引用格式：WANG Zhishe, SHAO Wenyu, YANG Fengbao, et al. Infrared and Visible Image Fusion Method via Interactive Attention-based Generative Adversarial Network[J]. Acta Photonica Sinica, 2022, 51(4):0410002
王志社,邵文禹,杨风暴,等.红外与可见光图像交互注意力生成对抗融合方法[J].光子学报,2022,51(4):0410002

红外与可见光图像交互注意力生成对抗融合方法

王志社[1],邵文禹[1],杨风暴[2],陈彦林[1]
(1 太原科技大学 应用科学学院,太原 030024)
(2 中北大学 信息与通信工程学院,太原 030051)

摘　要:为了解决生成对抗融合方法获得的融合图像不能同时保留红外图像典型目标和可见光图像纹理细节的问题,提出一种红外与可见光图像交互注意力生成对抗融合方法。首先,在生成网络模型中采用权重参数共享的双路编码器架构,利用多尺度聚合卷积模块提取源图像各自的深度特征;其次,在融合层设计上,利用交互注意力融合模型建立两类图像局部特征的全局依赖特性,获得的注意力图更聚焦于红外典型目标和可见光纹理细节,实现红外与可见光图像端到端融合。最后,在对抗网络模型中,采用双鉴别器均衡判定融合图像与源图像间的真假性,相互补偿的损失函数优化生成网络模型获得最佳的融合结果。与现有典型融合方法的对比实验结果表明,该方法能够获得更平衡的融合结果,在主观视觉描述和客观指标评价上都优于其他方法。

关键词:图像融合;交互注意力;生成对抗网络;深度学习;红外图像;可见光图像
中图分类号:TP391　　**文献标识码**:A　　**doi**:10.3788/gzxb20225104.0410002

0　引言

红外传感器对热源辐射敏感,通过捕捉物体发出的热辐射感知热源目标特性,但所获得的红外图像通常缺乏结构特征和纹理细节。相反,可见光传感器通过光反射成像,图像具有较高的空间分辨率和丰富的纹理细节,但不能有效突出目标特性,且易受到外界环境影响,特别在低照度的环境条件下,信息丢失严重。红外与可见光图像融合旨在综合两类传感器的优势,互补生成的融合图像具有更好的目标感知和场景表达,在目标跟踪[1]、目标检测[2]和行人重识别[3]等领域有广泛应用。

现有的传统图像融合方法大致可分为多尺度变换[4]、稀疏表示[5]、显著性融合[6]、子空间融合[7]和拟态融合[8]等。传统图像融合方法通常以相同的特征变换或特征表示提取图像特征,采用合适的融合规则进行合并,再通过反变换重构获得最终融合图像。由于红外与可见光传感器成像机制不同,红外图像以像素亮度表征目标特征,而可见光图像以边缘和梯度表征场景纹理。传统融合方法不考虑源图像的内在不同特性,采用相同的变换或表示模型无差别地提取图像特征,不可避免地造成融合性能低、视觉效果差的结果。此外,融合规则是人为设定的,且越来越复杂,计算成本高,限制了图像融合的实际应用。

目前,由于深度学习的卷积操作具有很强的特征提取能力,且可从大量数据中学习构建模型参数,深度学习成为图像融合领域最有潜力的方向[9]。深度学习融合方法可粗略分为卷积神经网络[10-17]和生成对抗融合方法[18-22]。文献[10-13]采用编码模型提取图像深度特征,设计相应的融合规则,再利用解码模型重构融合图像。特别地,文献[12]采用了空间注意力作为融合规则,且利用中间特征和补偿特征来提高深度特征表征能力。文献[13]构建了L_p正则化注意力融合模型,从通道和空间维度分别提取深度特征的注意力特征

基金项目:山西省基础研究计划资助项目(No.201901D111260),信息探测与处理山西省重点实验室开放基金(No.ISPT2020-4)
第一作者:王志社(1982—),男,副教授,博士,主要研究方向为图像融合、深度学习、机器视觉。Email: wangzs@tyust.edu.cn
收稿日期:2021-11-22;**录用日期**:2021-12-22

图。尽管这些方法取得了较好的融合结果，但都是非端到端融合网络，仍需人为设定融合规则。文献[18]提出生成对抗图像融合方法，由于采用单一的对抗机制，导致融合结果不平衡，偏向于红外图像，可见光图像纹理边缘信息丢失严重。文献[19]利用生成对抗机制将图像融合转变为多分类限定问题，虽能缓解融合不平衡问题，但融合图像中目标边缘模糊，纹理边缘信息依然缺失。

为此，本文提出一种红外与可见光图像交互注意力生成对抗融合方法，采用权重共享的双路编码网络结构分别提取源图像各自的深度特征，利用交互注意力融合模型建立深度特征的全局依赖特性，获得的注意力图像更聚焦于红外典型目标和可见光纹理细节。此外，双鉴别器和互补损失函数设计进一步优化生成对抗网络模型，使得融合图像能够同时保留更突出的红外图像典型目标和更清晰的可见光图像纹理细节，获得更好的图像融合性能。

1 融合方法

1.1 融合网络总体结构

交互注意力生成对抗融合原理框图如图1(a)所示。在生成网络模型中，红外和可见光图像作为输入源，通过编码-解码网络，双编码网络提取源图像各自的多尺度深度特征，交互注意力融合模型(Interactive Attention Fusion model, IAFM)建立多尺度局部特征的全局依赖特性，获得融合注意力图，最后经过解码网络重构获得融合图像。在对抗网络模型中，设计了面向红外和可见光图像的双鉴别器，能更均衡地判断融合图像与源图像的真假性，优化生成网络模型，使生成的融合图像更接近源图像的真实数据分布，最终获得更平衡的融合结果。

