激光与光电子学进展 先进成像

联合卷积神经网络和转换器的红外与可见光图像融合

杨阳，任振南*，李北辰
天津大学电气自动化与信息工程学院，天津 300072

摘要 为解决在红外与可见光图像融合领域中，卷积神经网络(CNN)无法建模源图像内部的全局语义相关性，对图像上下文信息利用不充分等问题，创新性地提出了一种联合CNN和转换器(Transformer)的图像融合模型。首先，为了弥补CNN无法建立长程依赖关系的缺陷，提出了联合CNN和Transformer的编码器，加强了对多个局部区域间相关性的特征提取，提高了模型对图像局部细节信息的提取能力。其次，提出了一种基于模态最大差异度的融合策略，强化融合过程中对源图像不同区域信息的自适应表达，提高了融合图像的对比度。最后，在TNO公开数据集上，联合多种对比算法对所提融合模型进行了实验验证。实验结果表明，在主观视觉效果和客观评价指标两方面的评估上，所提模型与现有融合方法相比都具有明显优势。此外，通过消融实验，分别对提出的联合编码器和融合策略进行了有效性分析，实验结果证明所提出的设计思想在红外可见光图像融合任务上是有效的。

关键词 图像融合；卷积神经网络；转换器；注意力机制；红外图像
中图分类号 TN911.73 **文献标志码** A **DOI**：10.3788/LOP222265

Infrared and Visible Image Fusion with Convolutional Neural Network and Transformer

Yang Yang, Ren Zhennan*, Li Beichen
School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract An innovative image fusion model combining convolutional neural network (CNN) and Transformer is proposed to address the issues of the CNN's inability to model the global semantic relevance within the source image and insufficient use of the image context information in infrared and visible image fusion field. First, to compensate for the shortcomings of CNN in establishing long-range dependencies, a combined CNN and Transformer encoder was proposed to improve the feature extraction of correlation between multiple local regions and improve the model's ability to extract local detailed information of images. Second, a fusion strategy based on the modal maximum disparity was proposed for better adaptive representation of information from various regions of the source image during the fusion process, enhancing the fused image's contrast. Finally, by comparing with multiple contrast methods, the fusion model developed in this research was experimentally confirmed using the TNO public dataset. The experimental results demonstrate that the suggested model has significant advantages over existing fusion approaches in terms of both subjective visual effects and objective evaluation metrics. Additionally, through ablation tests, the efficiency of the suggested combined encoder and fusion technique was examined separately. The findings of the experiments further support the effectiveness of the design concept for the infrared and visible image fusion assignments.

Key words image fusion; convolutional neural network; Transformer; attentional mechanism; infrared image

1 引 言

不同感光元件对不同波长电磁波的敏感程度不同，这就导致在利用单一成像设备对环境进行拍摄时，只能从单模态的图像(即来源于单一成像源的图像)的部分谱段对环境进行感知，从而仅能够获取对应该谱段的部分场景信息。例如，红外传感器能够捕获物体的热辐射能量，因此可以用来检测和追踪发热目标[1-2]，但无法获取纹理、背景等细节信息；可见光传感器可以通过接收物体的反射光来对场景细节、纹理特

收稿日期：2022-08-12；**修回日期**：2022-10-18；**录用日期**：2022-10-27；**网络首发日期**：2022-11-04
基金项目：国家重点研发计划(2021YFE0204200)、国家自然科学基金(62101378，62171318)
通信作者：*Ren2151311@163.com

征等进行描述[3]，但是无法在光线不足的情况下对目标进行探测。为了最大程度地保留场景信息，近年来研究人员开展了对多模态图像融合的研究[4-5]。得益于融合图像的信息量相较于任意单一模态的图像都更为丰富的优势，研究人员也同时对融合图像在多种任务上的应用展开了相关工作，这些应用领域不仅包括目标检测、目标跟踪等计算机视觉任务，也涵盖了遥感图像全色锐化、医疗影像处理、军事目标检测等工程应用。因此，研究红外与可见光的图像融合具有重要的科研价值和工程意义。

当前，现有的红外与可见光图像融合主要分为基于变换域的方法、基于空间域的方法和基于深度学习的方法等[6]。其中，基于变换域的方法出现最早，从金字塔变换[7-8]、小波变换[9-12]、稀疏表示[13]等思想出发，将图像转换到变换域中，通过对源图像进行分解和融合，再进行逆变换重构得到融合结果。基于空间域的方法通常侧重源图像中的显著性区域[14-15]，利用分块的思想在空间域上对源图像中不同区域的信息进行活跃度的度量[16]，以充分考虑局部区域内的像素相关性，然后利用多种图像处理方法实现图像融合。

相较于传统神经网络模型，深度神经网络由于层数的增加，具备了更强的特征提取能力，能够有效解决传统网络对数据表示能力弱的问题，因此近年来出现了许多基于卷积神经网络（CNN）模型的图像融合技术[17]。Li等[18]使用卷积神经网络代替传统的表示模型，对源图像进行特征提取和重构，通过采用密集连接网络结构及独特的融合策略获得了较好的融合结果。随后，Li等[19]对提取特征的卷积网络结构和融合策略分别进行了改进，利用多尺度的思想对源图像中的特征在粗粒度和细粒度的尺度上分别进行融合，再次提升了融合效果。基于生成对抗网络等图像生成的方法[20-21]，在人工设定的真实标记（ground truth）的引导下，通过生成器和鉴别器之间的对抗性学习，避免了融合规则的设计，在图像融合领域也取得了一定的创新性研究突破。然而，上述这些基于CNN的方法大多依赖卷积核与源图像之间的局部交互，这导致深度图像特征只能通过有限的感受野来建立。在这种情况下，模型忽略了不同区域之间的依赖性，因此无法利用图像中存在的上下文信息指导图像融合过程，融合图像的质量会明显降低。

针对上述问题，本文创新性地提出了一种联合CNN和Transformer的图像融合框架，该框架采用了编码器-解码器结构，并利用了在自然图像数据集上的预训练模型，提升了对多模态图像特征提取的有效性。其中，编码器利用了CNN局部空间上下文信息以及Transformer长距离上下文信息的提取能力，对多源图像的局部和全局特征进行了提取。在此基础上，对单模态图像提取的全局与局部特征进行了预融合，使得编码器获取了不同局部区域之间的信息差异。这种做法加强了对图像的局部细节信息的提取，提升了融合结果的整体质量。此外，本文还针对红外和可见光图像融合设计了一种基于模态最大差异度的融合策略。该融合策略能通过红外与可见光图像中的特征差异引导多模态数据融合，强化融合过程中对源图像不同区域信息的自适应表达，使融合图像更符合人类的视觉感知。实验结果表明，所提模型获得的融合图像能够在突出来自红外图像中热辐射目标的同时保留来自可见光图像中丰富的纹理细节等信息，其整体的亮度以及对比度较高，较对比算法有明显改善。

2 Transformer和注意力机制

2.1 视觉任务中的Transformer

Transformer最开始用于自然语言处理（NLP）领域中序列到序列之间的预测建模[22]，最近逐渐被开发到各种计算机视觉任务中。Dosovitskiy等[23]首次提出了用于图像的Transformer模型——视觉转换器（VIT），该模型将图像分为16×16像素的图像块，然后将不同图像块之间构成的伪序列关系输入Transformer模块来提取图像特征。VIT充分利用了Transformer的序列建模的特性，因此在图像分类、图像分割、图像生成等任务上都发挥出了较为明显的作用。许多研究[24-25]表明，图像融合性能的提升依赖于对源图像特征的表示。Transformer擅长建模图像的上下文，能够弥补CNN无法建模长距离依赖关系的不足，因此将CNN和Transformer以并行的方式相结合对图像进行特征提取，能够使融合网络更加全面地利用图像中的局部信息和全局信息。

