验证跨模态信息交互分支的有效性，设计了跨模态信息交互分支的消融实验（w/o I2 branch）。去除了网络架构Transformer中跨模态信息交互分支，仅保留自注意投影矩阵的生成，其余网络结构均保持不变。实验结果如图9（c）所示，相比去除跨模态信息交互分支的融合结果，本文方法保留了更多纹理细节，边缘更加清晰。这是由于去除跨模态信息交互分支后模型互补信息提取能力不足，导致光照不足条件下无法重现红外图像的锐利边缘。



(a) 红外图像　(b) 可见光图像　(c) w/o I2 branch　(d) w/o focused linear　(e) w/o FFT loss　(f) 本文

图 9　消融实验定性结果比较

Fig. 9　Comparison of qualitative results of ablation experiments

((a) infrared images; (b) visible images; (c) w/o I2 branch; (d) w/o focused linear; (e) w/o FFT loss; (f) ours)

表 3　消融实验定量结果比较

Table 3　Comparison of quantitative results of ablation experiments

方法	EN	SD	VIF	AG	EI	$Q^{AB/F}$
w/o FFT loss	5.619 3	27.966 3	0.498 1	3.377 5	35.374 6	0.623 5
w/o I2 branch	6.482 9	39.211 6	0.690 5	3.483 0	36.935 9	**0.686 9**
w/o focused linear	6.453 2	38.320 0	0.674 3	3.574 0	37.866 1	0.675 5
本文	**6.574 5**	**42.544 6**	**0.724 4**	**3.596 7**	**38.425 2**	0.680 3

注：加粗、下划线字体分别表示各列最优、次优结果。

如图 10 所示，通过将网络架构 Transformer 中的线性聚焦模块替换为 softmax 注意力模块，设计了线性聚焦模块的消融实验（w/o focused linear），以验证该模块在运行效率和融合质量方面的提升。需要注意的是，由于 focused linear 为线性复杂度，本文方法将单个像素作为 patch 序列，而基于 softmax 的消融方法为二次复杂度，为了保证模型能够在当前硬件条件的显存下运行，采用 16 × 16 的窗口划分 patch 序列。实验结果如图 9(d)所示，本文网络架构通过线性聚焦模块将 I2F Transformer 计算复杂度降至线性。本文方法与将此模块替换为 softmax 注意力模块的方法（w/o focused linear）相比，融合图像的视觉效果相当甚至更优。

为了验证快速傅里叶变换损失函数的有效性，设计了快速傅里叶变换损失函数的消融实验（w/o FFT loss），实验将本文提出的快速傅里叶变换损失函数替换为内容损失函数，损失函数参数保持不变，$\lambda_1 = 13$，$\lambda_2 = 7$，$\lambda_3 = 40$。内容损失函数为

$$L_{con} = \frac{1}{HW}\left(\left\| \boldsymbol{I}_f - \boldsymbol{I}_{ir} \right\|_1 + \left\| \boldsymbol{I}_f - \boldsymbol{I}_{vi} \right\|_1\right) \tag{18}$$

式中，H 和 W 分别表示图像的高和宽的像素数，$\|\cdot\|_1$ 表示 L1 范数。实验结果如图 9(e)所示，由于网络训练过程中缺乏傅里叶变换损失函数的约束，导致融合过程中重要频域信息丢失，融合结果整体背景亮度过低，纹理细节不清晰。

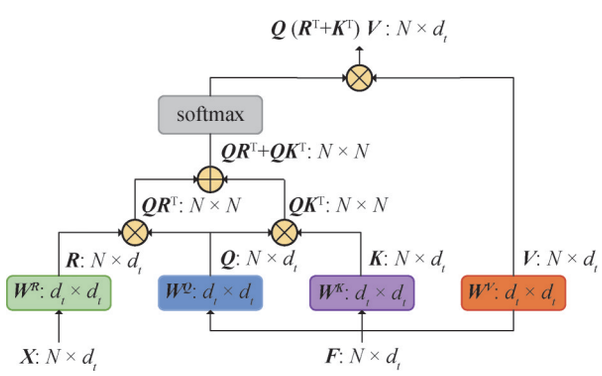


图 10　去除线性聚焦模块的注意力机制计算图

Fig. 10　The computation graph of MILF attention mechanism (w/o focused linear)

