DOI: 10.13382/j.jemi.B2407761

用于红外与微光图像融合的目标差分注意力和 Transformer 算法*

陈广秋 代宇航 段 锦 黄丹丹
(长春理工大学电子信息工程学院 长春 130022)

摘 要:针对当前红外与微光图像融合算法中易出现光谱信息缺失、目标边缘模糊等问题,提出了用于红外与微光图像融合的目标差分注意力和 Transformer 的融合算法。首先,利用残差结构构造一种微光重构网络,并利用 VGG-16 构建感知损失,最大程度保留微光图像中的背景纹理信息和亮度信息;而后,将卷积神经网络(CNN)与 Transformer 结合构建特征提取网络,提取图像的完整特征;同时,在目标差分注意力模块中,对红外图像和微光图像进行差分运算和特征提取,得到的红外差分图像通过通道注意力机制对目标特征进行增强,再与 CNN 特征提取网络的输出特征图进行逐元素相加,增强红外目标特征;然后,通过纹理保留模块捕捉特征的高频信息和低频信息,提升纹理细节的保留度;最后,利用特征重建网络重构出融合图像。实验结果表明,融合结果不仅更符合人眼视觉系统,在客观评价指标中 MI 和 VIF 分别比其他融合方法提升了 44.6%和 21.2%。

关键词: 深度学习;图像融合;红外图像;微光图像;CNN;Transformer
中图分类号: TN919.81;TP391.4 **文献标识码**: A **国家标准学科分类代码**: 520.20

Target differential attention and Transformer algorithm for infrared and low-light image fusion

Chen Guangqiu Dai Yuhang Duan Jin Huang Dandan
(School of Electronics and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China)

Abstract: Aiming at the problems of spectral information missing and target edge blurring in current infrared and low light level image fusion algorithms, a target difference attention algorithm and Transformer fusion algorithm for infrared and low light level image fusion are proposed. Firstly, a low-light level reconstruction network is constructed by using residual structure, and the perception loss is constructed by using VGG-16 to preserve the background texture and brightness information in the low-light level image to the maximum extent. Then, the feature extraction network is constructed by combining CNN and Transformer to extract the complete features of the image. At the same time, in the target differential attention module, the difference operation and feature extraction are carried out on the infrared image and low-light image, and the obtained infrared differential image is enhanced by the channel attention mechanism. Then the output feature map of CNN feature extraction network is added element by element to enhance the infrared target feature. Then, the high frequency and low frequency information of features are captured by gradient retention module to improve the retention of texture details. Finally, the feature reconstruction network is used to reconstruct the fused image. The experimental results show that the fusion results are not only more consistent with the human visual system, but also the objective evaluation indexes of MI and VIF are increased by 44.6% and 21.2%, respectively, compared with other fusion methods.

Keywords: deep learning; image fusion; infrared image; low-light image; CNN; Ttransformer

0 引 言

红外(infrared,IR)图像是由红外传感器通过接收地物反射或自身发散的红外辐射而得到的图像,并且外界光照条件对成像影响小,抗干扰能力强,但对热辐射信号外的场景亮度变化敏感性较差。微光(low-light,LL)图像可见性较红外图像强,通过目标场景的反射特性实现

收稿日期: 2024-08-13 Received Date: 2024-08-13
*基金项目:国家自然科学基金重大仪器专项(62127813)资助

成像,包含较为丰富的纹理信息和背景结构,但极容易受到环境光线影响。因此将具有空间和信息互补性的红外与微光图像融合应用非常广泛,对于难以辨析的空间场景具有重要应用价值。

近些年来,基于深度学习的图像融合方法发展迅速,因其具有很强的特征提取能力而得到了广泛的应用,其方法主要可分为 3 种,卷积神经网络(CNN)、自编码器网络和生成对抗网络。2018 年,Li 等[1]提出一种基于 DenseBlock 的红外和可见光图像融合算法(DenseFuse),DenseFuse 的编码器和解码器中的每层都使用了密集连接的网络架构,又利用加法策略和 l1 范数策略对红外特征和可见光特征进行融合。2019 年 Ma 等[2]首次将生成对抗网络引入图像融合领域(FusionGAN),该网络由生成器和鉴别器构成,利用两者的对抗性,源图像输入到生成器进行融合操作,而鉴别器作为生成器的“对手”,迫使生成器可以生成具有更多源图像细节信息的融合图像。2020 年,Xu 等[3]提出了一种通用的图像融合框架,名为 U2Fusion,使用 VGG-16 网络提取 5 个层次的特征图,随后向已训练的 DenseNet 网络输入特征,生成融合图像。同年 Li 等[4]将多尺度分解图像的思想融入网络,设计了基于巢连接的深度网络。2021 年,Zhang 等[5]引入压缩和分解网络,提出了 SDNet 用于实时图像融合的多功能压缩分解网络。2022 年,Tang 等[6]提出了一种基于照明感知来确定损失函数权重的算法(PIAFusion),网络主要利用光照感知网络判断白天概率和夜晚概率,并以这个概率作为损失函数的权重,但模型过于简单,无法在复杂环境中调整光照。2023 年,Tang 等[7]提出了一种基于视觉增强的红外和可见光图像融合框架(DIVFusion),并在其中设计了一种场景亮度解纠缠网络,用于提升两种模态的独特特征。虽然基于深度学习的融合方法弥补了传统融合方法的某些不足,但仍然存在细节信息丢失和红外目标不明显等问题。

为了解决这些问题,本文提出了用于红外与微光图像融合的目标差分注意力和 Transformer 的融合算法。首先,通过残差结构构造出微光重构网络,以保留微光图像中的背景纹理信息和亮度信息;而后,将 CNN 与 Transformer 结合构建特征提取网络,提取图像的完整特征;同时,在目标差分注意力模块中,对红外图像和微光图像进行差分运算和特征提取,再与 CNN 特征提取网络的输出特征图进行逐元素相加,然后,通过纹理保留模块提升纹理细节的保留度;最后,利用特征重建网络获得融合图像。实验结果表明,较已有文献的融合算法,本融合算法生成的融合图像更好地保留微光图像中的背景细节信息且红外目标显著,并在主观评价和客观评价方面均取得了较好的融合性能。

1 本文算法

1.1 TDATFuse 整体架构

本文所提出的红外与微光图像融合算法(TDATFuse)主要由微光重构网络(low-light reconstruction network,LLRNet)、CNN 特征提取网络(CNN feature extraction network,CFENet)、Transformer 特征提取网络(Transformer feature extraction network,TFENet)、目标差分注意力模块(target differential attention block,TDABlock)、纹理保留模块(texture preservation block,TPBlock)和特征重建网络(feature reconstruction network,FRNet)构成。TDATFuse 的整体架构如图 1 所示。

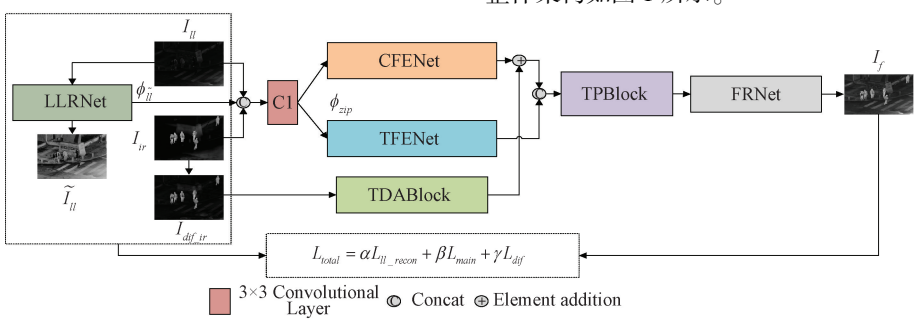


图 1 TDATFuse 的整体网络架构

Fig. 1 TDATFuse's overall network architecture

1.2 微光重构网络

由于微光图像能见度低,本文设计了微光重构网络重构微光图像,通过直方图均衡化后的图像 I_{hist} 指导微光重构图像 $\widetilde{I}_{ll}$ 的生成,并且重构后的特征图 $\varphi_{\widetilde{I}_{ll}}$ 直接参与特征提取。LLRNet 主要由残差块、卷积层和激活函数构成。微光图像 I_{ll} 先经过 1×1 的卷积增加特征通道数,随即进入 4 个残差块获取深层特征,随后经过卷积层得到微光重构图像 $\widetilde{I}_{ll}$。LLRNet 的整体架构如图 2 所示。