2.2 注意力机制

深度学习领域中注意力机制的本质是定位感兴趣的区域，对网络提取的深层特征进行重新校准，对某些特定区域给予更高权重。Hu等[26]提出了Squeeze and Excitation Net（SENet），通过压缩模块和激励模块在特征图的通道之间建立依赖关系，从而能够自适应地调整特征图通道之间的权重，实现通道维度上局部信息的整合。但是该网络只采取了通道注意力机制，没有涉及空间注意力模块，对特征图局部信息与全局信息的建模程度有限，无法全面提取特征图的关键语义信息。Woo等[27]提出了bottleneck attention module（BAM）和convolutional block attention module（CBAM）两种同时考虑特征图通道维度和空间维度的注意力模块，其中BAM网络中通道注意力模块和空间注意力模块采用并联结构，CBAM中两种注意力模块则采用串联结构，这两种结构均能够关注特征图中的区分性信息，有利于对关键语义信息的提取。

3 模型设计与融合方法

3.1 融合网络总体框架

所提图像融合模型如图1所示。该模型主要包括

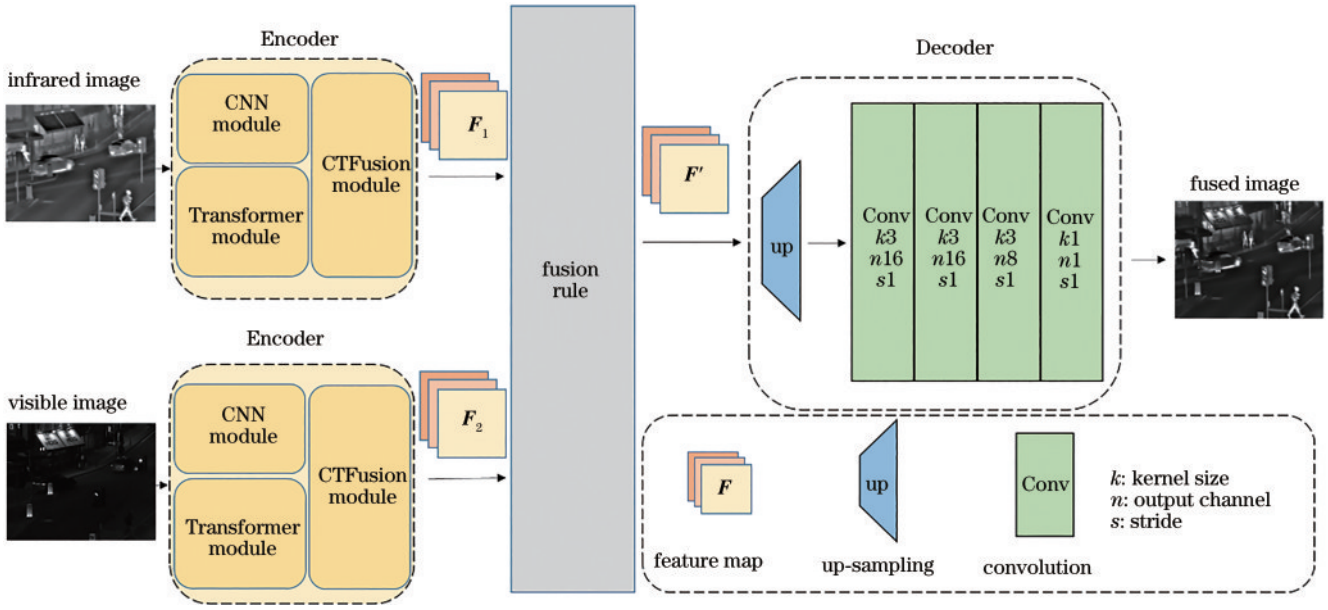

图 1　所提模型的整体结构

Fig. 1　Overall structure of the proposed model

3 个部分，分别是用于提取源图像特征的编码器（Encoder）网络、融合策略、用于重构图像的解码器（Decoder）网络。图像融合的整体流程为：红外图像和可见光图像分别被送入编码器网络中进行特征提取，得到各自的特征图 $\boldsymbol{F}_1$ 和 $\boldsymbol{F}_2$；随后通过特定的融合策略对这两种特征图进行运算，得到融合后的特征图 $\boldsymbol{F}'$；最后对这些融合后的特征图通过解码器进行重构，得到最终的融合结果。

融合模型中编码器网络由 3 个模块组成，分别为 CNN 模块、Transformer 模块及 CTFusion 模块。其中 CNN 模块和 Transformer 模块分别用来提取源图像中的局部特征和全局特征，在 CTFusion 模块中对这些特征信息进行整合，从而连接多个局部区域间的相关性，实现全局信息与局部信息的有效结合。解码器网络由 1 个上采样网络和 4 个卷积网络组成，卷积核中的参数如图 1 所示。特征图在经过卷积层后对其进行谱归一化（spectral normalization）操作[28]，即对模型中的权重矩阵进行二范数正则化，从而增强模型训练过程的稳定性。激活函数设定为带泄露线性整流函数（Leaky ReLU）。最后，卷积层输出的特征图通过 Tanh 函数后得到生成图像。

3.2　编码器结构

3.2.1　CNN 模块

CNN 模块被设计用来提取图像的局部特征，它由 1 个下采样操作和 4 个卷积层组成。下采样由步长为 2、卷积核尺寸为 3 的卷积操作来实现，使用镜像填充的方式保证特征图高和宽的尺寸减半。卷积层的具体参数如图 2 所示。特征图在经过每个卷积层后同样经过谱归一化操作和非线性激活操作，最后输入到 CTFusion 模块。

3.2.2　Transformer 模块

在 Transformer 模块中对源图像进行全局特征的提取，其首先经过 patch merging 层[29]进行双倍下采样，得到图像 $\boldsymbol{I}\in\mathbb{R}^{H\times W}$，该图像随后被裁剪成个数为 $N=\frac{H}{P}\times\frac{W}{P}$ 的图像块，其大小为 $p\times p$，这些图像块组成了序列 $\boldsymbol{x}_{\text{seq}}\in\mathbb{R}^{N\times P^2}$；这些序列通过线性投影（linear projection）层进行特征映射后，得到了编码后的特征序列 $\boldsymbol{z}\in\mathbb{R}^{N\times D}$，其中 D 为特征嵌入的维度；随后，该特征序列被送入 6 层具有多头自注意力（MSA）和多层感知机（MLP）的 Transformer 编码器中进行特征提取。其中自注意力机制（SA）是 Transformer 的核心原理，它能够通过聚合特征图中的全局信息来更新特征序列，原理公式为

$$\mathrm{SA}(\boldsymbol{z})=\mathrm{Softmax}(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{d}\,)\boldsymbol{V}\,,\tag{1}$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别代表查询（Query）、键（Key）、值（Value）；d 表示 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的维度；Softmax 函数能够根据 Query 和 Key 的查询结果对 Value 的重要程度提供从 0 到 1 的注意力值。$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 通过线性变换得来：

$$\boldsymbol{Q}=\boldsymbol{W}_{\mathrm{q}}\boldsymbol{z},\boldsymbol{K}=\boldsymbol{W}_{\mathrm{k}}\boldsymbol{z},\boldsymbol{V}=\boldsymbol{W}_{\mathrm{v}}\boldsymbol{z}\,,\tag{2}$$

式中：$\boldsymbol{W}_{\mathrm{q}}$、$\boldsymbol{W}_{\mathrm{k}}$ 和 $\boldsymbol{W}_{\mathrm{v}}$ 为可训练的变换矩阵。

特征序列在通过 MSA 模块和 MLP 模块之前首先进行层归一化（LayerNorm），并在之后应用残差连接（residual connection）进行特征增强。MLP 层由 2 个带有 GeLU 激活函数的线性层组成。在经过 6 层 Transformer 结构后，特征序列被重组为特征图后同样进入到 CTFusion 模块。