消融实验定量结果如表 3 所示。本文提出的完整网络架构在 EN、SD、VIF、AG 和 EI 上均取得最优值，$Q^{AB/F}$ 是基于噪声因素的质量评价指标，因此完整模型和去除信息交互分支的模型的 $Q^{AB/F}$ 数值（最优值）几乎一致。

相比于w/o I2 branch模型，本文方法中跨模态信息交互分支对于融合图像的AG和EI指标提升较大。相关定量指标的提升表明跨模态信息交互分支能够有效促进不同模态间的信息互补与特征融合，提高融合图像的清晰度。

本文方法与w/o focused linear模型对比表明，线性聚焦模块在提高模型计算效率的同时，融合图像质量显著提升，有关该模块运行效率的内容将在下文详细分析。

相比于w/o FFT loss模型，本文方法融合图像的EN、SD和VIF指标均有大幅度提升。这是由于图像频域变换后的幅度谱和相位谱能够突出目标结构强度和边缘信息，使得融合图像边缘纹理清晰度显著提高，取得了更优的视觉效果。

为进一步验证融合模型的计算效率，对本文方法和w/o focused linear模型架构进行了不同分辨率下的每秒浮点运算次数（floating point operations per second，FLOPs）和推理时间（inference time）对比。实验在RTX 3090 GPU上进行，实验结果如表4所示。

表4　本文方法与去除线性聚焦模块模型的每秒浮点运算次数和推理时间比较

Table 4　Comparison of FLOPs and inference time between the proposed method and module w/o focused linear

方法	分辨率/像素	FLOPs/G	推理时间/ms
w focused linear	32×32	3.5	6.2
w/o focused linear	32×32	9.6	16.0
w focused linear	64×64	6.3	6.5
w/o focused linear	64×64	25.6	45.4
w focused linear	96×96	14.2	12.4
w/o focused linear	96×96	111.9	220.6

两种模型在不同分辨率下的FLOPs和推理时间表现出显著差异。随着源图像分辨率的增加，本文模型架构的FLOPs增长相对平缓，而w/o focused linear架构的FLOPs呈现急剧增加趋势。源图像分辨率为96×96像素时，w/o focused linear架构的FLOPs接近本文模型架构计算量8倍。结果表明，本文方法处理高分辨率图像时具有更高的计算效率，能够更有效地利用计算资源，降低总体计算消耗。

推理时间的对比进一步验证了这一结论，随着分辨率的提升，本文模型的推理时间仅略有增加，而w/o focused linear模型的推理时间则显著上升。w/o focused linear模型在高分辨率下的性能瓶颈更加明显，导致推理效率大幅下降。

综上所述，消融实验的客观评估结果与主观评估结果保持一致，证明了本文方法提出的各个模块的有效性。

2.4　运行效率和时间复杂度分析

为比较不同图像融合方法的计算效率，本节实验测试了本文方法与对比方法的模型参数量及平均运行时间，所有测试均在NVIDIA GeForce RTX 3090 GPU上进行。

如表5所示，相比于RFN-Nest、TUFusion、SEDRFuse和SwinFuse方法，本文方法在模型参数量方面具有明显优势。计算效率方面，本文方法的推理时间优于大多数方法。本文提出的图像融合方法通过在网络架构中引入线性Transformer模块，相较于其他基于Transformer的方法（如SwinFusion和SwinFuse），在推理时间上优势明显。该优势源于线

表5　不同融合方法的模型参数量和推理时间比较

Table 5　Comparison of parameter and inference time for different fusion methods

方法	参数量/M	推理时间/s	
		MSRS	TNO
SeAFuison	0.16	0.22	0.06
SegMiF	0.99	0.82	0.28
CrossFuse	1.16	0.14	0.07
TarDAL	0.29	0.42	0.27
RFN-Nest	30.10	0.20	0.13
TUFusion	76.28	0.22	0.05
SwinFusion	0.93	2.44	1.39
SEDRFuse	3.31	2.22	2.23
GANMcC	1.83	0.64	0.71
U2Fusion	0.63	0.59	0.46
SwinFuse	8.09	0.48	0.34
FPDE	–	0.96	0.99
MSVD	–	0.37	0.40
本文	1.39	0.31	0.26