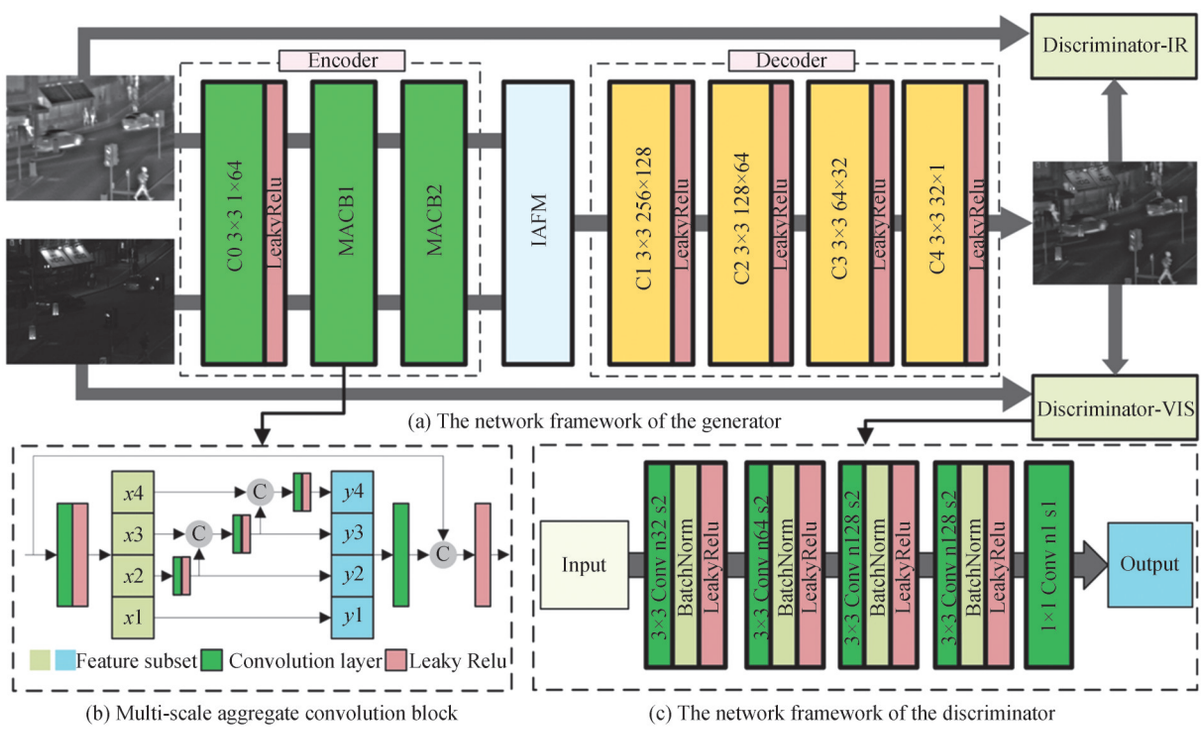


图1 交互注意力生成对抗融合原理

Fig.1 The principle of interactive attention-based generative adversarial network

生成网络模型包括编码部分、交互注意力融合模块和解码部分。编码部分由1个普通卷积层(C0)和2个多尺度聚合卷积模块(Multi-scale Aggregate Convolution block, MACB)构成，如图1(b)所示。普通卷积层提取图像的底层特征，而多尺度聚合卷积模块提取图像的高层特征。对于输入的底层特征，首先利用1×1卷积转换通道数，并将输入特征图分成4个特征子图($x1$,$x2$,$x3$,$x4$)。每个子图具有相同的空间大小，通道数为输入特征图的1/4。然后，除第一个特征子图$x1$以外，其他每个子图通过3×3卷积后，通道连接

（Concatenate）到下一个特征子图中。多尺度聚合卷积模块无需采用上采样或者下采样，以多视场聚合方式获取多尺度深度特征，尽可能保留有用的特征信息。解码部分由4个卷积核大小为3×3的普通卷积层组成。

对抗网络模型如图1（c）所示，采用红外（Discriminator-IR）和可见光（Discriminator-VIS）双鉴别器设计，网络结构相同，都由5个普通卷积层组成，其中前4个卷积层采用大小为3×3的卷积核，卷积步长为2，滤波器组的参数分别是32、64、128、128和1。采用BatchNorm对前4层普通卷积的输出数据进行归一化操作，加速收敛速度，避免出现梯度消失等问题。融合网络的卷积层均采用LeakyRelu函数作为激活函数，其他参数设定如图1所示。

1.2 交互注意力融合模型

交互注意力融合模型如图2所示，由通道注意力和空间注意力级联组成，从通道和空间维度上建立局部特征的全局依赖特性。对于输入的红外和可见光图像深度特征 $\boldsymbol{\Phi}_{\mathrm{I}}$ 和 $\boldsymbol{\Phi}_{\mathrm{V}}\in\boldsymbol{R}^{H\times W\times C}$，首先经过全局平均池化层，将深度特征转化为通道描述向量，获得相应的初始通道加权系数 $\varphi_{\mathrm{I}}^{\mathrm{ca}}$ 和 $\varphi_{\mathrm{V}}^{\mathrm{ca}}\in\boldsymbol{R}^{1\times 1\times C}$，即

$$\varphi_{\mathrm{I}}^{\mathrm{ca}}(c)=\mathrm{AvgPool}(\boldsymbol{\Phi}_{\mathrm{I}}) \tag{1}$$

$$\varphi_{\mathrm{V}}^{\mathrm{ca}}(c)=\mathrm{AvgPool}(\boldsymbol{\Phi}_{\mathrm{V}}) \tag{2}$$

式中，AvgPool(•)表示全局平均池化操作，$c=1,2,\cdots,C$ 表示通道索引。随后，利用Softmax操作获得最终的通道加权系数 $\beta_{\mathrm{I}}^{\mathrm{ca}}$ 和 $\beta_{\mathrm{V}}^{\mathrm{ca}}\in\boldsymbol{R}^{1\times 1\times C}$，即

$$\beta_{\mathrm{I}}^{\mathrm{ca}}(c)=\frac{\exp(\varphi_{\mathrm{I}}^{\mathrm{ca}}(c))}{\exp(\varphi_{\mathrm{I}}^{\mathrm{ca}}(c))+\exp(\varphi_{\mathrm{V}}^{\mathrm{ca}}(c))} \tag{3}$$

$$\beta_{\mathrm{V}}^{\mathrm{ca}}(c)=\frac{\exp(\varphi_{\mathrm{V}}^{\mathrm{ca}}(c))}{\exp(\varphi_{\mathrm{I}}^{\mathrm{ca}}(c))+\exp(\varphi_{\mathrm{V}}^{\mathrm{ca}}(c))} \tag{4}$$

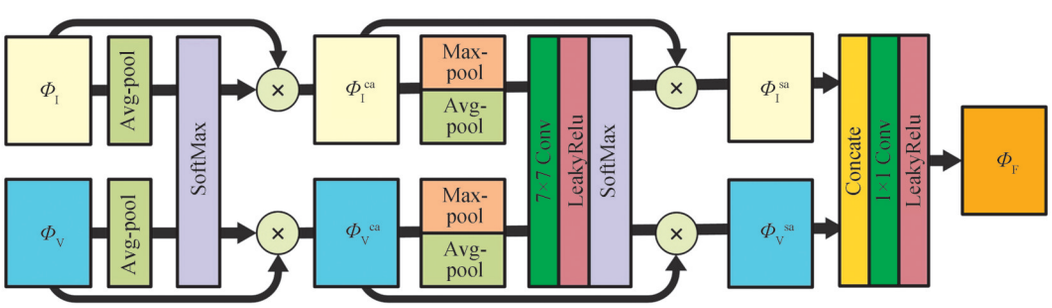


图2 交互注意力融合模型

Fig.2 Interactive attention fusion model

将最终通道加权系数与原始输入的深度特征进行元素相乘，获得红外和可见光图像的通道注意力图 $\boldsymbol{\Phi}_{\mathrm{I}}^{\mathrm{ca}}$ 和 $\boldsymbol{\Phi}_{\mathrm{V}}^{\mathrm{ca}}$ 分别为

$$\boldsymbol{\Phi}_{\mathrm{I}}^{\mathrm{ca}}(i,j)=\boldsymbol{\Phi}_{\mathrm{I}}(i,j)\times\beta_{\mathrm{I}}^{\mathrm{ca}}(c) \tag{5}$$

$$\boldsymbol{\Phi}_{\mathrm{V}}^{\mathrm{ca}}(i,j)=\boldsymbol{\Phi}_{\mathrm{V}}(i,j)\times\beta_{\mathrm{V}}^{\mathrm{ca}}(c) \tag{6}$$