3.2.3　CTFusion 模块

为了有效整合来自 CNN 模块和 Transformer 模块的特征，连接多个局部区域之间的相关性，使网络能够

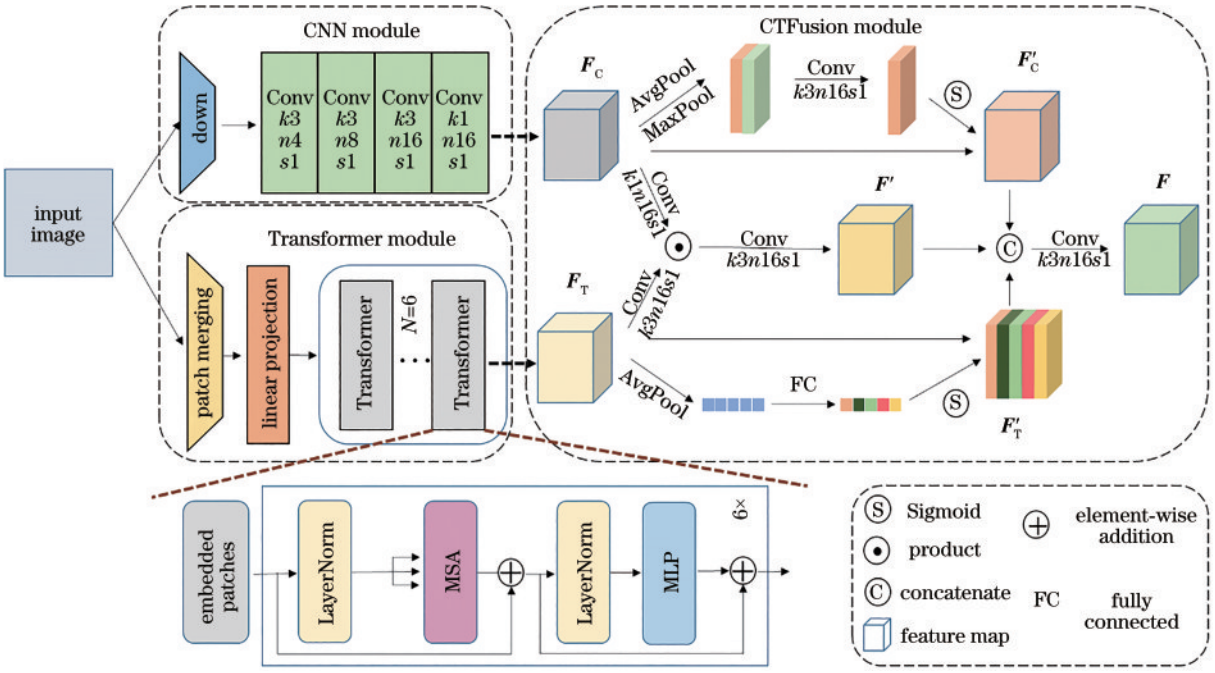

图 2　编码器的具体结构

Fig. 2　Concrete structure of the Encoder

同时关注到图像中的局部信息和全局信息，设计了CTFusion模块，如图 2 所示。具体来说，得到最终融合特征的表达式为

$$\begin{cases} \boldsymbol{F}'_{\mathrm{C}} = \mathrm{SpatialAtt}(\boldsymbol{F}_{\mathrm{C}}) \\ \boldsymbol{F}'_{\mathrm{T}} = \mathrm{ChannelAtt}(\boldsymbol{F}_{\mathrm{T}}) \\ \boldsymbol{F}' = \mathrm{Conv}\left[\mathrm{Conv}(\boldsymbol{F}'_{\mathrm{C}}) \odot \mathrm{Conv}(\boldsymbol{F}'_{\mathrm{T}})\right] \\ \boldsymbol{F} = \mathrm{Conv}\left[\mathrm{Concat}(\boldsymbol{F}'_{\mathrm{C}}, \boldsymbol{F}'_{\mathrm{T}}, \boldsymbol{F}')\right] \end{cases}, \tag{3}$$

式中：SpatialAtt为空间注意力机制[27]，这里被用来增强CNN分支提取到的局部特征并抑制无用信息；ChannelAtt为通道注意力机制[26]，这里被用来强化来自Transformer分支的全局信息。同时，CTFusion模块利用哈达玛乘积⊙与卷积操作来建模局部特征$\boldsymbol{F}_{\mathrm{C}}$和全局特征$\boldsymbol{F}_{\mathrm{T}}$之间的细粒度交互，从而获取不同局部区域之间的信息差异。最后，在通道维度对得到的特征图$\boldsymbol{F}'_{\mathrm{C}}$、$\boldsymbol{F}'_{\mathrm{T}}$、$\boldsymbol{F}'$进行拼接（Concat）和卷积运算，强化编码器对图像局部细节信息的提取，进而得到最终的融合特征图$\boldsymbol{F}$。

3.3　模型的训练

本模型的两个编码器网络参数共享（只需对单个编码器进行训练）。在训练过程中，融合策略被移除，单张训练图像被输入编码器网络得到特征图后，直接经过解码器网络重构回原图像。训练完成后，首先固定编码器和解码器的网络权重，并利用融合策略对编码器提取到的深度特征进行融合，然后通过解码器生成最后的融合图像。

为了训练编码器和解码器分解和重构图像的能力，需要最小化原图像和重构图像之间的误差，因此损失函数设置为

$$L = L_{\mathrm{MSE}} + \lambda L_{\mathrm{SSIM}}, \tag{4}$$

式中：L_{MSE}表示均方误差（MSE）损失；L_{SSIM}表示结构相似性（SSIM）损失；λ为调节这两项权重的超参数。

均方误差损失能够保证重构图像对原图像在像素级别上的重建精度，定义为

$$L_{\mathrm{MSE}} = \left\| \boldsymbol{I}_{\mathrm{re}} - \boldsymbol{I}_{\mathrm{or}} \right\|_2, \tag{5}$$

式中：$\boldsymbol{I}_{\mathrm{re}}$为重建图像；$\boldsymbol{I}_{\mathrm{or}}$为输入网络的原图像。结构相似性损失能够帮助模型更好地提取到原图像的结构信息，定义为

$$L_{\mathrm{SSIM}} = 1 - \mathrm{SSIM}(\boldsymbol{I}_{\mathrm{re}}, \boldsymbol{I}_{\mathrm{or}}), \tag{6}$$

式中：SSIM(·)为结构相似性运算[30]。

3.4　融合策略

为了使融合图像更好地突出红外图像中的发热目标以及保留可见光图像中背景纹理等信息，设计了一种新的融合策略，具体流程如图 3 所示。

为了反映红外与可见光图像中的显著特征，首先对红外特征图$\boldsymbol{F}_{\mathrm{ir}}$和可见光特征图$\boldsymbol{F}_{\mathrm{vi}}$作减法来计算二者之间的差异，然后对差值特征图与该差值特征图的最大差异度进行除法运算，得到相对差异权重$\boldsymbol{\mu}_1$和$\boldsymbol{\mu}_2$，公式分别为

$$\boldsymbol{\mu}_1 = \frac{\boldsymbol{F}_{\mathrm{vi}} - \boldsymbol{F}_{\mathrm{ir}}}{\max(\boldsymbol{F}_{\mathrm{vi}} - \boldsymbol{F}_{\mathrm{ir}})}, \tag{7}$$

$$\boldsymbol{\mu}_2 = \frac{\boldsymbol{F}_{\mathrm{ir}} - \boldsymbol{F}_{\mathrm{vi}}}{\max(\boldsymbol{F}_{\mathrm{ir}} - \boldsymbol{F}_{\mathrm{vi}})}, \tag{8}$$

式中：$\boldsymbol{F}_{\mathrm{vi}}$和$\boldsymbol{F}_{\mathrm{ir}}$分别为编码器输出的可见光图像和红外图像的特征图；max(·)对差值特征图进行全局深度最大池化操作，得到两种模态图像之间的最大差异度。