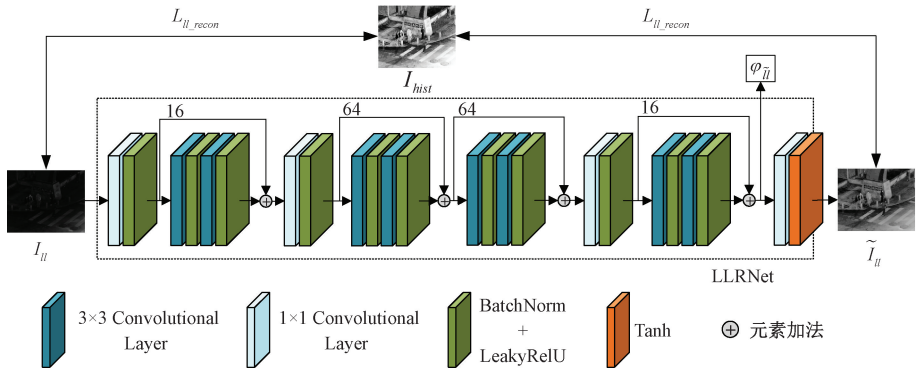


图 2　LLRNet 的整体架构

Fig. 2　LLRNet's overall architecture

1.3　CNN 特征提取网络

本文设计了 CNN 特征提取网络，其主要由 5 个 3×3 的卷积层和 LeakyRelu 激活函数组成。该网络对各层使用密集连接，经过公共卷积层的压缩特征 φ_{zip} 进入 CFENet，随后获得局部特征 φ_{part}。CFENet 的整体架构如图 3 所示。

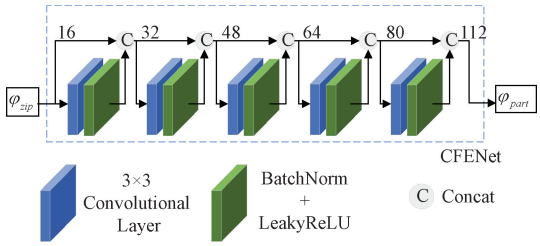


图 3　CFENet 的整体架构

Fig. 3　CFENet's overall architecture

1.4　Transformer 特征提取网络

TFENet 由 3 个 Transformer 块构成，输入为压缩特征 φ_{zip}，经过 TFENet 后得到全局特征 φ_{glob}。Transformer 模型通过自注意力机制有效建模长距离的依赖关系，使得网络高效捕捉到图像中不同区域的语义关系。TFENet 的整体架构如图 4 所示。

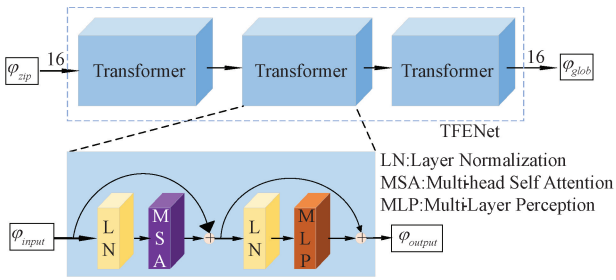


图 4　TFENet 的整体架构

Fig. 4　TFENet's overall architecture

每个输入到 Transformer 模块的特征 φ_{input} 被分割成不重叠 $M\times M$ 大小的窗口，从而生成 $(HW/M^2)\cdot M^2\cdot C$ 个特征图，其中 HW/M^2 为窗口总数。所有窗口都计算了自注意力机制。对于输入到 Transformer 模块的特征 $\varphi_{input}\in \boldsymbol{R}^{M^2\cdot C}$ 可以与 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 相对应，计算公式如式(1)所示。

$$\begin{aligned}\boldsymbol{Q}&=\varphi_{input}\cdot\boldsymbol{P}_Q\\ \boldsymbol{K}&=\varphi_{input}\cdot\boldsymbol{P}_K\\ \boldsymbol{V}&=\varphi_{input}\cdot\boldsymbol{P}_V\end{aligned}\tag{1}$$

式中：$\boldsymbol{P}_Q$、$\boldsymbol{P}_K$ 和 $\boldsymbol{P}_V$ 分别表示投影矩阵。窗口中的注意力度量通过自注意力方案计算并公式如式(2)所示。

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=S(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{d}+B)\boldsymbol{V}\tag{2}$$

式中：$S(\cdot)$ 表示 Softmax 操作；d 表示维度；B 表示可学习的相对位置编码。MLP 拥有两个 LN 层。

1.5　目标差分注意力模块

目标差分注意力模块如图 5 所示。

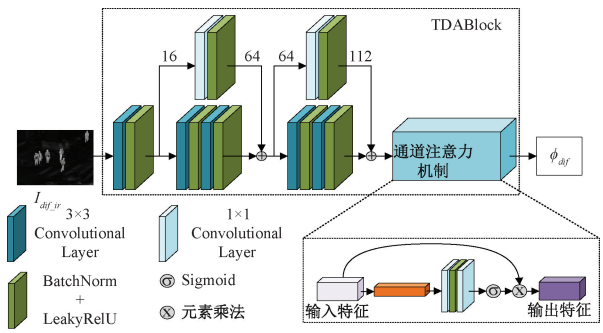


图 5　TDABlock 的整体架构

Fig. 5　TDABlock's overall architecture

通过对红外与微光图像进行差分操作得到红外差分图像，红外差分图像 I_{dif_ir} 经过卷积层和残差块得到深度特征，最后经过通道注意力机制得到最终的差分增强特征 φ_{dif}。

1.6　纹理保留模块

纹理保留模块用来增强特征的背景细节信息，φ_{f_cat} 为 TPBlock 的输入特征，表达式如式(3)所示。

$$\varphi_{f\_cat} = Concat((\varphi_{part} \oplus \varphi_{dif}), \varphi_{glob}) \tag{3}$$

式中：φ_{part} 为 CFENet 输出局部特征；φ_{dif} 为 TDABlock 输出的差分增强特征；φ_{glob} 为 TFENet 输出的全局特征；⊕表示元素相加；$Concat(\cdot)$ 表示级联操作。

TPBlock 中，利用 Laplacian 算子和 Sobel 算子得到梯度等信息，随即经过卷积和通道级联，以获取具有细粒度细节的深层特征 φ_{tex}。TPBlock 整体架构如图 6 所示。

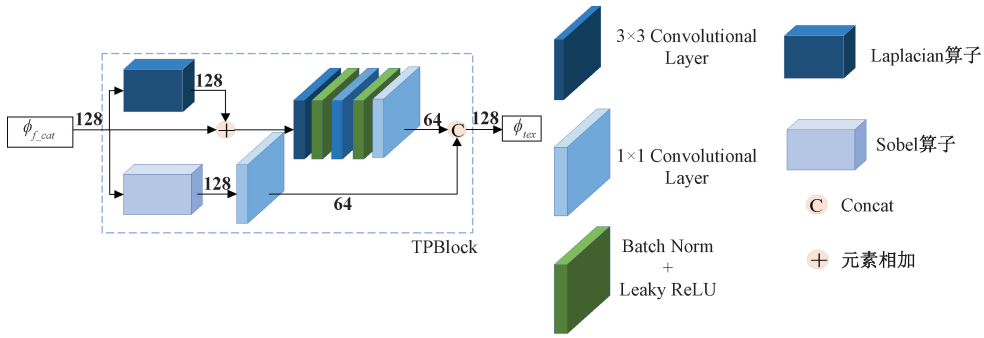


图 6　TPBlock 整体架构

Fig. 6　TPBlock's overall architecture

1.7　特征重建网络

特征重建网络 FRNet 的整体架构如图 7 所示。特征 φ_{tex} 进入特征重建网络，最终得到融合图像 I_f。

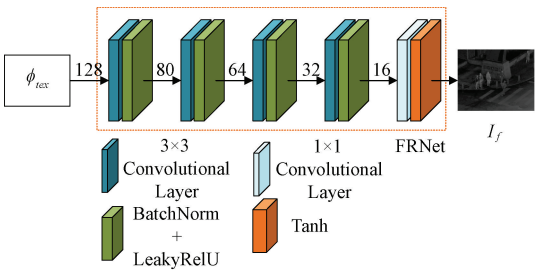


图 7　FRNet 的整体架构

Fig. 7　FRNet's overall architecture

2　实验训练及结果分析

在网络模型训练过程中，本文算法使用 PyTorch1.10.2 作为编程环境，在 NVIDIA RTX3090 GPU 和 Intel(R) Xeon(R) Platinum 8255C CPU @2.50 GHz 上进行实验。