注：“–”表示未提供相应实验结果。

性Transformer在处理图像数据时的高效性，它减少了模型在前向传播过程中的计算负担。此外，TUFusion方法虽然在推理时间上优于本文方法，但TUFusion在图像融合前，将源图像统一下采样为256 × 256像素的输入数据，并在输出端利用线性插值将融合结果上采样为原始尺寸，这种下采样策略不可避免地牺牲了融合图像的部分细节和质量，导致融合效果的下降。

总而言之，本文方法在获得优越融合性能的同时，也具备了高效的计算效率，显著降低了基于Transformer架构融合网络的计算复杂度。

3 结 论

本文提出一种基于信息交互线性Transformer的红外与可见光图像融合算法，将跨模态信息交互机制引入图像融合方法中，并通过I2F Transformer中聚焦线性模块实现融合网络的轻量化。由于图像融合任务中红外与可见光图像均来自于不同传感器，源图像既包含一致特征又包含互补特征。通过跨模态信息交互机制，可以充分整合源图像的互补特征，去除冗余信息。此外，在网络训练过程中，本文设计了一种基于快速傅里叶变换的损失函数，显著提升了融合图像纹理清晰度和视觉保真度。从实验结果来看，本文算法有效改善了现有方法在复杂场景下难以突出目标特征信息以及融合网络复杂度高、训练困难的问题。在MSRS和TNO数据集上的对比实验表明，与其他13种现有融合方法相比，本文所提方法在提升视觉效果、突出目标信息以及保留复杂场景下纹理细节信息等方面具有优越性。

本文方法在I2F Transformer框架内，通过引入一种聚焦函数来替代传统的softmax函数，实现了对模型时间复杂度的有效降低。然而，这种替换导致了模型全局特征感知能力减弱，进而一定程度上影响了融合图像的对比度。此外，本文方法尽管在时间效率上相比于其他基于Transformer架构的方法显著提升，但与基于卷积神经网络的方法相比，其时间效率仍有待提高。在未来的研究中可以聚焦于模型结构和训练策略的优化，减少模型训练和推理所需的时间，从而缩小与基于卷积神经网络的方法在时间效率方面的差距。

参考文献(References)

Amarsaikhan D, Saandar M, Ganzorig M, Blotevogel H H, Egshiglen E, Gantuyal R, Nergui B and Enkhjargal D. 2012. Comparison of multisource image fusion methods and land cover classification. International Journal of Remote Sensing, 33(8): 2532-2550 [DOI: 10.1080/01431161.2011.616552]

Chen M S. 2016. Image fusion of visual and infrared image based on NSCT and compressed sensing. Journal of Image and Graphics, 21(1): 39-44 (陈木生. 2016. 结合NSCT和压缩感知的红外与可见光图像融合. 中国图象图形学报, 21(1): 39-44) [DOI: 10.11834/jig.20160105]

Cui G M, Feng H J, Xu Z H, Li Q and Chen Y T. 2015. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. Optics Communications, 341: 199-209 [DOI: 10.1016/j.optcom.2014.12.032]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. [s. l.]: OpenReview.net

Gao X Q, Liu G, Xiao G, Bavirisetti D P and Shi K L. 2020. Fusion algorithm of infrared and visible images based on FPDE. Acta Automatica Sinica, 46(4): 796-804 (高雪琴, 刘刚, 肖刚, Bavirisetti D P, 史凯磊. 2020. 基于FPDE的红外与可见光图像融合算法. 自动化学报, 46(4): 796-804) [DOI: 10.16383/j.aas.2018.c180188]

Han D C, Pan X R, Han Y Z, Song S J and Huang G. 2023. Flatten transformer: vision transformer using focused linear attention//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 5938-5948 [DOI: 10.1109/ICCV51070.2023.00548]

Han Y, Cai Y Z, Cao Y and Xu X M. 2013. A new image fusion performance metric based on visual information fidelity. Information Fusion, 14(2): 127-135 [DOI: 10.1016/j.inffus.2011.08.002]

Harsanyi J C and Chang C I. 1994. Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach. IEEE Transactions on Geoscience and Remote Sensing, 32(4): 779-785 [DOI: 10.1109/36.298007]