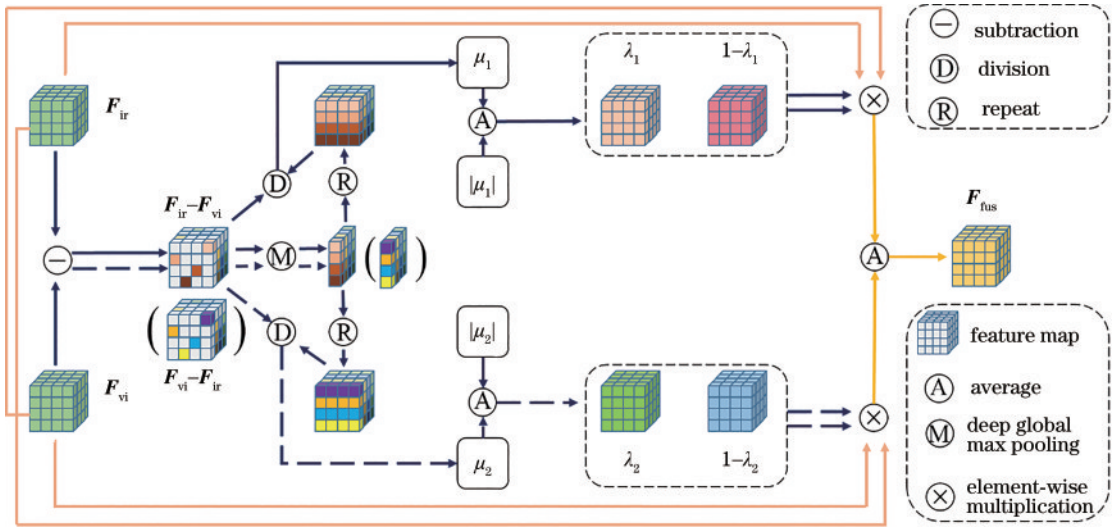

图 3　融合策略

Fig. 3　Fusion strategy

为了根据红外与可见光特征图中像素的重要性对其进行权重分配，将所得相对差异权重与其绝对值相加并求均值，结果作为最终的自适应加权图 λ_1 和 λ_2 来引导融合过程，

$$\lambda_1=\frac{\mu_1+|\mu_1|}{2},\tag{9}$$

$$\lambda_2=\frac{\mu_2+|\mu_2|}{2},\tag{10}$$

式中：λ_1 和 λ_2 与初始特征图具有相同维度。然后将该自适应加权图和初始特征图相乘，生成预融合特征图 $\boldsymbol{F}_1$ 和 $\boldsymbol{F}_2$，公式为

$$\boldsymbol{F}_1=\lambda_1\times\boldsymbol{F}_{vi}+(1-\lambda_1)\times\boldsymbol{F}_{ir},\tag{11}$$

$$\boldsymbol{F}_2=\lambda_2\times\boldsymbol{F}_{ir}+(1-\lambda_2)\times\boldsymbol{F}_{vi}。\tag{12}$$

最终对两个预融合特征图进行平均运算，得到最终的融合特征图 $\boldsymbol{F}$，公式为

$$\boldsymbol{F}=(\boldsymbol{F}_1+\boldsymbol{F}_2)/2。\tag{13}$$

4　实验与讨论

4.1　数据集与实验配置

利用 MS-COCO 数据集[31]对编码器和解码器网络以图像重建为目标进行了训练，经训练得到了编码器网络和解码器网络的权重；之后，利用该组权重初始化 2 个编码器和 1 个解码器，两个编码器分别用于对不同模态数据进行特征提取，解码器负责最终融合图像的输出；最终，在初始化后的模型上，利用 TNO 数据集进行了图像融合效果的测试。MS-COCO 数据集包含 82783 张不同场景的彩色自然图像，在训练过程中，这些图像被转换为灰度图像，同时尺寸也被调整为 256×256。TNO 公开数据集[32]包含了经过多波段摄像系统（如 Athena、DHV、FEL 系统）配准后的多对可见光和红外多源模态图像，其中红外波段包含近红外、长红外以及热红外等，这些图像具备种类多和外部环境变化显著的特点，使用较为广泛。

从 MS-COCO 数据集中选取 80000 张图像作为训练数据，从 TNO 数据集中选取了 15 组红外-可见光图像对作为测试数据。对测试数据按照 256×256 的尺寸进行了裁剪，最终得到了 367 组图像对作为模型的测试集。在测试过程结束后，对融合后的图像块在裁剪的位置进行拼接，得到了融合图像。

所提模型采用了 Adam 优化器进行了训练，设置初始学习率为 0.0001，通过指数衰减进行学习率的调节。损失函数中超参数 λ 设置为 10，Transformer 中多头自注意（MSA）的数量为 8，特征的嵌入维度为 256。每次训练选取 32 个样本，总共训练 80 个轮次（training epoch）。所提模型和所有的对比算法均利用了 NVIDIA GTX 3080 GPU 进行训练。

4.2　融合结果主观评价

人类视觉系统（HVS）是一个低通线性系统，在对融合实验结果进行主观评价时，需要同时考虑到亮度和对比度较色度信号的敏感程度，此外图像的边缘信息对主观视觉评价也十分重要[33]。

为了验证所提模型在红外与可见光图像上的融合效果，对所提模型、2 种传统融合模型、5 种基于深度学习的融合模型进行比较。传统融合模型包括基于金字塔变换的模型 RP[34]和基于小波变换的模型 Wavelet[35]；基于深度学习的融合模型包括基于卷积神经网络的模型 ResNet-ZCA[36]和 Dual-Branch[37]、基于自动编码器网络的模型 DenseFuse[18]、基于生成对抗网络的模型 FusionGAN[20]和 GANMcC[38]。各模型的部分融合结果如图 4～7 所示，其中粗线方框包围的区域强调了红外图像中的热辐射目标，细线方框包围的区域强调了可见光图像中的纹理细节，且在每幅融合图像的左下角放大了该区域中的信息。

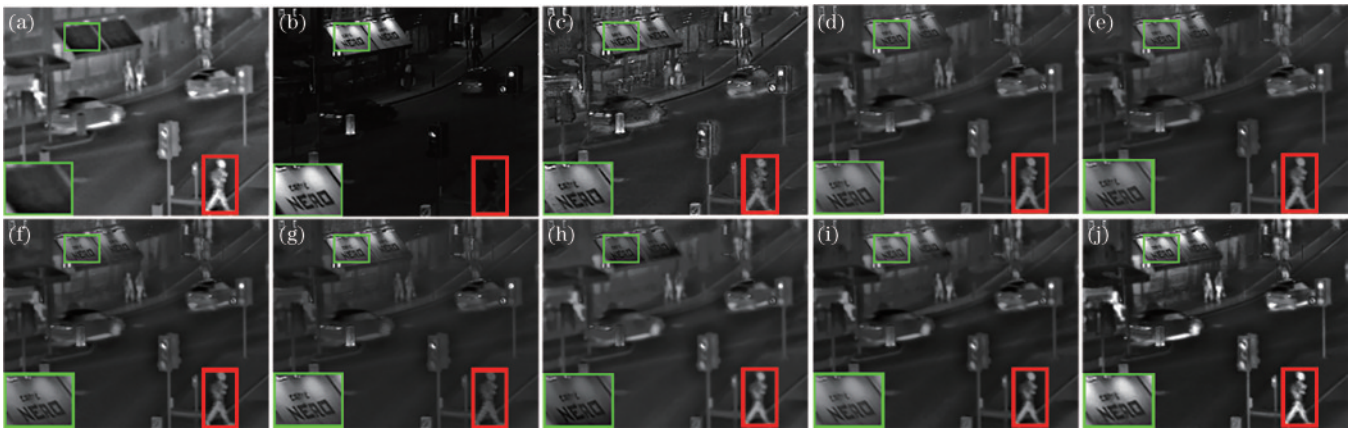

图 4 “Street”图像的融合结果。(a)红外图像;(b)可见光图像;(c)RP;(d)Wavelet;(e)ResNet-ZCA;(f)DenseFuse;(g)Dual-Branch;(h)FusionGAN;(i)GANMcC;(j)所提方法