在数据集方面，本文选取了 LLVIP[8] 和 KAIST[9] 数据集中训练集的 12 032 张红外和夜晚图像。首先，将训练图像转换为灰度图像，并将图像的尺寸重塑为 120×120。然后，将训练数据划分为大小为 32 的批次(batch_size=32)并设置 40 个 epoch 来进行训练。同时使用 Adam 优化器，将初始学习率设置为 1×10^{-2}。并引入学习率衰减策略，使学习率在每 10 个 epoch 后衰减为上一轮的 0.1 倍。在训练过程中学习率会根据训练的进展逐渐减小，以帮助网络稳定收敛到最优解。

在损失函数中，α 设置为 0.8，β 设置为 1，γ 设置为 0.8；α_1 设置为 1，α_2 设置为 1，α_3 设置为 10；β_1 设置为 100，β_2 设置为 0.42，β_3 设置为 0.7；γ_1 设置为 0.2，γ_2 设置为 0.2。

2.1　损失函数

在网络训练中本文设计了一种总损失函数 L_{total}，如式(4)所示。

$$L_{total} = \alpha L_{ll\_recon} + \beta L_{main} + \gamma L_{dif} \tag{4}$$

式中：L_{ll_recon} 表示微光重构损失函数；L_{main} 表示主干网络损失函数；L_{dif} 表示差分增强损失函数；α、β 和 γ 表示调节各个损失函数的参数。

1) 微光重构损失

直方图均衡化后的微光图像可以弥补微光图像背景信息不足的缺点，所以本文设计了微光重构损失函数 L_{ll_recon}，如式(5)所示。

$$L_{ll\_recon} = \alpha_1 L_{int} + \alpha_2 L_{struct} + \alpha_3 L_{per} \tag{5}$$

式中：L_{int} 表示强度损失；L_{struct} 表示结构损失；L_{per} 表示感知损失。

强度损失可以驱动微光重构网络尽可能多地使原始图像相似于直方图均衡化后的微光增强图像，强度损失 L_{int} 如式(6)所示。

$$L_{int} = \| \widetilde{I}_{ll} - I_{hist} \|_1 \tag{6}$$

式中：$\| \cdot \|_1$ 表示 l_1 范数；$\widetilde{I}_{ll}$ 表示微光重构图像；I_{hist} 表示微光增强图像。

此外本文使用结构相似度作为结构损失，结构损失 L_{struct} 如式(7)所示。

$$L_{struct} = 1 - SSIM(\widetilde{I}_{ll}, I_{hist}) \tag{7}$$

式中：$SSIM(\cdot)$ 表示结构相似度计算公式；$\widetilde{I}_{ll}$ 表示微光

重构图像；I_{hist} 表示微光增强图像。

在微光重构损失中，采用 VGG-16 网络提取特征构建感知损失，感知损失 L_{per} 如式(8)所示。

$$L_{per} = \| VGG(\widetilde{I_{ll}}) - VGG(I_{hist}) \|_2^2 \tag{8}$$

式中：$\| \cdot \|_2$ 表示 l_2 范数；$VGG(\cdot)$ 表示预训练的 VGG-16 模型；$\widetilde{I_{ll}}$ 为微光重构图像；I_{hist} 为微光增强图像。为了有效地提高后续高级视觉任务的性能，感知损失只选择 VGG-16 网络的最后两层来实现深度特征提取。

2) 主干损失

为使融合图像保留一定的原图像特征，本文设计了主干网络损失函数 L_{main}，具体如式(9)所示。

$$L_{main} = L_{SSIM} + L_{recon} + \beta_1 L_{gra} \tag{9}$$

式中：L_{SSIM} 表示结构相似度损失；L_{recon} 表示重建损失；L_{gra} 表示梯度损失。

此外本文设计了平衡融合图像亮度的结构相似度损失，使图像中保留更多黑暗区域中的细节信息，结构相似度损失 L_{SSIM} 如式(10)所示。

$$L_{SSIM} = \beta_2(1 - SSIM(I_f, I_{ir})) + (1-\beta_2)(\beta_3(1 - SSIM(I_f, \widetilde{I_{ll}})) + (1-\beta_3)(1 - SSIM(I_f, I_{ll}))) \tag{10}$$

式中：I_f 表示融合图像；I_{ir} 表示红外图像；I_{ll} 表示微光图像；β_2 是调节红外图像与微光的比例参数；β_3 为调节微光图像与微光重构图像的比例参数。

本文基于平衡亮度的思想设计了重建损失函数，重建损失 L_{recon} 如式(11)所示。

$$L_{recon} = \beta_2 \| I_f - I_{ir} \|_1 + (1-\beta_2)(\beta_3 \| I_f - \widetilde{I_{ll}} \|_1 + (1-\beta_3) \| I_f - I_{ll} \|_1) \tag{11}$$

梯度损失 L_{gra} 如式(12)所示。

$$L_{gra} = \| \nabla I_f - \mathrm{Max}(\nabla I_{ir}, \nabla I_{ll}) \|_1 \tag{12}$$

式中：$\nabla(\cdot)$ 表示梯度算子；$\mathrm{Max}(\cdot)$ 表示最大值操作。

3) 差分增强损失

最后，本文设计了一种差分增强损失函数 L_{dif}，旨在引导差分注意力模块增强融合图像中的红外目标，差分增强损失函数 L_{dif} 如式(13)所示。

$$L_{dif} = \gamma_1 L_{dif\_SSIM} + \gamma_2 L_{dif\_int} \tag{13}$$

式中：L_{dif_SSIM} 表示差分结构相似度损失；L_{dif_int} 表示差分强度损失；γ_1 和 γ_2 为调节参数。差分结构相似度损失 L_{dif_SSIM} 如式(14)所示。

$$L_{dif\_SSIM} = (1 - SSIM(I_f, I_{dif\_ir})) \tag{14}$$

式中：I_{dif_ir} 表示红外差分图像。差分强度损失 L_{dif_int} 如式(15)所示。

$$L_{dif\_int} = \| I_f - I_{dif\_ir} \|_1 \tag{15}$$

2.2　结果分析

1) 测试数据集，本文算法在公开的红外与可见光数据集 LLVIP 与 M3FD[10] 上选取夜晚中的可见光图像作为微光图像和其对应的红外图像进行测试，分别包含 30 和 59 对红外图像与微光图像。

2) 对比实验与评价指标，为了比较所提出算法与最先进算法的融合性能，选取了 7 种具有代表性的算法与本文算法进行对比，包括 CVT[11]、Wavelet[12]、U2Fusion[3]、DenseFuse[4]、FusionGAN[2]、TarDAL[10] 和 IFCNN[13]。图 8 和 9 所示为各类算法在 LLVIP 数据集上的对比实验，图 10 和 11 所示为各类算法在 M3FD 数据集上的对比实验。其中 Wavele 和 CVT 是传统融合算法，U2Fusion、DenseFuse、FusionGAN、IFCNN 和 TarDAL 都属于深度学习融合算法，在对比实验中，本文借鉴了原始设置参数，并且使用评价指标客观评估融合结果，评价指标有信息熵(EN)[14]、空间频率(SF)[15]、标准差(SD)[16]、平均梯度(AG)[17]、基于伪影的指标 $N^{AB/F}$[18] 和相关系数(CC)[19]，表 1~4 分别为本文算法与其他算法在 LLVIP 数据集和 M3FD 数据集上的客观评价指标对比结果。其中除 $N^{AB/F}$ 外，其余指标均为分数越高融合效果越佳。表 1 中黑色粗体为最优值，下划线为次优值。