Jian L H, Yang X M, Liu Z, Jeon G, Gao M L and Chisholm D. 2021. SEDRFuse: a symmetric encoder - decoder with residual block network for infrared and visible image fusion. IEEE Transactions on Instrumentation and Measurement, 70: #5002215 [DOI: 10.1109/TIM.2020.3022438]

Jiang L M, Dai B, Wu W and Loy C C. 2021. Focal frequency loss for image reconstruction and synthesis//Proceedings of 2021 IEEE/

CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 13899-13909 [DOI: 10.1109/ICCV48922.2021.01366]

Li G L, Xiang W H, Zhang S L and Zhang B X. 2022. Infrared and visible image fusion algorithm based on residual network and attention mechanism. Unmanned Systems Technology, 5(2): 9-21 (李国梁, 向文豪, 张顺利, 张博勋. 2022. 基于残差网络和注意力机制的红外与可见光图像融合算法. 无人系统技术, 5(2): 9-21) [DOI: 10.19942/j.issn.2096-5915.2022.2.012]

Li H and Wu X J. 2019. DenseFuse: a fusion approach to infrared and visible images. IEEE Transactions on Image Processing, 28(5): 2614-2623 [DOI: 10.1109/TIP.2018.2887342]

Li H and Wu X J. 2024. CrossFuse: a novel cross attention mechanism based infrared and visible image fusion approach. Information Fusion, 103: #102147 [DOI: 10.1016/j.inffus.2023.102147]

Li H, Wu X J and Kittler J. 2021. RFN-Nest: an end-to-end residual fusion network for infrared and visible images. Information Fusion, 73: 72-86 [DOI: 10.1016/j.inffus.2021.02.023]

Liu J Y, Fan X, Huang Z B, Wu G Y, Liu R S, Zhong W and Luo Z X. 2022. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5792-5801 [DOI: 10.1109/CVPR52688.2022.00571]

Liu J Y, Liu Z, Wu G Y, Ma L, Liu R S, Zhong W, Luo Z X and Fan X. 2023. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 8081-8090 [DOI: 10.1109/ICCV51070.2023.00745]

Lu X Q, Zhang B H, Zhao Y, Liu H and Pei H Q. 2014. The infrared and visible image fusion algorithm based on target separation and sparse representation. Infrared Physics and Technology, 67: 397-407 [DOI: 10.1016/j.infrared.2014.09.007]

Ma J Y, Tang L F, Fan F, Huang J, Mei X G and Ma Y. 2022. SwinFusion: cross-domain long-range learning for general image fusion via swin transformer. IEEE/CAA Journal of Automatica Sinica, 9(7): 1200-1217 [DOI: 10.1109/JAS.2022.105686]

Ma J Y, Xu H, Jiang J J, Mei X G and Zhang X P. 2020. DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. IEEE Transactions on Image Processing, 29: 4980-4995 [DOI: 10.1109/TIP.2020.2977573]

Ma J Y, Yu W, Liang P W, Li C and Jiang J J. 2019. FusionGAN: a generative adversarial network for infrared and visible image fusion. Information Fusion, 48: 11-26 [DOI: 10.1016/j. inffus. 2018.09.004]

Ma J Y, Zhang H, Shao Z F, Liang P W 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: #5005014 [DOI: 10.1109/TIM.2020.3038013]

Naidu V P S. 2011. Image fusion technique using multi-resolution singular value decomposition. Defence Science Journal, 61(5): 479-484 [DOI: 10.14429/dsj.61.705]

Oppenheim A V and Lim J S. 1981. The importance of phase in signals. Proceedings of the IEEE, 69(5): 529-541 [DOI: 10.1109/PROC.1981.12022]

Paramanandham N and Rajendiran K. 2018. Infrared and visible image fusion using discrete cosine transform and swarm intelligence for surveillance applications. Infrared Physics and Technology, 88: 13-22 [DOI: 10.1016/j.infrared.2017.11.006]

Qu L H, Liu S L, Wang M N and Song Z J. 2022. TransMEF: a transformer-based multi-exposure image fusion framework using self-supervised multi-task learning//Proceedings of the 36th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 2126-2134 [DOI: 10.1609/aaai.v36i2.20109]

Rao Y J. 1997. In-fibre Bragg grating sensors. Measurement Science and Technology, 8(4): 355-375 [DOI: 10.1088/0957-0233/8/4/002]