接着，将红外和可见光图像的通道注意力图作为空间注意力模型的输入特征，首先经过最大池化和平均池化层，经过通道连接和7×7卷积层后，获得初始空间加权系数 $\varphi_{\mathrm{I}}^{\mathrm{sa}}$ 和 $\varphi_{\mathrm{V}}^{\mathrm{sa}}\in\boldsymbol{R}^{M\times N\times 1}$，即

$$\varphi_{\mathrm{I}}^{\mathrm{sa}}(i,j)=\mathrm{Conv}(\mathrm{Concate}[\mathrm{MaxPool}(\boldsymbol{\Phi}_{\mathrm{I}}^{\mathrm{ca}}),\ \mathrm{AvgPool}(\boldsymbol{\Phi}_{\mathrm{I}}^{\mathrm{ca}})]) \tag{7}$$

$$\varphi_{\mathrm{V}}^{\mathrm{sa}}(i,j)=\mathrm{Conv}(\mathrm{Concate}[\mathrm{MaxPool}(\boldsymbol{\Phi}_{\mathrm{V}}^{\mathrm{ca}}),\ \mathrm{AvgPool}(\boldsymbol{\Phi}_{\mathrm{V}}^{\mathrm{ca}})]) \tag{8}$$

式中，MaxPool(•)表示最大池化，Concate[•]表示通道连接。利用Softmax操作获得最终的空间加权系数 $\beta_{\mathrm{I}}^{\mathrm{sa}}$ 和 $\beta_{\mathrm{V}}^{\mathrm{sa}}\in\boldsymbol{R}^{M\times N\times 1}$，即

$$\beta_{\mathrm{I}}^{\mathrm{sa}}(i,j)=\frac{\exp(\varphi_{\mathrm{I}}^{\mathrm{sa}}(i,j))}{\exp(\varphi_{\mathrm{I}}^{\mathrm{sa}}(i,j))+\exp(\varphi_{\mathrm{V}}^{\mathrm{sa}}(i,j))} \tag{9}$$

$$\beta_{\mathrm{V}}^{\mathrm{sa}}(i,j)=\frac{\exp(\varphi_{\mathrm{V}}^{\mathrm{sa}}(i,j))}{\exp(\varphi_{\mathrm{I}}^{\mathrm{sa}}(i,j))+\exp(\varphi_{\mathrm{V}}^{\mathrm{sa}}(i,j))} \tag{10}$$

将最终空间加权系数与通道注意力图进行元素相乘，获得红外和可见光图像的空间注意力图$\boldsymbol{\Phi}_{\mathrm{I}}^{\mathrm{sa}}$和$\boldsymbol{\Phi}_{\mathrm{V}}^{\mathrm{sa}}$分别为

$$\boldsymbol{\Phi}_{\mathrm{I}}^{\mathrm{sa}}(i,j)=\boldsymbol{\Phi}_{\mathrm{I}}^{\mathrm{ca}}(i,j)\times\beta_{\mathrm{I}}^{\mathrm{sa}}(i,j) \tag{11}$$

$$\boldsymbol{\Phi}_{\mathrm{V}}^{\mathrm{sa}}(i,j)=\boldsymbol{\Phi}_{\mathrm{V}}^{\mathrm{ca}}(i,j)\times\beta_{\mathrm{V}}^{\mathrm{sa}}(i,j) \tag{12}$$

最后，将红外和可见光图像的空间注意力图进行通道连接和1×1卷积层后，获得最终的融合注意力图$\boldsymbol{\Phi}_{\mathrm{F}}$为

$$\boldsymbol{\Phi}_{\mathrm{F}}(i,j)=\mathrm{Conv}(\mathrm{Concate}[\boldsymbol{\Phi}_{\mathrm{I}}^{\mathrm{sa}}(i,j),\ \boldsymbol{\Phi}_{\mathrm{V}}^{\mathrm{sa}}(i,j)]) \tag{13}$$

1.3 损失函数

生成网络模型的损失函数L_{G}由对抗损失L_{adv}和内容损失L_{con}两部分构成，即

$$L_{\mathrm{G}}=L_{\mathrm{adv}}+L_{\mathrm{con}} \tag{14}$$

虽然红外图像以像素亮度表征目标特征，可见光图像以边缘和梯度表征场景细节，但事实上，可见光图像也存在一定的亮度分布信息。因此，采用Frobenius范数分别约束融合图像与红外、可见光图像间的数据分布相似度，保留红外与可见光图像像素强度，且通过比例系数突出红外目标的亮度信息。考虑到Frobenius范数会放大融合图像与源图像之间的灰度差异，导致可见光图像的纹理细节信息损失，又采用L1范数进一步约束融合图像与可见光图像的相似性，保留可见光的纹理细节信息。因此，两个损失函数设计是相互补偿的，使整个损失函数平衡，生成的融合图像在突出红外目标亮度前提下，保留了更加丰富的可见光图像纹理细节信息。内容损失函数L_{con}可表示为

$$L_{\mathrm{con}}=\frac{1}{HW}\left[(\beta\|I_{\mathrm{f}}-I_{\mathrm{ir}}\|_{\mathrm{F}}^{2}+\|I_{\mathrm{f}}-I_{\mathrm{vis}}\|_{\mathrm{F}}^{2})+\|I_{\mathrm{f}}-I_{\mathrm{vis}}\|_{1}\right] \tag{15}$$

式中，H、W分别表示源图像的高和宽，β为调整系数且取值大于1，I_{f}、I_{ir}和I_{vis}分别表示融合图像、红外图像和可见光图像，$\|\cdot\|_{\mathrm{F}}$为Frobenius范数，$\|\cdot\|_{1}$表示L1范数。

此外，对抗损失函数可表示为

$$L_{\mathrm{adv}}=-\frac{1}{N}\sum_{i=1}^{N}[D_{\mathrm{ir}}(I_{\mathrm{f}})]-\frac{1}{N}\sum_{i=1}^{N}[D_{\mathrm{vis}}(I_{\mathrm{f}})] \tag{16}$$

式中，N表示融合图像数量，$D_{\mathrm{ir}}(\cdot)$与$D_{\mathrm{vis}}(\cdot)$表示两个鉴别器的输出结果。

在对抗网络模型中，设计了红外(Discriminator-IR)和可见光(Discriminator-VIS)双鉴别器，通过鉴别损失函数可以平衡判定融合图像与源图像的真假性，进而与生成网络模型形成对抗博弈，使生成融合图像更趋向于源图像的真实数据分布。红外和可见光图像的鉴别器损失函数可表示为

$$L_{D_{\mathrm{ir}}}=-\frac{1}{N}\sum_{i=1}^{N}D_{\mathrm{ir}}(I_{\mathrm{ir}})+\frac{1}{N}\sum_{i=1}^{N}D_{\mathrm{ir}}(I_{\mathrm{f}})+\lambda\frac{1}{N}\sum_{i=1}^{N}(\|\nabla D_{\mathrm{ir}}(I_{\mathrm{ir}})\|_{2}-1) \tag{17}$$

$$L_{D_{\mathrm{vis}}}=-\frac{1}{N}\sum_{i=1}^{N}D_{\mathrm{vis}}(I_{\mathrm{vis}})+\frac{1}{N}\sum_{i=1}^{N}D_{\mathrm{vis}}(I_{\mathrm{f}})+\lambda\frac{1}{N}\sum_{i=1}^{N}(\|\nabla D_{\mathrm{vis}}(I_{\mathrm{vis}})\|_{2}-1) \tag{18}$$