TDATFuse 与其他融合算法在 LLVIP 数据集上的第 1 组融合图像对比结果如图 8 所示。CVT 生成的融合图像中“人物”目标较为突出，图 8 中红框的“栅栏”少部分特征被保留，且红外目标的边缘及周围有不规则伪影，且图像清晰度较低。Wavelet 生成的融合图像中，红外目标亮度较低，图 8 中“斑马线”的整体特征与周围背景对比度低，边缘轮廓模糊。在 U2Fusion 和 IFCNN 生成的融合图像中，虽然保留了红外图像中关键特征，但融合结果仍缺乏重要细节信息。例如，在红框框选的“栅栏”区域，整体轮廓仍表征不清，总体上，这两种融合技术所生成的图像背景都表征不足，并且场景亮度都较为昏暗。DenseFuse 的融合图像中，虽然保留足够的红外特征和背景信息，但该算法在某些暗区域的纹理细节获取能力表现不佳，甚至还丢失了“斑马线”的轮廓细节，这些问题都对整体图像质量产生了较多的负面影响。通过对 FusionGAN 和 TarDAL 的结果进行观察，可以看出 FusionGAN 的融合图像整体较模糊，仅体现了部分红外目标的信息，且目标区域的亮度过于发散，与之相比，TarDAL 生成的图像能够明显地展示黄框框选的目标“人物”的位置和边界，但缺乏原始图像的背景纹理信息，除“人物”外的其他场景对象的目标特征均被丢失。本文算法在保证红外目标清晰度的前提下，与其他几种算法相比，TDATFuse 的融合图像在红外目标的清晰度上表现较好，同时保留了丰富的背景纹理特征，使得“斑马线”、“栅栏”等建筑物细节及边缘均被充分表达。尽管在明暗交界处对比度稍显不足，但目标对象的轮廓获得了明显表征，且整体图像的清晰度和亮度得到显著提升。

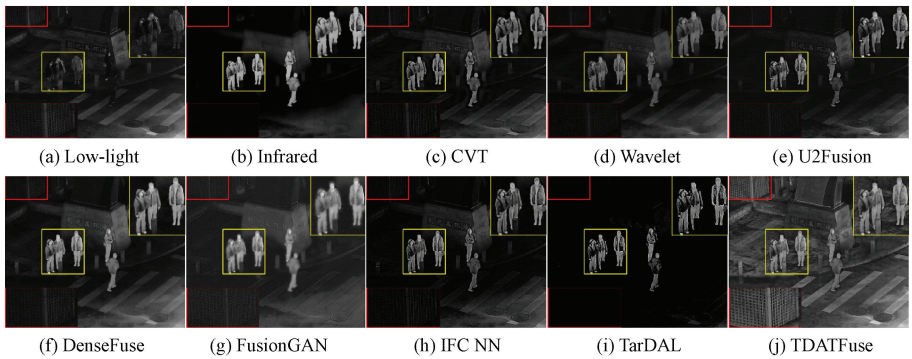


图 8　LLVIP 数据集的第 1 组融合图像对比结果

Fig. 8　The first set of fusion image comparison results of the LLVIP dataset

由表 1 可知,在指标 AG 中,TDATFuse 的融合图像获得较高分数,比 IFCNN 算法提升 1.997 5,这代表图像保留较多纹理信息。EN、SF 和 SD 这 3 项指标相比次优值分别提高 9.1%、40.5%和 8.8%。TDATFuse 整体仍优于大部分对比算法。

表 1　LLVIP 数据集的第 1 组融合图像客观评价指标对比结果

Table 1　Comparative results of the first set of objective evaluation indicators of fusion images in LLVIP dataset

算法	EN	SF	SD	AG	$N^{AB/F}$	CC
CVT	5.900 0	10.727 9	25.227 1	2.734 2	0.030 0	0.768 6
Wavelet	5.620 8	6.706 0	22.102 2	1.706 4	0.087 0	0.784 1
U2Fusion	4.984 4	9.082 3	22.342 2	2.041 7	0.037 4	0.763 4
DenseFuse	6.135 2	9.445 8	30.261 0	2.428 6	0.029 3	**0.786 3**
FusionGAN	5.749 2	7.401 7	25.388 2	2.070 3	0.086 4	0.691 1
IFCNN	5.667 2	11.221 6	25.232 7	2.862 0	**0.020 7**	0.772 9
TarDAL	1.463 9	10.759 9	23.022 9	2.124 9	0.042 4	0.457 9
TDATFuse	**6.715 4**	**15.765 3**	**32.936 8**	**4.859 5**	0.026 7	0.784 4

TDATFuse 与其他融合算法在 LLVIP 数据集的第 2 组融合图像对比结果如图 9 所示。首先,在图像左侧的背景较暗处,U2Fusion 和 TarDAL 的融合结果均缺少微光图像左侧的纹理细节,TarDAL 仅保留了小部分黄框“人物”特征,U2Fusion 比 TarDAL 保留了一些的细节信息,可视性略大于 TarDAL。IFCNN 的融合图像包含了更多的微光图像背景纹理特征,但是“栅栏”等建筑物边缘模糊,且左侧黑暗背景的细节信息仍有缺失。

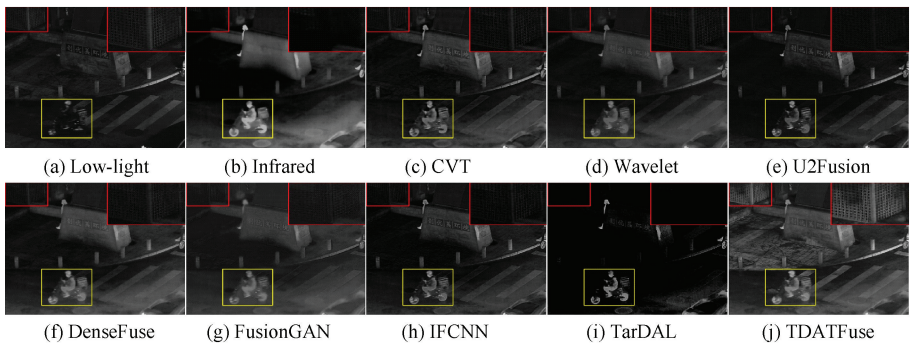


图 9　LLVIP 数据集的第 2 组图像融合结果

Fig. 9　The second set of fusion image comparison results of the LLVIP dataset

Wavelet 和 CVT 的融合结果保留了微光图像的大部分细节，CVT 的红外目标边缘更加锐利，而 Wavelet 的红外目标边缘比较模糊，从主观视觉看，CVT 的融合效果优于 Wavelet。FusionGAN 的融合图像仅保留少部分红外信息和背景纹理信息，除“人物”外，周围“栅栏”、“路面”等建筑物对象过于模糊，且伴随伪影出现，使得图像缺失了源图像的关键特征，融合效果较差。在 DenseFuse 的融合图像中可以清晰地观察到红外目标与周围背景对比度较高，融合效果在几种对比算法中最优，但红框的较暗背景处的部分细节不够突出，且建筑物的清晰度明显低于“人物”。

从 TDATFuse 的融合图像中可看出，图像增加“人物”清晰度的同时，保留了丰富的背景纹理信息，使得道路“斑马线”及周围建筑物的细节特征均被良好表达，且融合图像具有合理亮度。

由表 2 可知，TDATFuse 在 5 项指标中为最佳，对比于其他算法，评价指标 EN、SF、SD 和 CC 分别提升 3.9%、8.9%、6.9%和 2.7%，AG 提升最多，为 58.3%。

表 2　LLVIP 数据集的第 2 组融合图像客观评价指标对比结果

Table 2　Comparative results of the second set of objective evaluation indicators of fusion images in LLVIP dataset

算法	EN	SF	SD	AG	$N^{AB/F}$	CC
CVT	6.343 2	10.836 1	28.682 1	2.947 0	0.027 5	0.834 3
Wavelet	6.186 1	6.627 0	26.721 8	1.843 8	0.058 5	0.818 9
U2Fusion	5.895 0	9.592 3	29.309 4	2.489 2	0.025 7	0.823 4
DenseFuse	6.557 6	8.884 2	31.988 3	2.502 9	0.029 8	0.836 5
FusionGAN	6.305 7	8.187 2	27.540 5	2.203 7	0.056 9	0.758 1
IFCNN	6.290 6	11.126 6	28.775 2	3.015 9	**0.019 1**	0.836 7
TarDAL	2.953 7	12.383 2	25.574 5	2.926 7	0.040 0	0.582 9
TDATFuse	**6.813 8**	**13.487 8**	**34.213 6**	**3.942 3**	0.025 3	**0.859 6**