Roberts J W, Van Aardt J A and Ahmed F B. 2008. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. Journal of Applied Remote Sensing, 2(1): #023522 [DOI: 10.1117/1.2945910]

Tang L F, Yuan J T and Ma J Y. 2022. Image fusion in the loop of high-level vision tasks: a semantic-aware real-time infrared and visible image fusion network. Information Fusion, 82: 28-42 [DOI: 10.1016/j.inffus.2021.12.004]

Tang L F, Zhang H, Xu H and Ma J Y. 2023. Deep learning-based image fusion: a survey. Journal of Image and Graphics, 28(1): 3-36 (唐霖峰, 张浩, 徐涵, 马佳义. 2023. 基于深度学习的图像融合方法综述. 中国图象图形学报, 28(1): 3-36) [DOI: 10.11834/jig.220422]

Tang W, He F Z and Liu Y. 2023. YDTR: infrared and visible image fusion via Y-shape dynamic transformer. IEEE Transactions on Multimedia, 25: 5413-5428 [DOI: 10.1109/TMM.2022.3192661]

Toet A. 2017. The TNO multiband image data collection. Data in Brief, 15: 249-251 [DOI: 10.1016/j.dib.2017.09.038]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wang Z S, Chen Y L, Shao W Y, Li H and Zhang L. 2022. SwinFuse: a residual swin transformer fusion network for infrared and visible images. IEEE Transactions on Instrumentation and Measurement, 71: #5016412 [DOI: 10.1109/TIM.2022.3191664]

Wu M H, Ma Y, Huang J, Fan F and Dai X B. 2020. A new patch-based two-scale decomposition for infrared and visible image fusion. Infrared Physics and Technology, 110: #103362 [DOI: 10.

1016/j.infrared.2020.103362］

Xu H，Ma J Y，Jiang J J，Guo X J and Ling H B. 2022. U2Fusion：a unified unsupervised image fusion network. IEEE Transactions on Pattern Analysis and Machine Intelligence，44（1）：502-518［DOI：10.1109/TPAMI.2020.3012548］

Xydeas C S and Petrović V. 2000. Objective image fusion performance measure. Electronics Letters，36(4)：308-309

Yang P，Gao L F and Zi L L. 2021. Image fusion method of convolution sparsity and detail saliency map analysis. Journal of Image and Graphics，26(10)：2433-2449（杨培，高雷阜，訾玲玲. 2021. 卷积稀疏与细节显著图解析的图像融合. 中国图象图形学报，26(10)：2433-2449）［DOI：10.11834/jig.200205］

Zhang B H，Lu X Q，Pei H Q and Zhao Y. 2015. A fusion algorithm for infrared and visible images based on saliency analysis and non-subsampled Shearlet transform. Infrared Physics and Technology，73：286-297［DOI：10.1016/j.infrared.2015.10.004］

Zhang Y，Liu Y，Sun P，Yan H，Zhao X L and Zhang L. 2020. IFCNN：a general image fusion framework based on convolutional neural network. Information Fusion，54：99-118［DOI：10.1016/j.inffus.2019.07.011］

Zhao Y Y，Zheng Q C，Zhu P H，Zhang X and Ma W P. 2024. TUFusion：a transformer-based universal fusion algorithm for multimodal images. IEEE Transactions on Circuits and Systems for Video Technology，34(3)：1712-1725［DOI：10.1109/TCSVT.2023.3296745］

Zhao Z X，Bai H W，Zhang J S，Zhang Y L，Xu S，Lin Z D，Timofte R and Van Gool L. 2023. CDDFuse：correlation-driven dual-branch feature decomposition for multi-modality image fusion//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：5906-5916［DOI：10.1109/CVPR52729.2023.00572］

作者简介

杨天宇，男，硕士研究生，主要研究方向为图像处理与模式识别。E-mail：2023211474@stu.ppsuc.edu.cn

霍宏涛，通信作者，男，教授，博士生导师，主要研究方向为图像处理与模式识别、遥感应用技术和图像取证。

E-mail：huohongtao@ppsuc.edu.cn

刘晓文，男，博士研究生，主要研究方向为模式识别和计算机视觉。E-mail：liu_x_wen@163.com

郑博文，男，硕士研究生，主要研究方向为图像处理和图像融合。E-mail：2022211499@stu.ppsuc.edu