式中，∇表示梯度算子，第一项表示源图像的鉴别器损失，第二项表示融合图像的鉴别器损失，前两项表示源图像与融合图像的Wasserstein距离，最后一项为梯度惩罚，限制鉴别器的学习能力，λ为正则化参数。

2 实验验证

2.1 训练与测试参数设定

在训练过程中，由于红外与可见光图像数据集有限，在TNO数据集上采用滑动步长为12，将原始图像对尺寸裁剪为256×256，灰度值范围转换为[0,1]，以获得10 653组红外与可见光图像对作为训练数据集。此外，采用Adam优化器更新网络模型参数，Batchsize和Epoch分别设置为4和6。生成网络模型和对抗网络模型的学习率分别设置为1×10^{-4}和4×10^{-4}，且对应的训练次数分别设置为1和2。在损失函数参数设置中，平衡因子β为3.5，正则化参数λ设置为10。实验测试平台采用Intel i9-10850k CPU、64 GB内存和NVIDIA GeForce GTX 3090显卡，训练和测试环境为Python和PyTorch平台。

在测试过程中，从TNO[23]、Roadscene[24]数据集分别选取25和30组红外和可见光图像、以及Nato_camp序列作为测试数据。本文方法与现有的9种典型融合方法进行比较，包括WLS[6]、DenseFuse[10]、IFCNN[11]、SEDRFuse[12]、U2Fusion[15]、PMGI[16]、FusionGAN[18]、GANMcC[19]和RFN-Nest[14]。客观评价采用8个融合评价指标，分别为平均梯度（Average Gradient，AG）、标准差（Standard Deviation，SD）、互信息（Mutual Information，MI）、相位一致性（Phase Congruency，PC）、非线性相关信息熵（Nonlinear Correlation Information Entropy，NCIE）、空间频率（Spatial Frequency，SF）、多尺度结构相似性（Multi-Scale Structural Similarity Index Measure，MS_SSIM）和视觉信息保真度（Visual Information Fidelity，VIF）。在客观评价中，评价指标数值越大表明融合性能越好。此外，还采用平均指标提高率（Average Metric Improvement Rate，AMIR）来量化指标提高程度，其公式表示为

$$\mathrm{AMIR}=\frac{M_{\mathrm{ours}}-\frac{1}{N}\sum M_{\mathrm{other}}}{\frac{1}{N}\sum M_{\mathrm{other}}} \tag{19}$$

式中，M_{ours}和M_{other}分别表示本文方法和其他对比方法取得的客观指标值，N表示对比方法个数。

2.2 消融实验

为了验证交互注意力融合模型的有效性，将与无注意力模型（记作No_atten）、仅有通道注意力模型（记作Only_CA）、仅有空间注意力模型（记作Only_SA）和空间级联通道注意力模型（记作SA_CA）进行比较。实验采用TNO数据集25组图像和8个评价指标。图3给出了5种模型的融合对比结果。从结果可以看出，No_atten模型融合结果既丢失了红外目标亮度信息，又缺失了可见光的纹理细节。Only_CA和Only_SA模型能够保留红外典型目标，但可见光图像的纹理细节依然有所缺失。相比之下，SA_CA和交互注意力模型取得了更平衡的融合结果，同时保留了红外图像的典型目标和可见光图像的纹理细节，从主观上来看，两者之间的差异不明显，要从客观评价上来比较两者的融合性能。

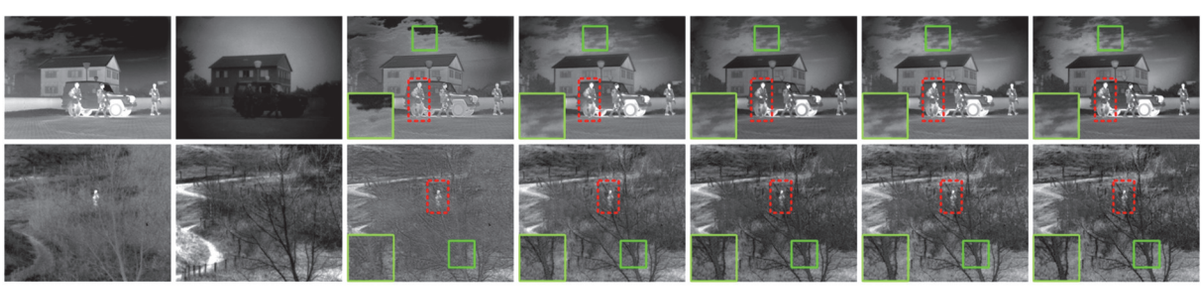


图3 5种融合模型的主观对比结果

Fig.3 The subjective comparison results of five fusion models

表1给出了5种模型的客观评价对比结果，最优值和次优值分别以加粗和下划线标注。可以看出，Only_CA和Only_SA模型显著好于No_atten模型，表明注意力机制可以有效提高图像融合性能。此外，SA_CA模型和交互注意力模型的融合性能高于Only_CA和Only_SA模型，表明交互的通道和空间注意力模型显著好于单个注意力模型。本文方法取得了指标AG、MI、NCIE、SF和VIF的最优值，而SA_CA模型取得了指标AG、SD、PC、NCIE、SF和VIF的次优值。对比其他4个模型，本文的交互注意力模型取得了最