TDATFuse 与其他融合算法在 M3FD 数据集上的第 1 组对比结果如图 10 所示。CVT 的红外目标具有较强的对比度，但是车辆周围伪影严重，边缘细节不明显。Wavelet 突出了红外目标，但对比度低于 CVT，“大楼”等建筑物的场景背景信息不够丰富，车辆纹理信息丢失严重。U2Fusion 和 IFCNN 的融合图像中红外目标较为突出，但伪影问题没有被良好解决，图像上方的“天空”场景无法被真实呈现，且 IFCNN 融合结果在背景信息表达能力低于 U2Fusion。FusionGAN 算法保留了一定的红外目标特征，但丢失了过多的细节信息，使得车身细节和轮廓信息无法被精准表达。DenseFuse 的融合结果在突出红外目标的基础上过多地关注目标和背景的边缘信息，忽略了源图像的纹理细节。TarDAL 融合结果保留了更多的纹理信息，但是目标和背景的亮度表达程度失衡，导致部分场景过暗或过亮。

TDATFuse 的融合图像中，目标车辆与周围环境对比度合理，背景细节清晰且无明显的伪影出现。

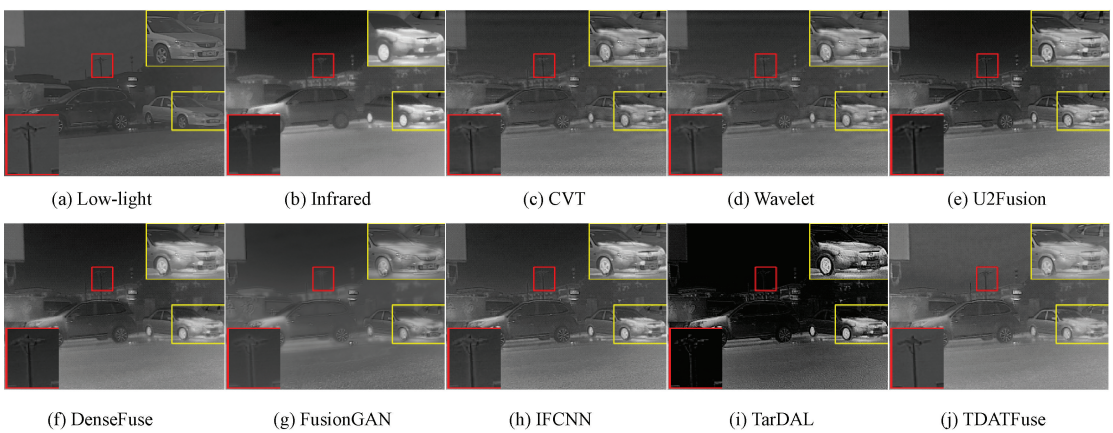


图 10　M3FD 数据集的第 1 组融合图像对比结果

Fig. 10　The first set of fusion image comparison results of the M3FD dataset

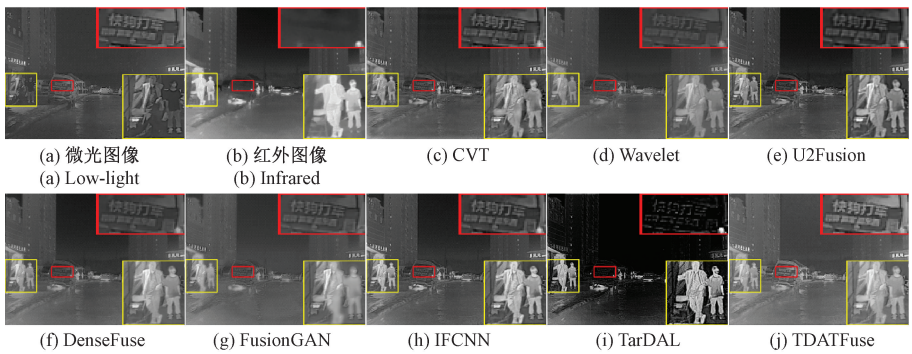
由表 3 可知，TDATFuse 在 $N^{AB/F}$ 外，其余评价指标均为最优，但仅与 IFCNN 算法相差 0.000 8，与排名第 2 的 TarDAL 算法相比，TDATFuse 在评价指标 SF 中提升了 17.0%。

表3　M3FD数据集的第1组融合图像客观评价指标对比结果

Table 3　Comparative results of the first set of objective evaluation indicators of fusion images in M3FD dataset

算法	EN	SF	SD	AG	$N^{AB/F}$	CC
CVT	6.833 6	11.795 6	30.646 6	3.139 2	0.020 0	0.847 8
Wavelet	6.687 1	6.492 1	28.707 1	1.738 2	0.059 3	0.862 8
U2Fusion	6.941 3	10.422 7	34.848 2	3.078 9	0.015 1	0.861 8
DenseFuse	6.865 6	7.622 5	33.526 4	2.070 5	0.039 8	0.864 8
FusionGAN	6.477 0	7.355 0	26.491 8	1.773 3	0.055 9	0.815 8
IFCNN	6.898 4	12.661 6	33.925 1	3.458 4	**0.013 6**	0.853 5
TarDAL	6.130 0	11.770 1	20.758 1	3.328 4	0.032 6	0.746 1
TDATFuse	**7.108 2**	**14.815 3**	**37.657 5**	**3.687 9**	0.014 4	**0.867 5**

TDATFuse与其他融合算法在M3FD数据集上的第2组对比结果如图11所示。CVT和Wavelet融合图像中红框框选的字体略显模糊。U2Fusion的融合结果中,红外目标与背景对比度较高,但是红框区域的“文字”亮度低。FusionGAN红框区域的“文字”融合效果优于DenseFuse,但黄框内的红外目标以及边缘信息模糊不清,且细节丢失严重。IFCNN的融合图像效果较好,但信息丰富程度不敌TDATFuse。TDATFuse的融合图像具有明显对比度,包含了丰富的红外目标信息,“人物”和“文字”清晰,融合图像更加符合人类视觉感官。