Fig. 4 Fusion results of the “Street” image. (a) Infrared image; (b) visible image; (c) RP; (d) Wavelet; (e) ResNet-ZCA; (f) DenseFuse; (g) Dual-Branch; (h) FusionGAN; (i) GANMcC; (j) proposed method

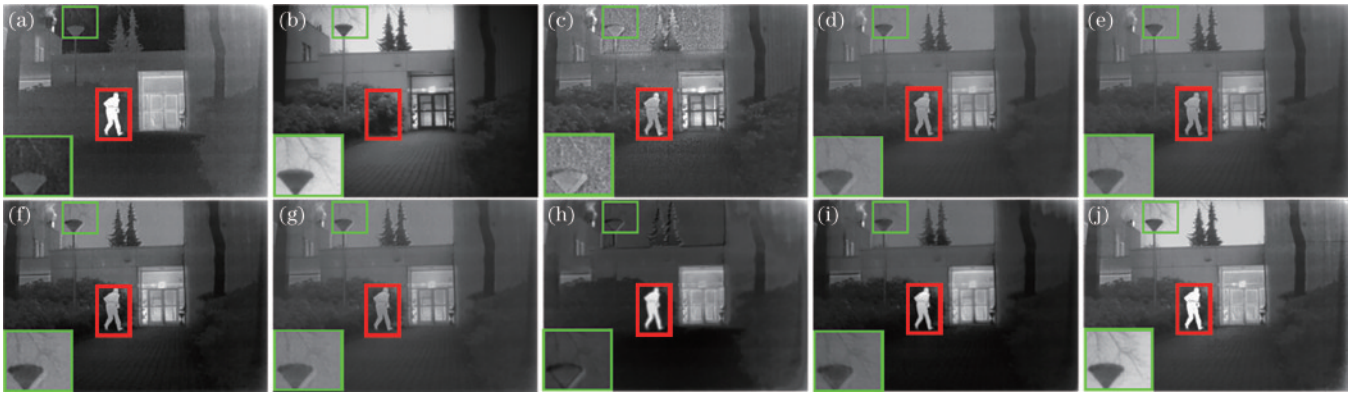

图 5 “Kaptein_1123”图像的融合结果。(a)红外图像;(b)可见光图像;(c)RP;(d)Wavelet;(e)ResNet-ZCA;(f)DenseFuse;(g)Dual-Branch;(h)FusionGAN;(i)GANMcC;(j)所提方法

Fig. 5 Fusion results of the “Kaptein_1123” image. (a) Infrared image; (b) visible image; (c) RP; (d) Wavelet; (e) ResNet-ZCA; (f) DenseFuse; (g) Dual-Branch; (h) FusionGAN; (i) GANMcC; (j) proposed method

图 6 “Kaptein_1654”图像的融合结果。(a)红外图像;(b)可见光图像;(c)RP;(d)Wavelet;(e)ResNet-ZCA;(f)DenseFuse;(g)Dual-Branch;(h)FusionGAN;(i)GANMcC;(j)所提方法

Fig. 6 Fusion results of the “Kaptein_1654” image. (a) Infrared image; (b) visible image; (c) RP; (d) Wavelet; (e) ResNet-ZCA; (f) DenseFuse; (g) Dual-Branch; (h) FusionGAN; (i) GANMcC; (j) proposed method

根据图 4~7 的融合图像结果可以看出,相较于用于对比实验的大部分融合方法,所提方法的融合图像的对比度更高,所提方法不仅较好地保留来自可见光原始图像中的纹理和结构等信息,而且更好地描述了原始红外图像的温度变化情况,对可见光及红外图像信息的保留程度更好。在图 4 的融合结果中,其他模

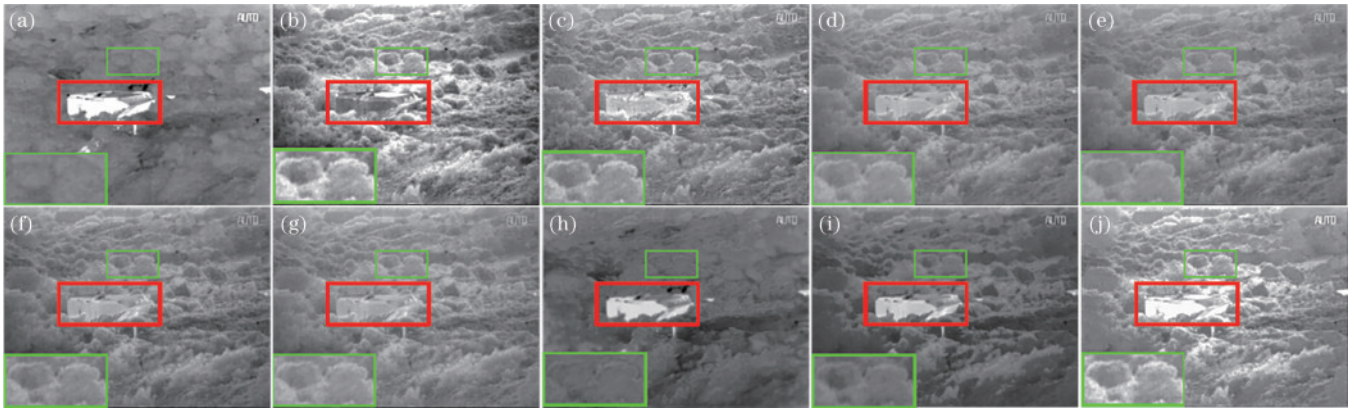

图 7　“Bunker”图像的融合结果。(a)红外图像;(b)可见光图像;(c)RP;(d)Wavelet;(e)ResNet-ZCA;(f)DenseFuse;(g)Dual-Branch;(h)FusionGAN;(i)GANMcC;(j)所提方法

Fig. 7　Fusion results of the “Bunker” image. (a) Infrared image; (b) visible image; (c) RP; (d) Wavelet; (e) ResNet-ZCA; (f) DenseFuse; (g) Dual-Branch; (h) FusionGAN; (i) GANMcC; (j) proposed method

型处理的粗线方框中的行人目标相较于原始红外图像在亮度上均有较大程度降低,而所提模型依然能够突出该目标的热辐射信息,这表明所提模型对红外图像信息的保留程度较好,此外,所提模型处理的细线方框中的灯光及文字等细节丰富程度也较其他算法更好,这表明所提模型能够对可见光信息进行有效的提取和重建。在图 5 中:RP 算法的融合结果中背景部分存在较大噪声;Wavelet、ResNet-ZCA 和 Dual-Branch 的融合图像的对比度较低,偏向于红外图像和可见光图像取平均的结果;FusionGAN 和 GANMcC 的融合图像中可见光图像中细节信息丢失较多,融合结果偏向于原始红外图像;虽然 DenseFuse 对于红外图像中的热辐射目标和可见光图像中的细节信息均有较为完整的保存,但是所提模型的融合图像有更好的视觉效果和更丰富的细节信息。图 6 和图 7 的融合结果与图 5 类似,Wavelet、ResNet-ZCA、DenseFuse 和 Dual-Branch 虽然对行人和碉堡等红外热辐射目标以及路灯、灌木等纹理细节有所保留,但是图像对比度较低,各个特征无法得到显著表达;RP 模型的融合结果要优于以上四种模型,能够较好地保存图 6 细线方框中的路灯和图 7 粗线方框中的碉堡;FusionGAN 和 GANMcC 无法有效保留来自可见光图像中的各种细节信息,融合结果与原始红外图像的相似程度更高。总体而言,所提模型一方面能够突出行人及碉堡等发热目标,保留了更丰富的红外信息,另一方面对路灯、帐篷、雨伞、灌木等细节区域进行了较好的重建,对可见光图像中的信息实现了更有效的表达。