表1 TNO数据集的5种融合模型的客观对比结果

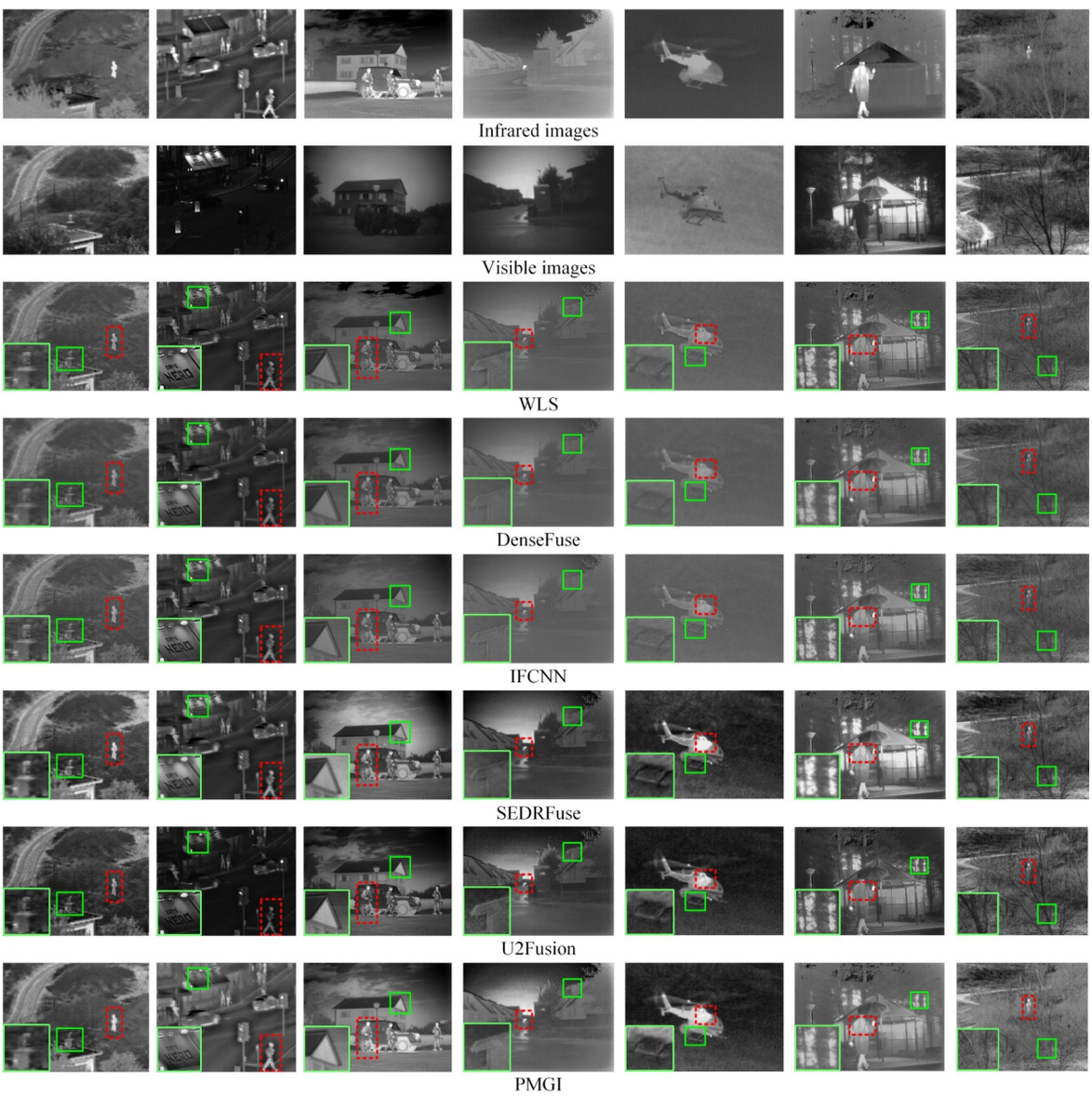
Table 1 The objective comparison results of five fusion models on the TNO dataset

Models	AG	SD	MI	PC	NCIE	SF	MS_SSIM	VIF
No_atten	5.464 33	29.894 71	1.942 38	0.237 18	0.804 27	10.010 08	0.872 68	0.395 57
Only_CA	5.153 56	36.361 13	2.657 49	0.308 27	0.806 42	9.620 20	**0.892 71**	0.409 20
Only_SA	4.649 28	**37.827 80**	2.968 02	**0.316 78**	0.807 41	8.807 06	0.879 70	0.403 27
SA_CA	5.547 57	37.047 94	2.958 20	0.315 43	0.807 46	10.554 43	0.875 66	0.426 09
Ours	**5.572 04**	36.768 19	**3.047 12**	0.309 84	**0.807 91**	**10.588 10**	0.863 14	**0.426 42**

优的融合结果。

2.3 TNO数据集实验验证

为了验证本文方法的优越性，对TNO数据集进行实验验证，选取其中7组典型红外和可见光图像作为主观评价，包括Nato_camp、Street、Soldiers_with_jeep、Movie_01、Helicopter、Kaptein_1654和Sandpath。图4给出了7组图像的主观评价对比结果。为了便于直观观察，典型的红外目标和纹理细节分别以虚线框和实线框标注，且对纹理细节进行局部放大。可以看出，传统融合方法WLS在一定程度上保留了可见光的纹理细节，但是典型的红外目标信息丢失严重，存在较为严重的伪影现象。深度学习融合方法DenseFuse和IFCNN，由于采用加权平均的融合规则，获得融合图像倾向于保留可见光的纹理细节，典型的红外目标信息依然缺失严重。相比之下，SEDRFuse、U2Fusion和PMGI取得了相对满意的效果。尽管这样，这些方法依然不能有效保留红外目标亮度特性，目标特性不突出。FusionGAN和GANMcC的融合结果倾向于红外图像，能够保留红外图像的典型目标，但目标边缘模糊，且可见光的纹理细节丢失严重。RFN-Nest虽然采用两阶段训练，但所获得的结果倾向于保留更多的纹理细节，红外图像的目标信息严重丢失。总的来说，本文方法能够有效保留红外图像的典型目标和可见光图像的纹理细节，达到更平衡的融合结果，获得最优的视觉效果。



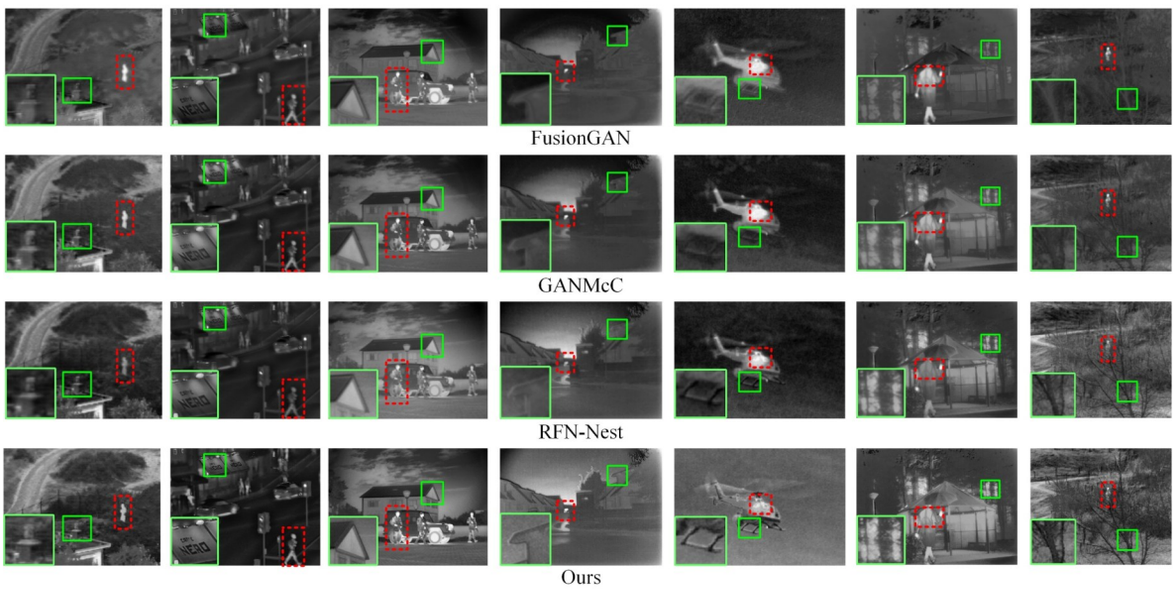


图4　TNO数据集上不同融合方法主观评价对比结果

Fig.4　The subjective comparison results of different fusion methods on the TNO dataset