(a) 微光图像 (a) Low-light　(b) 红外图像 (b) Infrared　(c) CVT　(d) Wavelet　(e) U2Fusion

(f) DenseFuse　(g) FusionGAN　(h) IFCNN　(i) TarDAL　(j) TDATFuse

图11　M3FD数据集的第2组融合图像对比结果

Fig. 11　The second set of fusion image comparison results of the M3FD dataset

由表4可知,TDATFuse在EN、SF、SD指标中成绩最佳,AG与CC略低于IFCNN,但仍处于这几种算法中较为优秀的水平。

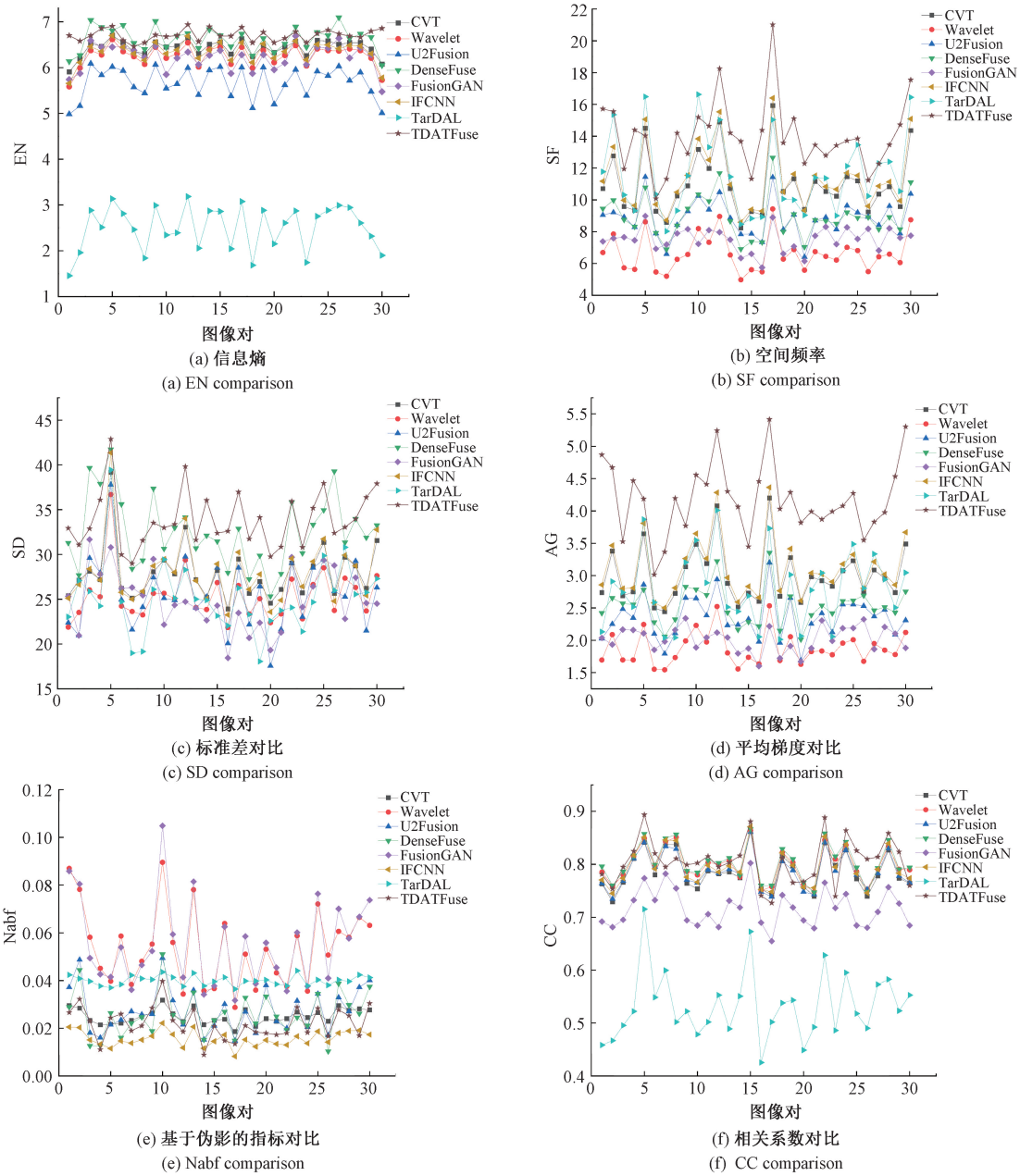
表4　M3FD数据集的第2组融合图像客观评价指标对比结果

Table 4　Comparative results of the second set of objective evaluation indicators of fusion images in M3FD dataset

算法	EN	SF	SD	AG	$N^{AB/F}$	CC
CVT	6.737 2	10.543 9	28.017 2	2.774 4	0.022 5	0.847 8
Wavelet	6.480 9	6.000 1	25.326 9	1.558 7	0.079 8	0.863 8
U2Fusion	6.768 2	9.902 0	31.212 7	2.821 8	0.018 1	0.861 8
DenseFuse	6.709 8	7.321 8	30.723 0	1.934 3	0.048 4	**0.864 8**
FusionGAN	6.505 7	7.508 2	27.875 7	1.692 3	0.078 3	0.815 8
IFCNN	6.907 7	11.184 6	32.792 0	**3.085 9**	**0.018 6**	0.853 5
TarDAL	6.064 1	9.877 6	23.934 2	2.013 3	0.042 6	0.746 1
TDATFuse	**6.917 0**	**12.129 3**	**33.237 5**	3.065 1	0.026 2	0.864 6

本文使用现有的7种算法与TDATFuse在数据集LLVIP和M3FD上进行客观评价指标对比实验。LLVIP数据集的客观评价指标对比实验折线图如图12所示。可以看出,TDATFuse的大部分融合结果的客观评价指标都处在一个较高的水平。

由表5可知,TDATFuse在数据集LLVIP中,EN、SF、SD、AG、CC五项指标均为最优值,分别比次优值提升了1.16%、20.64%、2.55%、34.20%、0.12%。



(a) 信息熵
(a) EN comparison
(b) 空间频率
(b) SF comparison
(c) 标准差对比
(c) SD comparison
(d) 平均梯度对比
(d) AG comparison
(e) 基于伪影的指标对比
(e) Nabf comparison
(f) 相关系数对比
(f) CC comparison

图 12　LLVIP 数据集中 30 对图像融合结果的客观评价指标

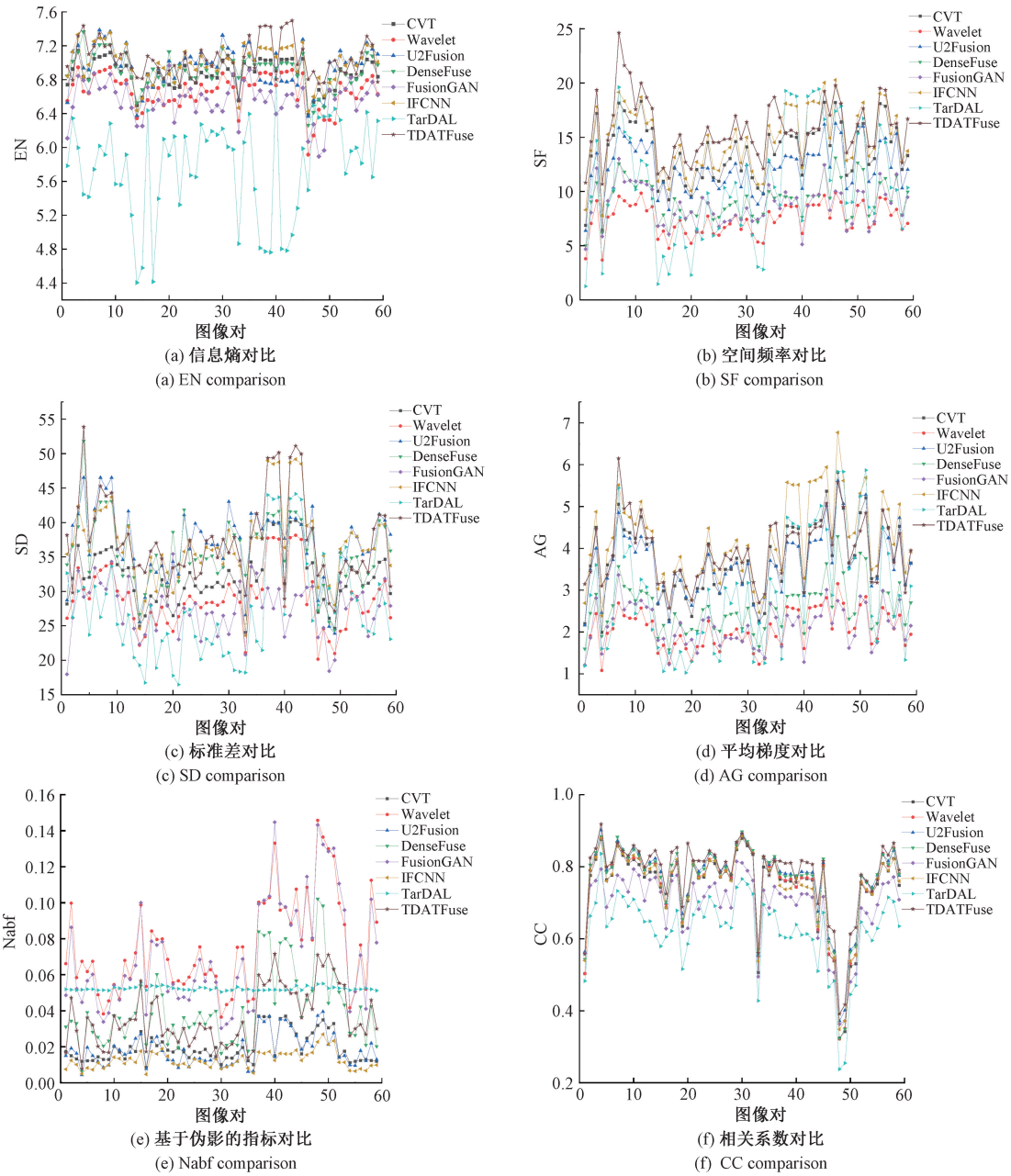
Fig. 12　LLVIP dataset 30 objective evaluation indicators of image fusion result