4.3　融合结果客观评价

除主观评价分析,还选用 6 个图像客观评价指标,基于融合图像和多模态源图像,对不同融合方法的融合性能进行定量评估[39-40]。

1) 面向融合图像质量的信息熵(En)

$$V_{\mathrm{En}}=-\sum_{i}P(F_i)\log_2 P(F_i),\tag{14}$$

式中:$P(F_i)$是融合图像的灰度直方图。En 越大,证明融合结果中信息量越多,融合方法的性能越好。

2) 标准差(SD)

$$V_{\mathrm{SD}}=\sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left[F(i,j)-\mu\right]^2},\tag{15}$$

式中:$F(i,j)$表示在大小为 $M\times N$ 的图像上(i,j)位置处的像素强度;μ 表示该图像上所有像素强度的均值。标准差越大,表明融合图像中具有越分散的灰度级分布,融合结果的质量越好。

3) 空间频率(SF)

$$V_{\mathrm{SF}}=\sqrt{{F_{\mathrm{R}}}^2+{F_{\mathrm{C}}}^2},\tag{16}$$

式中:F_{R} 和 F_{C} 分别为图像的行频率和列频率。SF 反映图像灰度的变化率,也可用于反映图像的清晰度。一般地,SF 值越大,空间频率越高,图像越清晰。

4) 互信息(MI)

$$V_{\mathrm{MI}\,X,Y}=\sum_{x,y}p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)},\tag{17}$$

$$V_{\mathrm{MI}}=V_{\mathrm{MI\,ir,\,fus}}+V_{\mathrm{MI\,vi,\,fus}},\tag{18}$$

式中:$p_{X,Y}(x,y)$表示图像 X 和 Y 之间的联合分布;$p_X(x)$和$p_Y(y)$分别表示二者的边缘分布互信息,用来表征两张图像之间的相关性。对于图像融合任务来说,总的 MI 为对两张原始图像和融合图像分别计算互信息值 $V_{\mathrm{MI\,ir,\,fus}}$ 和 $V_{\mathrm{MI\,vi,\,fus}}$ 再相加的结果。MI 值越高,融合结果含有越多来自原始图像中的信息。

5) 差异相关性之和(SCD)

SCD[41]是一种通过计算源图像与融合图像之间的差值图像然后再与源图像计算相关系数之和的图像融合评价方法,该指标越大,与主观目视的评价越一致。

6) 新型图像质量评估指标 Q_abf

Q_abf[42]能够通过计算源图像与融合图像之间的局部度量来估计源图像中的显著信息在融合图像中的表现程度,该指标越高,融合图像对于源图像中的显著

特征保留得越完整。

利用 6 个评价指标进行融合结果的性能评估，不同融合模型在 TNO 数据集上的客观评价结果如图 8 所示。

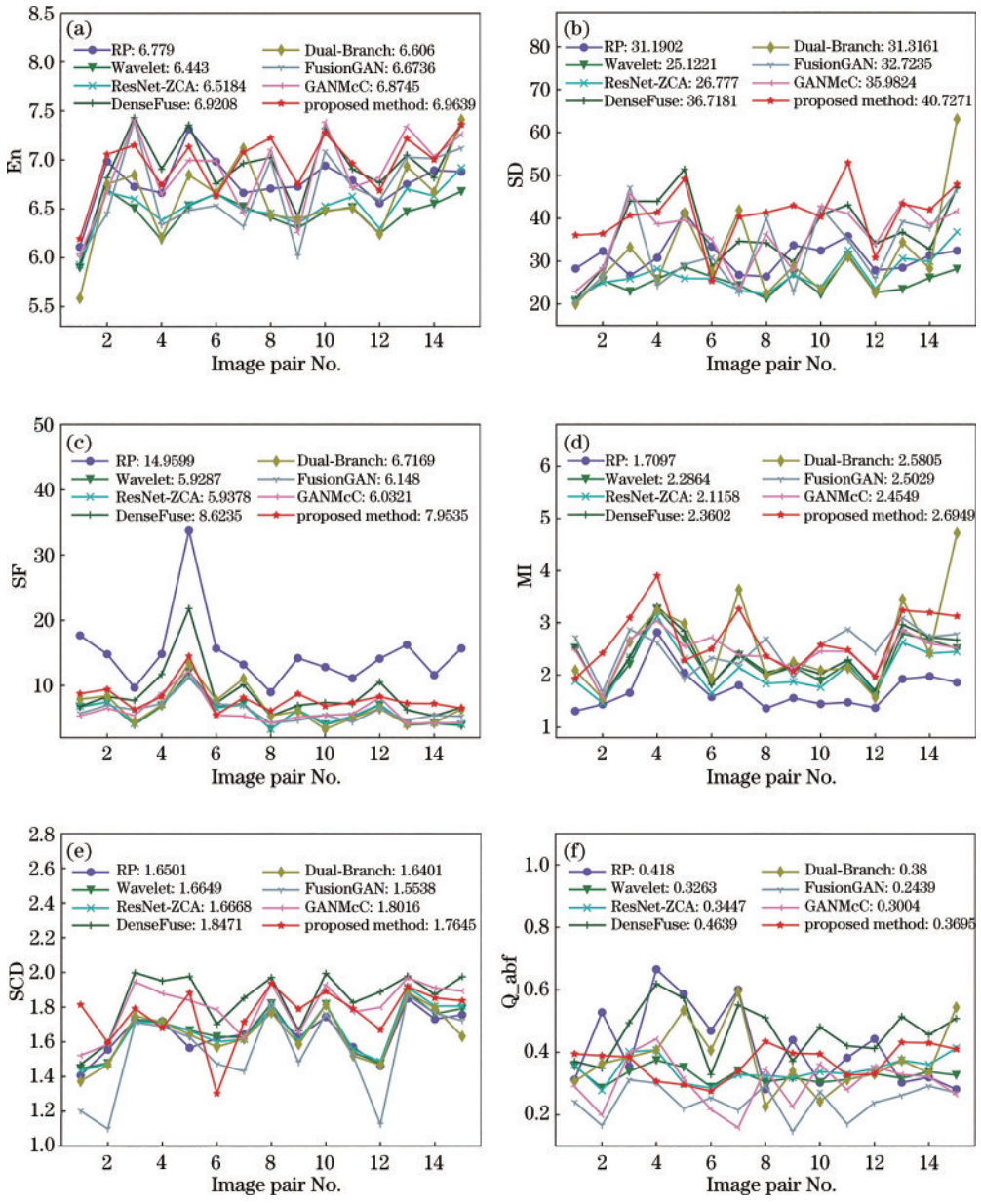

图 8　不同融合模型在 TNO 数据集上的 6 项客观指标。(a)En；(b)SD；(c)SF；(d)MI；(e)SCD；(f)Q_abf

Fig. 8　Six objective metrics of different fusion models on TNO dataset. (a) En; (b) SD; (c) SF; (d) MI; (e) SCD; (f) Q_abf

从图 8 可以看出，相较于其他 7 种融合方法，所提方法的信息熵(En)和标准差(SD)这两个指标均处于最优值，空间频率(SF)指标仅次于 RP 和 DenseFuse 的结果，表现了所提模型的有效性。所提模型结合了深度神经网络与 Transformer 网络，因而特征提取和图像表示表现更优秀，所提方法得到的融合图像所含信息量较多，灰度级分布较大，综合视觉效果更好。

此外，所提模型的互信息(MI)、差异相关性之和(SCD)和新型客观指标 Q_abf 分别处于第 1、第 3 和第 4 的位置，表明所提模型更好地保留了红外图像中的显著目标信息和可见光图像中纹理细节等信息，这点与融合图像的主观结果相一致。FusionGAN 和 GANMcC 等基于生成对抗网络的方法的 Q_abf 和 SCD 指标分别优于所提模型，这是由于这两个融合模型对红外图像的拟合能力足够高，但这也导致这类模型对可见光图像的表示能力不足，这点也可以从主观图像中看出。所提模型的 MI 达最优值，表明相较于 ResNet-ZCA 和 DenseFuse 等需要中间融合过程的方法，所提融合策略能够从源图像提取的特征中保留更多的有效信息。