表2给出了TNO数据集的客观评价指标对比结果，可以看出，本文方法获得了指标MI、PC、NCIE和VIF的最优平均值，指标AG、SD、SF的次优平均值，指标平均提高率分别为34.548%、21.416%、54.385%、33.056%、0.436%、33.735%、22.384%，指标MS_SSIM降低了1.670%，整体指标平均提高率为24.786%。为了进一步验证本文方法的有效性，表3给出了Nato_camp序列的客观评价指标对比结果。本文方法取得了指标SD、MI、PC和NCIE的最优值，指标VIF的次优值，指标平均提高率分别为18.593%、29.815%、78.433%、32.005%、0.505%、23.208%、10.656%，指标MS_SSIM降低了2.565%，整体指标平均提高率为23.831%。此外，可以发现本文方法在指标MS_SSIM未取得最优值，这是因为采用交互注意力融合模块，获得的融合图像既要保留红外图像的典型目标又要保留可见光图像的纹理细节，交互特征融合可能弱化了某些结构或边缘特征，导致指标MS_SSIM取得了相对较低的值。尽管如此，从多指标评价的结果来看，本文融合方法取得了最优的融合性能，客观评价与上述的主观评价一致。

表2　不同融合方法在TNO数据集上的客观评价指标对比结果

Table 2　The objective comparison results of different fusion methods on the TNO dataset

Methods	AG	SD	MI	PC	NCIE	SF	MS_SSIM	VIF
WLS	5.417 71	24.860 27	1.742 86	0.280 99	0.803 91	**10.793 84**	0.909 88	0.378 66
DenseFuse	3.193 31	22.857 69	2.035 89	0.290 28	0.804 51	6.094 43	0.876 96	0.330 90
IFCNN	5.467 53	24.067 12	1.801 91	0.278 68	0.804 04	10.131 41	0.904 74	0.389 20
SEDRFuse	3.544 11	**40.793 02**	2.110 14	0.172 27	0.804 62	6.794 46	0.895 05	0.316 82
U2Fusion	**5.615 15**	33.596 08	1.760 99	0.261 20	0.803 92	10.436 55	**0.919 82**	0.420 78
PMGI	4.695 36	34.768 16	2.108 01	0.243 11	0.804 82	8.999 78	0.887 96	0.390 59
FusionGAN	3.073 57	26.820 44	2.166 52	0.102 64	0.805 03	5.982 47	0.734 49	0.248 69
GANMcC	3.139 83	29.929 73	2.108 64	0.232 71	0.804 52	6.009 63	0.859 15	0.305 10
RFN－Nest	3.125 21	34.853 73	1.928 51	0.233 90	0.804 28	6.012 69	0.912 17	0.355 10
Ours	5.572 04	36.768 19	**3.047 12**	**0.309 84**	**0.807 91**	10.588 10	0.863 14	**0.426 42**
AMIR	34.548%	21.416%	54.385%	33.056%	0.436%	33.735%	－1.670%	22.384%

从评价结果来看，最优的MI、NCIE和PC指标表明本文方法能够从源图像提取更多特征信息，保留到融合结果上。这是因为本文方法采用双路编码-解码生成网络模型，多尺度聚合卷积模块提取了多尺度特征，能够有效地表征图像特征信息。最优的AG、SF指标表明本文方法的融合结果保留了更多的边缘和纹理特征信息，说明双鉴别器能够平衡融合图像与源图像的真实数据分布，相互补充的损失函数进一步平衡

了融合结果。此外，最优的SD和VIF指标表明本文方法具有最高的对比度和视觉效果，这是因为采用交互注意力融合模型能够从通道和空间维度上对局部特征进行建模，获取局部特征的全局依赖特性，使得注意力图更聚焦于红外图像的目标特性和可见光的纹理细节。主客观实验验证了方法的有效性，表明本文方法取得了较好的融合性能，优于其他9种典型融合方法。

表3 不同融合方法在Nato_camp序列上的客观评价指标对比结果

Table 3 The objective comparison results of different fusion methods on the Nato-camp sequence

Methods	AG	SD	MI	PC	NCIE	SF	MS_SSIM	VIF
WLS	6.119 13	24.220 50	1.475 49	0.244 28	0.803 25	11.297 25	0.906 43	0.309 24
DenseFuse	3.814 31	22.715 99	1.579 94	0.258 93	0.803 41	7.130 29	0.873 69	0.286 36
IFCNN	**6.651 62**	23.563 42	1.500 51	0.237 27	0.803 25	**12.168 38**	0.902 03	0.315 65
SEDRFuse	4.993 49	38.476 49	2.059 20	0.252 53	0.804 31	9.803 58	0.934 91	0.316 76
U2Fusion	5.961 61	34.031 64	1.593 96	0.259 71	0.803 42	11.554 53	**0.936 11**	**0.339 02**
PMGI	5.274 55	34.230 81	2.080 61	0.214 98	0.804 35	9.987 53	0.876 01	0.318 94
FusionGAN	3.115 54	25.355 11	2.431 53	0.082 94	0.805 50	6.404 57	0.670 09	0.207 64
GANMcC	4.333 90	33.617 44	1.994 60	0.228 22	0.804 04	8.180 66	0.900 37	0.278 32
RFN-Nest	3.568 69	35.239 55	1.687 92	0.264 75	0.803 65	7.057 25	0.925 03	0.304 43
Ours	5.775 83	**39.153 66**	**3.252 20**	**0.299 74**	**0.807 97**	11.442 50	0.857 93	0.329 06
AMIR	18.593%	29.815%	78.433%	32.005%	0.505%	23.208%	−2.565%	10.656%

2.4 Roadscene数据集实验验证

为进一步验证该融合方法的有效性，从Roadscene数据集中选取了30组红外和可见光图像进行实验验证。图5、6给出了“FLIR_06422”和“FLIR_07210”的主观评价对比结果，可以看出，对于红外图像的典型目标，如虚线框标注的行人和路灯，WLS、DenseFuse、IFCNN、U2Fusion和RFN-Nest的融合结果偏向于可见光图像，能保留可见光图像纹理细节信息，但红外图像的目标不突出，亮度特性丢失严重。对于可见光图像的细节特征，如实线框标注的地面和标志牌上“STOP”字样，SEDRFuse、PMGI、FusionGAN和GANMcC的融合结果偏向于红外图像，能够保留红外图像的典型目标，但细节信息损失严重。对比之下，本文方法的融合结果既能保留红外图像的典型目标，又能保留可见光图像的纹理细节，获得了最佳的视觉效果，更符合人类视觉系统。

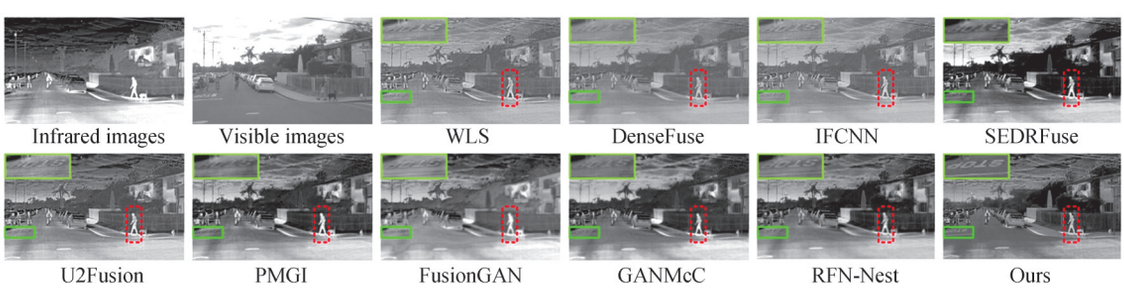


图5 FLIR_06422的不同融合方法主观评价对比结果

Fig.5 The subjective comparison results of different fusion methods for FLIR_06422

