DOI: 10.13382/j.jemi.B2407197

结合语义分割与跨模态差分特征补偿的红外与可见光图像融合方法*

李天放　孙一宸　于明鑫　董明利
（北京信息科技大学仪器科学与光电工程学院　北京　100192）

摘　要：针对现有红外与可见光图像融合模型在深层特征提取时细节信息丢失、显著目标轮廓模糊的问题，提出一种结合语义分割与跨模态差分特征补偿（CMDFC）的红外与可见光图像融合方法。通过具有卷积注意力机制（CBAM）的跨模态差分特征补偿模块，叠加不同模态的互补特征信息至原始特征中进行深层特征提取，引入语义分割网络对融合图像进行像素级别的分类操作构造语义损失来约束融合网络，并使用解码器重构融合图像。在公开数据集上进行融合实验的结果表明，相较于对照模型中的最优指标，所选的 5 种指标均有不同程度的提高，其中互信息（MI）和视觉信息保真度（VIF）分别提高了 4.41%和 4.25%，说明本文所提出的模型生成的融合图像更清晰，与源图像相关性更强，该方法有效缓解了红外与可见光图像融合过程中特征细节信息丢失的问题，增强了生成图像的视觉效果和对比度。

关键词：图像融合；语义分割；注意力机制；跨模态差分特征补偿
中图分类号：TP391　　**文献标识码**：A　　**国家标准学科分类代码**：520.60

Infrared and visible image fusion method integrating semantic segmentation and cross-modality differential feature compensation

Li Tianfang　Sun Yichen　Yu Mingxin　Dong Mingli
(School of Instrument Science and Opto-Electronics Engineering, Beijing Information Science and Technology University, Beijing 100192, China)

Abstract: To address the issues of detail information loss and blurred salient target contours in existing infrared and visible image fusion models during deep feature extraction, we propose an infrared and visible image fusion method that combines semantic segmentation with cross-modality differential feature compensation (CMDFC). By incorporating a cross-modality differential feature compensation module with a convolutional block attention module (CBAM), complementary features from different modalities are integrated into the original features for deep feature extraction. Additionally, a semantic segmentation network is introduced to perform pixel-level classification on the fused image, constructing a semantic loss to constrain the fusion network, and a decoder is used to reconstruct the fused image. Experimental results on public datasets show that compared to the best metrics of the reference models, the proposed model achieves various degrees of improvement in five selected metrics, with mutual information (MI) and visual information fidelity (VIF) increased of 4.41% and 4.25%, respectively. These results indicate that the proposed model generates clearer fused images with stronger correlation to the source images, effectively mitigating the issue of feature detail loss during the fusion process and enhancing the visual quality and contrast of the generated images.

Keywords: image fusion; semantic segmentation; attentional mechanism; cross-modality differential feature compensation

收稿日期：2024-01-12　　Received Date: 2024-01-12
*基金项目：北京市教委科技计划一般项目（KM202011232007）、高校学科人才引进计划（D17021）、北京信息科技内涵发展项目（2019KYNH204）资助

0　引　言

受不同传感器成像机理的差异影响与理论方面的限制,从单一传感器捕获到的图像中很难获取目标场景的全面、完整的信息[1]。红外传感器通过捕捉目标发出的热辐射能够实现全天候工作,但生成的红外图像分辨率较低,容易受到噪声的干扰,且日间图像细节模糊,边缘不明显,缺乏纹理细节。可见光传感器在正常光照条件下能够显示物体清晰、丰富的纹理信息,具有高对比度和细致且真实的外观,但对环境较为敏感,易受恶劣天气与光照强度的影响。为了尽可能满足不同场景下信息获取的需求,图像融合技术得到了发展。相比于对纹理细节敏感的可见光图像,红外图像对热源目标敏感,图像融合技术利用可见光和红外图像的互补性,取长补短,以提高图像的视觉感知能力。这种图像在恶劣天气和暗淡环境中能够突出目标并丰富细节信息,提供更可靠、更准确的场景信息。