4.4　消融实验及模型有效性分析

在所提模型的基础上设计了 3 种模型，以开展消融实验验证模型各组成部分的有效性，训练集和测试集与 3.1 节中设定的数据集保持一致，部分融合结果如图 9 所示。

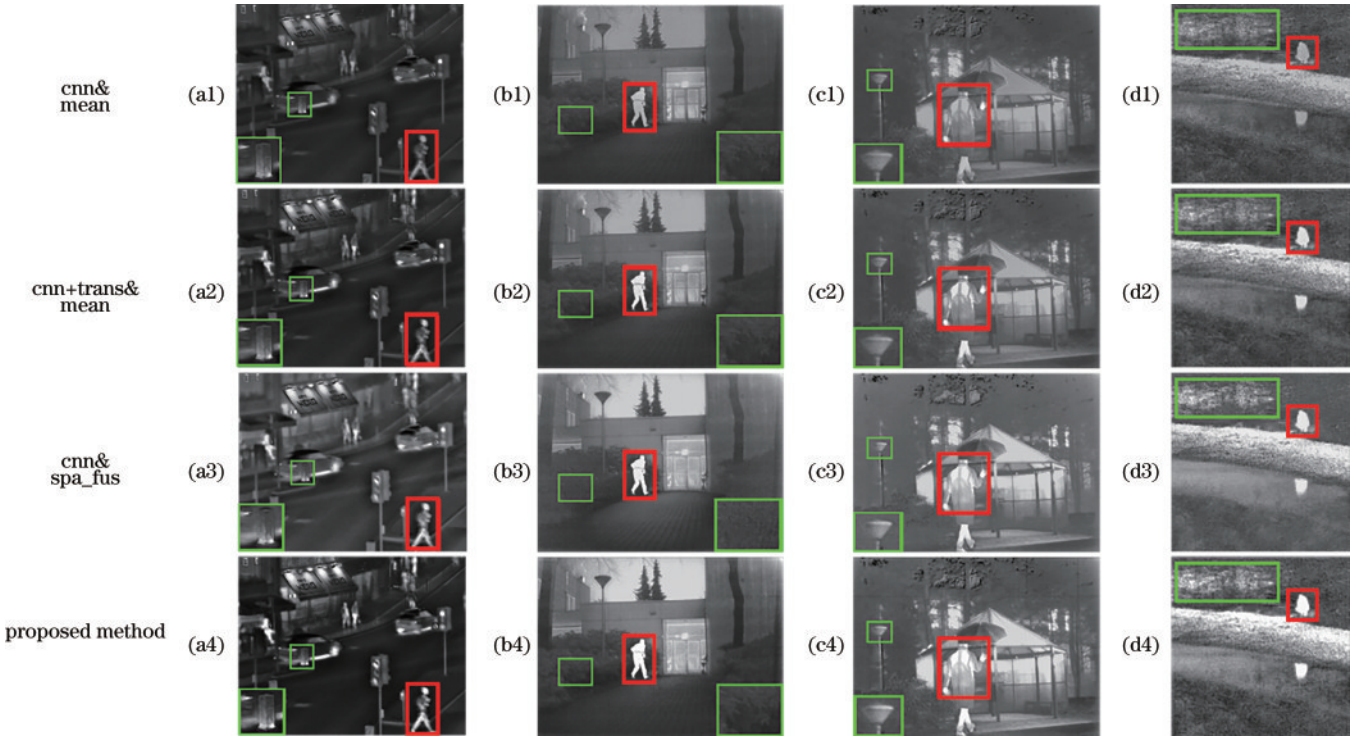

图 9　消融实验的主观结果

Fig. 9　Subjective results of the ablation experiment

模型 A(cnn&mean):只使用 CNN 提取特征并且选择对特征图取平均操作的融合方式。该模型验证了所提 Transformer 及 CTFusion 模块和所提融合策略协同的作用。

模型 B(cnn+trans&mean):同时使用 CNN 和 Transformer 进行增强式的特征提取并且选择求平均的融合方式。该模型验证了所提 Transformer 和 CTFusion 模块的作用。

模型 C(cnn &spa_fus):只使用 CNN 进行特征提取并且选择所提融合策略的融合方式。该模型验证了所提融合策略的作用。

根据图 9 的结果可以看出:模型 A 融合结果的对比度整体偏低,热辐射目标不明显,树枝、灌木丛、墙壁等纹理细节虽然能够保存下来,但边缘较为模糊,整体视觉效果较差。这表明 CNN 模块虽然能够一定程度地提取到红外图像中热辐射目标的亮度信息和可见光图像中树木等局部纹理信息,但是提取能力有限,无法将源图像中关键信息完整传输到融合图像中。相较于变种模型 A 仅使用 CNN 进行特征提取的情况,模型 B 融合结果中热辐射目标较为显著,而且背景纹理也有着一定程度的锐化,这是因为 Transformer 能够关注到图像的上下文,其特有的自注意机制能够强化 CNN 对红外图像中热辐射目标的特征提取和可见光图像中背景纹理的局部细节提取,进而增强对源图像的理解,提升融合图像的整体质量。与模型 A 相比,模型 C 使用了所设计的融合策略,从结果可以看出,由于考虑到了红外与可见光图像同一区域的模态信息最大差异度,因此能够自适应地依据不同模态图像的显著区域来进行融合,最终增强了融合图像中整体的对比度,如热辐射目标等。然而某些局部区域的边缘纹理较为模糊,如图 9(a3)、(b3)、(c3)中细线方框所示,灯柱、灌木丛和路灯等部分存在边缘信息丢失的情况。通过在特征提取部分加上 Transformer 可以解决这一问题,如图 9 第 4 行图片所示,所提模型既能显著突出来自红外图像的热辐射目标,又能完整保留来自可见光图像中丰富的背景纹理等细节信息,图像整体较为自然,更加符合人类的视觉感知。这表明利用 CNN 和 Transformer 优化特征提取过程,能够使网络更好地关注图像的局部和全局信息,增强源图像的特征表达,所设计的融合策略能够显著提升图像对比度,使融合图像保留源图像中更丰富的信息。同时所提特征提取网络和图像融合策略能够生成更符合人类视觉感知且含有源图像中丰富特征的融合图像。

4 种模型的客观指标如表 1 所示,可以看出,无论在模型 A 的设定基础上添加 Transformer 来强化有效特征提取过程,或者采用所提出的融合策略提升融合过程中的信息保留程度,都能够显著提高 6 种客观指标。所提模型的 En、SD、SF 和 SCD 均取得了最优值,这表明同时使用 Transformer 和所提融合策略能够使融合图像包含更多的信息和更大的灰度级分布,空间频率指标(SF)和差异相关性之和(SCD)高,意味着本融合图像更符合人眼的视觉特性,这点与得到的主观结果相一致。所提模型的 MI 和 Q_abf 指标取得次优,仅次于模型 C,表明了融合图像中含有大量来自源图

表 1 消融实验的客观结果平均值

Table 1 Average value of objective results of the ablation experiment

Model	En	SD	SF	MI	SCD	Q_abf
cnn&mean	6.5546	28.2087	6.0295	2.2380	1.7105	0.3059
cnn+trans&mean	6.6574	31.7763	6.5209	2.2648	1.7249	0.3110
cnn&spa_fus	6.7775	35.1202	7.4665	3.2954	1.6879	0.4115
Proposed method	6.9639	40.7271	7.9535	2.6949	1.7645	0.3695

像中的显著特征和细节信息，这表明所提基于模态最大差异度的融合策略能够对源图像中的显著特征信息进行有效保留，提升了融合图像的整体质量。

5 结 论

提出了一种联合CNN和Transformer的红外可见光图像融合网络模型。该模型采用了自动编码器网络结构，通过并行的方式同时利用CNN和Transformer进行特征提取，解决了现有融合模型仅利用CNN无法建模源图像内部的全局语义相关性以及对源图像上下文信息利用不充分等问题。此外，基于红外和可见光特征图之间的最大差异度，设计了一种新的融合策略，根据二者特征图中像素的重要性对其进行权重分配，实现对源图像中的显著特征进行自适应保留的目的。而后，在TNO公开数据集中进行了充分的对比和消融实验，结果表明在主观评价和客观指标上，所提模型均优于现有的图像融合模型。同时所提模型在Transformer上的尝试，也拓展了基于深度学习的图像融合领域的发展思路。

参 考 文 献

[1] 何自芬，陈光晨，陈俊松，等. 多尺度特征融合轻量化夜间红外行人实时检测[J]. 中国激光，2022，49(17)：1709002.
He Z F, Chen G C, Chen J S, et al. Multi-scale feature fusion lightweight real-time infrared pedestrian detection at night[J]. Chinese Journal of Lasers, 2022, 49(17): 1709002.