表4给出了Roadscene数据集的客观评价指标对比结果，本文融合方法在指标MI、PC、NCIE和VIF上取得了最优值，在指标AG和SF上取得了次优值，指标平均提高率分别为23.610%、0.923%、24.547%、38.148%、0.224%、26.580%、21.948%，指标MS_SSIM降低了4.427%，整体指标平均提高率为16.466%。客观实验结果验证本方法具有显著的融合性能。从主、客观评价结果来看，本文融合方法在2个数据集和1个序列上的融合性能都优于其他典型融合方法，表明了本文方法具有较强的鲁棒性和优越性。此外，为了进一步验证融合计算效率，传统方法WLS在CPU上进行测试，而深度学习方法都在GPU上测试。表5给出了不同融合方法计算效率的对比结果。本文方法的计算效率略低于DenseFuse和IFCNN，这是因为这两个方法采用了加权平均的融合规则。综合实验分析结果表明本文方法在取得更佳融合性能的同时，还具有

较高的计算效率。

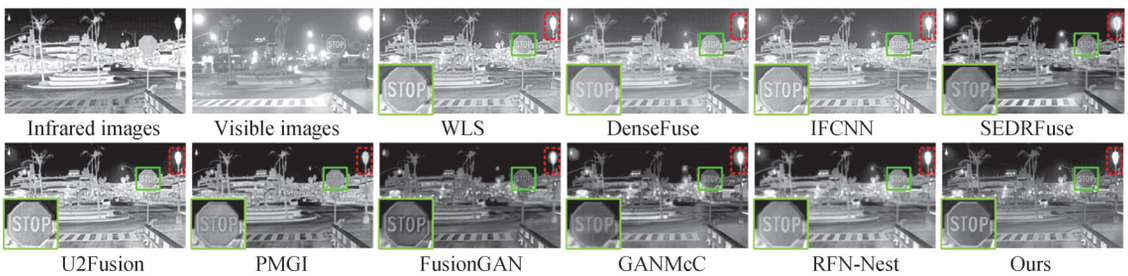


图6　FLIR_07210的不同融合方法主观评价对比结果

Fig.6　The subjective comparison results of different fusion methods for FLIR_07210

表4　不同融合方法在Roadscene数据集上的客观评价指标对比结果

Table 4　The objective comparison results of different fusion methods on the Roadscene dataset

Methods	AG	SD	MI	PC	NCIE	SF	MS_SSIM	VIF
WLS	5.883 14	29.056 59	2.361 24	0.342 41	0.806 11	12.347 23	0.885 22	0.311 46
DenseFuse	3.652 56	27.115 48	2.776 46	0.350 99	0.807 31	7.679 22	0.831 36	0.266 34
IFCNN	5.964 53	28.818 46	2.483 70	0.334 82	0.806 36	12.521 83	0.873 55	0.293 26
SEDRFuse	5.337 39	**46.910 59**	3.017 21	0.280 03	0.807 34	11.115 36	0.895 57	0.276 69
U2Fusion	**6.514 43**	45.098 18	3.001 63	0.348 45	0.807 95	**13.724 65**	**0.912 77**	0.315 35
PMGI	5.027 04	45.020 31	2.681 32	0.304 32	0.808 45	10.472 62	0.894 09	0.286 20
FusionGAN	3.919 09	39.887 01	2.672 07	0.110 38	0.807 26	8.450 70	0.732 31	0.190 09
GANMcC	4.076 88	42.787 94	2.805 20	0.280 49	0.806 81	8.451 23	0.862 17	0.243 64
RFN－Nest	3.830 42	44.210 13	2.887 68	0.262 81	0.806 80	8.115 65	0.876 61	0.265 58
Ours	6.071 37	39.124 91	**3.416 26**	**0.401 35**	**0.808 96**	13.062 80	0.825 99	**0.331 78**
AMIR	23.610%	0.923%	24.547%	38.148%	0.224%	26.580%	－4.247%	21.948%

表5　不同融合方法计算效率对比结果（单位：秒）

Table 5　The comparison results of computation efficiency for different fusion methods (units: s)

Dataset	WLS	Dense-Fuse	IFCNN	SEDR-Fuse	U2Fusion	PMGI	Fusion-GAN	GAN-McC	RFN-Nest	Ours
TNO	2.359	0.085	0.046	1.148	1.722	0.588	2.184	4.445	0.177	0.151
Sequence	0.455	0.029	0.015	1.158	0.659	0.181	0.678	1.329	0.068	0.059
Roadscene	1.087	0.048	0.024	1.137	0.992	0.293	1.135	2.271	0.096	0.087

3　结论

本文提出了一种红外与可见光图像交互注意力生成对抗融合方法，设计了双路编码-解码的生成网络模型，构造多尺度聚合卷积模块，有效提取源图像各自的深度特征；构建了交互注意力融合模型，建立了局部特征的全局依赖特性，使注意力图更聚焦于红外典型目标和可见纹理细节。在对抗网络模型中，设计了双鉴别器来判定融合图像与源图像间的真假性，互补的损失函数优化生成网络模型获得最佳的融合结果。实验结果表明，与其他9种典型融合方法相比，本文方法能够取得更平衡的主观视觉融合结果，在TNO、Nato_camp序列和Roadscene数据集上客观指标分别提高了24.786%、23.831%、16.466%，获得了最优的融合性能，且具有较高的计算效率和较强的鲁棒性。下一步工作将注意力机制引入对抗网络模型中，进一步提高红外与可见光图像融合性能。

参考文献

[1] ZHANG Xingchen, YE Ping, LEUNG H, et al. Object fusion tracking based on visible and infrared images: a comprehensive review[J]. Information Fusion, 2020, 63: 166-187.

[2] TU Zhengzheng, LI Zhun, LI Chenglong, et al. Multi-interactive dual-decoder for RGB-thermal salient object detection