表 5　LLVIP 数据集的平均定量值

Table 5　The average quantitative value of the LLVIP dataset

算法	EN	SF	SD	AG	$N^{AB/F}$	CC
CVT	6. 454 0	10. 948 1	27. 898 6	3. 004 3	0. 025 2	0. 787 2
Wavelet	6. 273 3	6. 669 4	25. 422 8	1. 881 8	0. 054 0	0. 802 0
U2Fusion	5. 673 0	8. 864 2	25. 645 0	2. 349 9	0. 027 8	0. 788 9
DenseFuse	6. 637 3	8. 935 5	32. 540 0	2. 529 0	0. 026 1	0. 804 5
FusionGAN	6. 250 8	7. 501 3	25. 022 3	2. 018 9	0. 055 4	0. 717 2
IFCNN	6. 357 0	11. 367 1	28. 072 5	3. 108 6	**0. 015 9**	0. 796 0
TarDAL	2. 505 4	11. 614 7	25. 207 9	2. 796 4	0. 040 1	0. 532 4
TDATFuse	**6. 714 1**	**14. 011 7**	**33. 371 1**	**4. 171 6**	0. 022 3	**0. 805 5**

M3FD 数据集上进行客观评价指标对比实验折线如图 13 所示。可以看出，TDATFuse 的大部分融合结果的客观评价指标都处在较高水平。



(a) 信息熵对比
(a) EN comparison

(b) 空间频率对比
(b) SF comparison

(c) 标准差对比
(c) SD comparison

(d) 平均梯度对比
(d) AG comparison

(e) 基于伪影的指标对比
(e) Nabf comparison

(f) 相关系数对比
(f) CC comparison

图 13　M3FD 数据集中 59 对图像融合结果的客观评价指标

Fig. 13　M3FD dataset 59 objective evaluation indicators of image fusion results

由表 6 可知，在数据集 M3FD 中，EN、SF、SD、CC 四项指标均为最优值，分别比次优值提升了 0.73%、6.14%、0.66%、1.69%。

EN 和 CC 较高表示融合图像包含信息更加丰富。SF、SD、AG 较高说明融合图像有着较强的细节信息表达能力，而且对比度更佳，融合效果更加自然。但是，由于 TDATFuse 针对微光图像背景纹理信息量少的问题引入了微光重构网络，这种架构可能导致融合图像中某些区域的边缘强度超过源图像，在视觉上会被误判为伪影，所以 $N^{AB/F}$ 这个指标会相对较高。

2.3　消融实验

由于 Transformer 具有很好的捕获全局特征的能力，可以提取图像中丰富的各部分特征信息，而目标差分注意力模块可以增强融合图像中的红外目标，使红外目标更加明显，所以为了验证 Transformer 特征提取网络和目

标差分注意力模块对实验结果的影响,对融合网络结构进行 3 组消融实验[20],分别为图 14(c)两个模块均不使用、图 14(d)仅引入 TFENet、图 14(e)仅引入 TDABlock、图 14(f)TDATFuse。消融实验对比结果如图 14 所示。

表 6　M3FD 数据集的平均定量值

Table 6　The average quantitative value of the M3FD dataset

算法	EN	SF	SD	AG	$N^{AB/F}$	CC
CVT	6.847 1	13.774 0	32.048 8	3.739 9	**0.012 2**	0.740 0
Wavelet	6.641 3	7.474 0	28.959 2	2.036 5	0.089 0	0.759 6
U2Fusion	6.945 1	12.258 1	37.004 3	3.636 2	0.013 6	0.765 4
DenseFuse	6.901 2	9.271 1	35.893 5	2.554 2	0.020 5	0.759 5
FusionGAN	6.539 0	8.757 4	27.597 9	2.121 8	0.077 8	0.691 3
IFCNN	6.987 1	14.922 3	37.185 4	**4.197 8**	0.097 6	0.746 5
TarDAL	5.717 2	10.426 7	27.229 5	2.681 4	0.051 5	0.624 4
TDATFuse	**7.038 2**	**15.839 1**	**37.431 7**	3.923 6	0.029 4	**0.778 3**

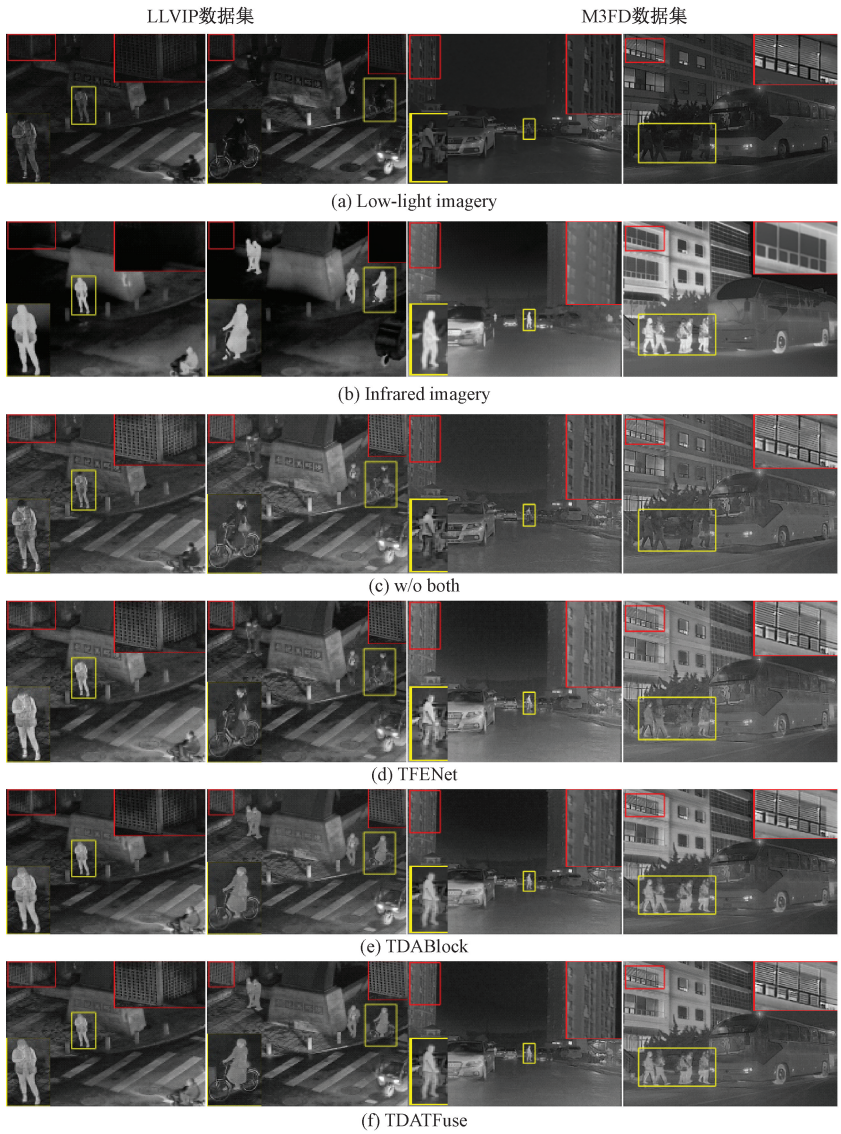


图 14　消融实验对比结果

Fig. 14　The results of ablation experiment were compared

图14中,本文的消融实验旨在分析Transformer特征提取网络和目标差分注意力模块对融合效果的影响。可以观察到,在没有TFENet和TDABlock(w/o Both)的情况下,红外特征缺失严重,红外目标在融合后的图像中不够突出(图14(c)中黄框),图像整体趋向于微光图像,且局部细节无法与整个图像建立联系,导致部分纹理信息融合的不够自然(图14(c)中红框)。

相比于上述情况,仅加入TFENet的情况下(图14(d)),融合网络可以很好地构建长程依赖关系,使得融合图像更加自然,但是红外目标仍不够显著(图14(d)中的黄框部分)。随后,对仅保留了TDABlock的融合网络进行实验,来验证其对红外目标的增强效果(图14(e)),在图14(e)中黄框可以看出,融合图像中的红外特征显著增加,不过在保留背景纹理上处理较差(图14(e)中红框),部分纹理信息丢失,导致融合结果无法从源图像中提取详细特征。