[2] 李畅，杨德东，宋鹏，等. 基于全局感知孪生网络的红外目标跟踪[J]. 光学学报，2021，41(6)：0615002.
Li C, Yang D D, Song P, et al. Global-aware siamese network for thermal infrared object tracking[J]. Acta Optica Sinica, 2021, 41(6): 0615002.

[3] 冯玉芳，殷宏，卢厚清，等. 基于改进全卷积神经网络的红外与可见光图像融合方法[J]. 计算机工程，2020，46(8)：243-249，257.
Feng Y F, Yin H, Lu H Q, et al. Infrared and visible light image fusion method based on improved fully convolutional neural network[J]. Computer Engineering, 2020, 46(8): 243-249, 257.

[4] 陈潮起，孟祥超，邵枫，等. 一种基于多尺度低秩分解的红外与可见光图像融合方法[J]. 光学学报，2020，40(11)：1110001.
Chen C Q, Meng X C, Shao F, et al. Infrared and visible image fusion method based on multiscale low-rank decomposition[J]. Acta Optica Sinica, 2020, 40(11): 1110001.

[5] 唐超影，浦世亮，叶鹏钊，等. 基于卷积神经网络的低照度可见光与近红外图像融合[J]. 光学学报，2020，40(16)：1610001.
Tang C Y, Pu S L, Ye P Z, et al. Fusion of low-illuminance visible and near-infrared images based on convolutional neural networks[J]. Acta Optica Sinica, 2020, 40(16): 1610001.

[6] Liu Y, Chen X, Wang Z F, et al. Deep learning for pixel-level image fusion: recent advances and future prospects[J]. Information Fusion, 2018, 42: 158-173.

[7] Burt P, Adelson E. The Laplacian pyramid as a compact image code[J]. IEEE Transactions on Communications, 1983, 31(4): 532-540.

[8] Toet A. Hierarchical image fusion[J]. Machine Vision and Applications, 1990, 3(1): 1-11.

[9] Nencini F, Garzelli A, Baronti S, et al. Remote sensing image fusion using the curvelet transform[J]. Information Fusion, 2007, 8(2): 143-156.

[10] Lewis J J, O'Callaghan R J, Nikolov S G, et al. Pixel- and region-based image fusion with complex wavelets[J]. Information Fusion, 2007, 8(2): 119-130.

[11] Zhao C H, Guo Y T, Wang Y L. A fast fusion scheme for infrared and visible light images in NSCT domain[J]. Infrared Physics & Technology, 2015, 72: 266-275.

[12] Naidu V P S, Raol J R. Pixel-level image fusion using wavelets and principal component analysis[J]. Defence Science Journal, 2008, 58(3): 338-352.

[13] Liu Y, Chen X, Ward R K, et al. Image fusion with convolutional sparse representation[J]. IEEE Signal Processing Letters, 2016, 23(12): 1882-1886.

[14] Ma J L, Zhou Z Q, Wang B, et al. Infrared and visible image fusion based on visual saliency map and weighted least square optimization[J]. Infrared Physics & Technology, 2017, 82: 8-17.

[15] 傅志中，王雪，李晓峰，等. 基于视觉显著性和NSCT的红外与可见光图像融合[J]. 电子科技大学学报，2017，46(2)：357-362.
Fu Z Z, Wang X, Li X F, et al. Infrared and visible image fusion based on visual saliency and NSCT[J]. Journal of University of Electronic Science and Technology of China, 2017, 46(2): 357-362.

[16] Huang W, Jing Z L. Evaluation of focus measures in multi-focus image fusion[J]. Pattern Recognition Letters, 2007, 28(4): 493-500.

[17] 江泽涛，何玉婷. 基于卷积自编码器和残差块的红外与可见光图像融合方法[J]. 光学学报，2019，39(10)：

1015001.
Jiang Z T, He Y T. Infrared and visible image fusion method based on convolutional auto-encoder and residual block[J]. Acta Optica Sinica, 2019, 39(10): 1015001.

[18] 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.

[19] Li H, Wu X J, Durrani T. NestFuse: an infrared and visible image fusion architecture based on nest connection and spatial/channel attention models[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 69(12): 9645-9656.

[20] Ma J Y, Yu W, Liang P W, et al. FusionGAN: a generative adversarial network for infrared and visible image fusion[J]. Information Fusion, 2019, 48: 11-26.

[21] Xu H, Liang P W, Yu W, et al. Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence, August 10-16, 2019, Macao, China. New York: ACM Press, 2019: 3954-3960.

[22] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4-9, 2017, Long Beach, CA, USA. New York: ACM Press, 2017: 6000-6010.

[23] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB/OL]. (2020-10-22)[2022-06-05]. https://arxiv.org/abs/2010.11929.

[24] Xu H, Ma J Y, Jiang J J, et al. U2Fusion: a unified unsupervised image fusion network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(1): 502-518.

[25] 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.

[26] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7132-7141.

[27] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[28] Miyato T, Kataoka T, Koyama M, et al. Spectral normalization for generative adversarial networks[EB/OL]. (2018-02-16)[2022-05-06]. https://arxiv.org/abs/1802.05957.

[29] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2021: 9992-10002.

[30] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[31] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.

[32] Toet A. TNO image fusion dataset[DB/OL]. (2014-04-26)[2020-04-20]. https://figshare.com/articles/TN_Image_Fusion_Dataset/1008029.

[33] Zuo Y J, Liu J H, Bai G B, et al. Airborne infrared and visible image fusion combined with region segmentation[J]. Sensors, 2017, 17(5): 1127.

[34] Toet A. Image fusion by a ratio of low-pass pyramid[J]. Pattern Recognition Letters, 1989, 9(4): 245-253.

[35] Qu G H, Zhang D L, Yan P F, et al. Medical image fusion by wavelet transform modulus maxima[J]. Optics Express, 2001, 9(4): 184-190.

[36] Li H, Wu X J, Durrani T S. Infrared and visible image fusion with ResNet and zero-phase component analysis[J]. Infrared Physics & Technology, 2019, 102: 103039.

[37] Fu Y, Wu X J. A dual-branch network for infrared and visible image fusion[C]//2020 25th International Conference on Pattern Recognition (ICPR), January 10-15, 2021, Milan, Italy. New York: IEEE Press, 2021: 10675-10680.

[38] Ma J Y, Zhang H, Shao Z F, et al. GANMcC: a generative adversarial network with multiclassification constraints for infrared and visible image fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 5005014.

[39] Liu Z, Blasch E, Xue Z Y, et al. Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(1): 94-109.

[40] 胡良梅, 高隽, 何柯峰. 图像融合质量评价方法的研究[J]. 电子学报, 2004, 32(S1): 218-221.
Hu L M, Gao J, He K F. Research on quality measures for image fusion[J]. Acta Electronica Sinica, 2004, 32(S1): 218-221.

[41] Aslantas V, Bendes E. A new image quality metric for image fusion: the sum of the correlations of differences[J]. AEU-International Journal of Electronics and Communications, 2015, 69(12): 1890-1896.

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