[J]. IEEE Transactions on Image Processing, 2021, 30: 5678-5691.
[3] FENG Zhanxiang, LAI Jianhuang, XIE Xiaohua. Learning modality-specific representations for visible-infrared person reidentification[J]. IEEE Transactions on Image Processing, 2020, 29: 579-590.
[4] WANG Zhishe, XU Jiawei, JIANG Xiaolin, et al. Infrared and visible image fusion via hybrid decomposition of NSCT and morphological sequential toggle operator[J]. Optik, 2020, 201(1) : 163497.
[5] JIANG Zetao, JIANG Qi, HUANG Yongsong, et al. Infrared and low-light-level visible light enhancement image fusion method based on latent low-rank representation and composite filtering[J]. Acta Photonica Sinica, 2020, 49(4): 0410001.
江泽涛, 蒋琦, 黄永松, 等. 基于潜在低秩表示与复合滤波的红外与弱可见光增强图像融合方法[J]. 光子学报, 2020, 49(4): 0410001.
[6] MA Jinlei, ZHOU Zhiqiang, WANG Bo, 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.
[7] KONG Weiwei, LEI Yang, ZHAO Huaixun. Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization[J]. Infrared Physics & Technology, 2014, 67: 161-172.
[8] LV Sheng, YANG Fengbao, JI Linna, et al. Combination fusion of multi-types mimic variables of infrared intensity and polarization image[J]. Infrared and Laser Engineering, 2018, 47(5): 504005.
吕胜, 杨风暴, 吉琳娜,等. 红外光强与偏振图像多类拟态变元组合融合[J]. 红外与激光工程, 2018, 47(5): 504005.
[9] ZHANG Hang, XU Han, TIAN Xin, et al. Image fusion meets deep learning: a survey and perspective[J]. Information Fusion, 2021, 76: 323-336.
[10] LI Hui, WU Xiaojun. DenseFuse: a fusion approach to infrared and visible images[J]. IEEE Transactions on Image Processing, 2019, 28(5): 2614-2623.
[11] ZHANG Yu, LIU Yu, SUN Peng, et al. IFCNN: a general image fusion framework based on convolutional neural network[J]. Information Fusion, 2020, 54: 99-118.
[12] JIAN Lihua, YANG Xiaomin, LIU Zheng, et al. SEDRFuse: A symmetric encoder-decoder with residual block network for infrared and visible image fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70:1-15.
[13] WANG Zhishe, WANG Junyao, WU Yuanyuan, et al. UNFusion: a unified multi-scale densely connected network for infrared and visible image fusion[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021. DOI: 10.1109/TCSVT.2021,3109895.
[14] LI Hui, WU Xiaojun, KITTLERB J. RFN-Nest: an end-to-end residual fusion network for infrared and visible images [J]. Information Fusion, 2021, 73: 72-86.
[15] XU Han, MA Jiayi, JIANG Junjun, et al. U2fusion: a unified unsupervised image fusion network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(1):502-518.
[16] ZHANG Hao, XU Han, XIAO Yang, et al. Rethinking the image fusion: a fast unified image fusion network based on proportional maintenance of gradient and intensity[C]. Proceedings of AAAI Conference on Artificial Intelligence, 2020, 34(7): 12797-12804.
[17] LONG Yongzhi, JIA Haitao, ZHONG Yida, et al. RXDNFuse: a aggregated residual dense network for infrared and visible image fusion[J]. Information Fusion, 2021, 69: 128-141.
[18] MA Jiayi, YU Wei, LIANG Pengwei, et al. FusionGAN: a generative adversarial network for infrared and visible image fusion[J]. Information Fusion, 2019, 48: 11-26.
[19] MA Jiayi, ZHANG Hang, SHAO Zhenfeng, 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: 1-14.
[20] TANG Lili, LIU Gang, XIAO Gang. Infrared and visible image fusion method based on dual-path cascade adversarial mechanism[J]. Acta Photonica Sinica,2021, 50(9): 0910004.
唐丽丽,刘刚,肖刚. 基于双路级联对抗机制的红外与可见光图像融合方法[J]. 光子学报, 2021, 50(9): 0910004.
[21] MA Jiayi, XU Han, JIANG Junjun, et al. DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion[J]. IEEE Transactions on Image Processing, 2020, 29: 4980-4995.
[22] FU Yu, WU Xiaojun, DURRANI T. Image fusion based on generative adversarial network consistent with perception[J]. Information Fusion, 2021, 72: 110-125.
[23] TOET A. TNO image fusion dataset[EB/OL](2018-09-15)[2021-11-22]. https://figshare.com/articles/TN Image Fusion Dataset/1008029.
[24] XU Han. Roadscene database [EB/OL][2021-11-22]. https://github.com/hanna-xu/RoadScene.

Infrared and Visible Image Fusion Method via Interactive Attention-based Generative Adversarial Network

WANG Zhishe[1], SHAO Wenyu[1], YANG Fengbao[2], CHEN Yanlin[1]
(1 *School of Applied Science, Taiyuan University of Science and Technology, Taiyuan* 030024, *China*)
(2 *School of Information and Communication Engineering, North University of China, Taiyuan* 030051, *China*)

Abstract: Infrared sensors can capture prominent target characteristics by thermal radiation imaging, however the obtained infrared images usually lack structural features and texture details. On the contrary, visible sensors can obtain rich scene information by light reflection imaging, the obtained visible images have high spatial resolution and rich texture details, but cannot effectively perceive target characteristics, especially in low illumination environmental conditions. Infrared and visible image fusion aims to integrate the advantages of the two types of sensors to generate a composite image with better target perception and superior scene representation, which is widely applied for object tracking, object detection and pedestrian re-recognition. The existing generative adversarial network-based fusion methods only make use of convolution operation to extract local features, but do not consider their long-range dependence, which is easy to cause the fusion imbalance, resulting in the fusion image cannot retain typical targets of infrared image and texture details of visible image at the same time. To this end, an end-to-end infrared and visible image fusion method via interactive attention-based generative adversarial network is proposed. Firstly, in the generative network model, we adopt a dual-path encoder architecture with weight parameters sharing to extract the respective multi-scale deep features of source images, where the first normal convolution layer is used to extract low-level features, and two multi-scale aggregation convolution models are adopted to extract high-level features. By aggregating multiple available receptive fields, our multi-scale dual-path encoder network can efficiently extract more meaningful information for fusion tasks without down-sample or up-sample operations. Secondly, in the fusion layer, we design an interactive attention fusion model, which is cascading channel and spatial attention models, to establish the global dependence of their local features from the channel and spatial dimensions. The obtained attention maps can refine multi-scale feature maps to more focus on typical infrared targets and visible texture details, so that the fused results achieve better visual results. Finally, in the adversarial network model, we propose two discriminators, such as Discriminator-IR and Discriminator-VIS, to balance the truth-falsity between fusion image and source images. Besides, we introduce the mutually-compensated loss function to supervise the entire network, which can gradually optimize the generative network model to obtain the best fused result. In the ablation study and verified experiments, the TNO and Roadscene datasets and eight evaluation metrics are proposed to demonstrate the effectiveness and superiority of the proposed method. The ablation experimental results of the interactive attention fusion model indicate that our model can effectively establish the global dependency of local features compared with other four models, and further improve infrared and visible image fusion performance. In addition, compared with other nine the state-of-the-art fusion methods, such as WLS, DenseFuse, IFCNN, SEDRFuse, U2Fusion, PMGI, FusionGAN, GANMcC and RFN-Nest, the proposed method can achieve more balanced fusion results in retaining the typical targets of infrared image and rich texture details of visible image, and has a better visual effect, which is more suitable for the human visual system. Meanwhile, from a multi-index evaluation perspective, the proposed method has better image fusion performance, higher computational efficiency and stronger robustness than other state-of-the-art fusion methods.

Key words: Image fusion; Interactive attention; Generative adversarial network; Deep learning; Infrared image; Visible image
OCIS Codes: 100.2000; 200.4260; 350.2260; 100.4996

Foundation item: The Fundamental Research Program of Shanxi Province (No. 201901D111260), the Open Foundation of Shanxi Key Laboratory of Signal Capturing & Processing (No. ISPT2020-4)