最后,本文使用的TDATFuse算法包含了Transformer特征提取网络和目标差分注意力模块,由图14(f)可以看到,融合结果在保证红外目标显著的同时(图14(f)中黄框),另整体图像包含丰富且清晰的背景纹理信息(图14(f)中红框),上述分析说明TDATFuse在主观评价上得到较好的视觉效果。本文从客观评价指标对TDATFuse的融合效果进行评判,并利用EN、SF、SD、AG、$N^{AB/F}$和CC 6个评价指标,分别在LLVIP和M3FD数据集上对本消融实验做客观评价,如表7和8所示。

表7　LLVIP数据集的消融实验客观评价指标

Table 7　Objective evaluation index of LLVIP data set ablation experiment

消融模块	EN	SF	SD	AG	$N^{AB/F}$	CC
TFENet	6.650 6	11.966 8	32.403 4	3.610 3	0.025 8	0.765 5
TDABlock	6.447 0	12.446 6	28.984 7	3.708 1	0.031 0	0.713 7
w/o Both	6.576 7	**14.458 5**	30.420 6	3.972 4	0.023 7	0.676 1
TDATFuse	**6.714 1**	14.011 7	**33.371 1**	**4.171 6**	**0.022 3**	**0.805 5**

表8　M3FD数据集的消融实验客观评价指标

Table 8　Objective evaluation index of M3FD data set ablation experiment

消融模块	EN	SF	SD	AG	$N^{AB/F}$	CC
TFENet	6.826 7	14.428 2	36.199 2	3.673 1	0.037 3	0.765 9
TDABlock	6.838 9	14.281 8	35.154 8	3.665 9	0.040 9	**0.788 6**
w/o Both	6.610 6	15.290 1	31.614 4	3.805 0	0.045 8	0.747 4
TDATFuse	**7.038 2**	**15.839 1**	**37.431 7**	**3.923 6**	**0.029 4**	0.778 3

表7和8为TDATFuse与各种网络结构的定量比较。综上所述,与其他3种融合网络相比,完整的结构具有明显优势。

3　结　论

本文提出了用于红外与微光图像融合的目标差分注意力和Transformer算法。该算法主要由微光重构网络、目标差分注意力模块、CNN特征提取网络、Transformer特征提取网络、纹理保留模块和特征重建网络六部分构成。首先,微光图像进入微光重构网络提升亮度、丰富细节信息,随后利用CNN特征提取网络和Transformer特征提取网络提取局部特征和全局特征,同时,利用目标差分注意力模块,增强红外目标特征,然后利用纹理保留模块最大程度的保留特征的背景纹理信息,最后通过重建网络得到最终的融合图像。实验结果表明该算法在主观视觉评价和客观指标评价中优于已有文献中具有代表性的算法,达到了较好的融合效果。在今后的工作中,本团队将继续探索图像融合领域,并着重开发特征提取及融合关键步骤,使其良好用于复杂的道路场景。

参考文献

[1] LI H, WU X J. DenseFuse: A fusion approach to infrared and visible images[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2614-2623.

[2] MA J, YU W, LIANG P, et al. FusionGAN: A generative adversarial network for infrared and visible image fusion[J]. Information Fusion, 2019, 48: 11-26.

[3] XU H, MA J, JIANG J, et al. U2Fusion: A unified unsupervised image fusion network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(1): 502-518.

[4] 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.

[5] ZHANG H, MA J. SDNet: A versatile squeeze-and-deco-mposition network for real-time image fusion [J]. Interna-tional Journal of Computer Vision, 2021, 129: 2761-2785.

[6] TANG L, YUAN J, ZHANG H, et al. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware [J]. Information Fusion, 2022, 83: 79-92.

[7] TANG L, XIANG X, ZHANG H, et al. DIVFusion: Darkness-free infrared and visible image fusion [J]. Information Fusion, 2023, 91: 477-493.

[8] JIA X, ZHU C, LI M, et al. LLVIP: A visible-infrared paired dataset for low-light vision[C]. Proceedings of the IEE-E/CVF International Conference on Computer Vision, 2021: 3496-3504.

[9] HWANG S, PARK J, KIM N, et al. Multispectral pedestrian detection: Benchmark datasetand baseline [C]. Proceedings-of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1037-1045.

[10] LIU J, FAN X, HUANG Z, et al. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 5802-5811.

[11] NENCINI F, GARZELLI A, BARONTI S, et al. Remote sensing image fusion using the curvelet transform [J]. Infor-Mation Fusion, 2007, 8(2): 143-156.

[12] MA J, CHEN C, LI C, et al. Infrared and visible image fusion via gradient transfer and total variation minimization [J]. Information Fusion, 2016, 31: 100-109.

[13] 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.

[14] MA J, TANG L, XU M, et al. STDFusionNet: An infrared and visible image fusion network based on salient target detection [J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-13.

[15] 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.

[16] ZHOU H, WU W, ZHANG Y, et al. Semantic supervised infrared and visible image fusion via a dual-discriminator generative adversarial Network[J]. IEEE Transactions on Multimedi-a, 2021, 2(14): 1-1.

[17] LI H, WU X, TARIQ D, et al. 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.

[18] LIU T, LAM K, ZHAO R, et al. Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 1(9): 1-1.

[19] MAO X, SHEN C, YANG Y B. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections [J]. Advances in Neural Information Processing Systems, 2016, 29: 2810-2818.

[20] 陈广秋, 温奇璋, 尹文卿, 等. 用于红外与可见光图像融合的注意力残差密集融合网络[J]. 电子测量与仪器学报, 2023, 37(8): 182-193.
CHEN G Q, WEN Q ZH, YIN W Q, et al. Attentional residual dense connection fusion network for infrared and visible image fusion [J]. Journal of Electronic Measurement and Instrumentation, 2023, 37 (8): 182-193.

作者简介

陈广秋, 1999 年于吉林大学获得学士学位, 2006 年于吉林大学获得硕士学位, 2015 年于吉林大学获得博士学位, 现为长春理工大学副教授, 主要研究方向为图像处理与机器视觉。
E-mail: gaungqiu_chen@126.com

Chen Guangqiu received his B. Sc. degree from Jilin University in 1999, M. Sc. degree from Jilin University in 2006 and Ph. D. degree from Jilin University in 2015, respectively. Now he is an associate professor in Changchun University of Science and Technology. His main research interests include image processing and machine vision.

代宇航, 2022 年于吉林建筑大学获得学士学位, 现为长春理工大学硕士研究生, 主要研究方向为图像处理与机器视觉。
E-mail: 2909119265@qq.com

Dai Yuhang received B. Sc. degree from Jilin Jianzhu University in 2022. She is now a M. Sc. candidate at Changchun University of Science and Technology. Her main research interests include image processing and machine vision.

段锦(通信作者),1993 年于北京理工大学获得学士学位,1998 年于沈阳工业学院获得硕士学位,2004 年于吉林大学获得博士学位,现为长春理工大学教授,主要研究方向为偏振成像探测、图像处理与模式识别、数字光学环境仿真。

E-mail: duanjin@ vip. sina. com

Duan Jin (Corresponding author) received his B. Sc. degree from Beijing Institute of Technology in 1993, M. Sc. degree from Shenyang Institute of Technology in 1998 and Ph. D. degree from Jilin University in 2004, respectively. Now he is a professor in Changchun University of Science and Technology. His main research interests include polarization imaging detection, image processing and pattern recognition, digital optical environment simulation.

黄丹丹,2007 年于长春理工大学大学获得学士学位,2009 年于东北大学获得硕士学位,2014 年于大连理工大学获得博士学位,现为长春理工大学讲师,主要研究方向为计算机视觉和机器学习。

E-mail: hdd@ cust. edu. cn

Huang Dandan received her B. Sc. degree from Changchun University of Science and Technology in 2007, M. Sc. degree from Northeastern University in 2009 and Ph. D. degree from Dalian University of Technology in 2014, respectively. Now she is a lecturer in Changchun University of Science and Technology. Her main research interests include computer vision and machine learning.