在过去几十年里,图像融合技术的发展使得图像视觉质量得到提高,主要分为传统方法和基于深度学习的方法两大类。传统方法通常是在空间域直接对像素进行操作,或者将图像转换到变换域,利用变换域的特定性质进行特征提取和图像融合[2]。根据理论不同,传统图像融合技术可分为基于多尺度变换的方法[3]、基于稀疏表示的方法[4]、基于子空间的方法[5]、基于优化的方法[6]以及基于显著性的方法[7-8]。尽管传统图像融合方法已发展多年,但由于其对先验知识的依赖,在未知场景下难以保持良好的效果。此外,随着融合规则变得越来越复杂,手工设计方法变得繁琐复杂,使得这些方法在今后的发展中难以有更大的提升。在基于深度学习的方法中,具有代表性的方法主要分为以下 3 类:基于自编码器(auto-encoder, AE)的方法、基于卷积神经网络(convolutional neural network, CNN)的方法以及基于生成对抗网络(generative adversarial network, GAN)的方法。

近年来,基于深度学习的图像融合技术飞速发展,并能生成令人满意的融合图像。在文献[9]中,多尺度密集连接被利用以实现特征提取后的复用,然而在融合过程中可能导致细节丢失。文献[10-11]使用多尺度编码器-解码器架构和嵌套连接,提高了性能,但仍未摆脱需要手工设计规则的缺点。针对这一问题,文献[12]提出一种基于卷积神经网络的图像融合模型,成功解决了手工设计规则的问题。此外,文献[13]将生成对抗网络引入图像融合中,同样实现了摆脱手工设计规则的目的。然而,由于单一鉴别器会破坏红外和可见光数据之间的平衡,导致红外图像细节的丢失。另一方面,文献[14]用像素强度衡量源图像的显著性特征,但是网络结构单一,导致信息丢失。为了解决这一问题,文献[15]引入了语义分割网络以增强融合图像中的语义信息,从而满足生成图像用于后续视觉任务的需求。

综上所述,在融合深层特征时,若采用了不恰当的融合规则,则会造成红外和可见光图像的特征信息丢失,使生成图像缺失语义信息,进而出现图像信息不完整、细节模糊、清晰度不佳、视觉感知质量低等情况。针对上述问题,提出一种结合语义分割与跨模态差分特征补偿的红外与可见光图像融合方法(SSCM-Fusion),主要工作有:

1)使用跨模态差分特征补偿模块提取红外和可见光图像的差分信息,同时引入注意力机制以提升模型对图像细节的保留能力;

2)引入语义分割网络,通过语义损失将高级语义信息回馈到融合网络,以实现对模型的约束;

3)采用联合训练策略,对融合网络和语义分割网络进行迭代训练,以增强语义分割网络对融合网络的拟合效果。

1　SSCM-Fusion 网络模型

1.1　网络总体框架

模型的整体结构如图 1 所示。给定一组配准的红外图像 $\boldsymbol{I}_{ir} \in \boldsymbol{R}^{H\times W\times 1}$、彩色可见光图像 $\boldsymbol{I}_{vi} \in \boldsymbol{R}^{H\times W\times 3}$ 和语义标签 $\boldsymbol{L}_s \in (1,\boldsymbol{C})^{H\times W}$,其中 H 和 W 分别表示图片的高度和宽度,$\boldsymbol{C}$ 表示对象类别的数量。为了保留融合图像中的颜色信息,首先将可见光图像转换到 YCbCr 色彩空间,再对两种图像的 Y 通道进行融合,最后将生成图像与可见光图像的 Cb 和 Cr 通道级联,还原为 RGB 色彩图像。使用语义分割网络对生成图像进行分割,并结合语义标签计算语义损失。在迭代训练过程中,对融合网络和语义分割网络进行优化,同时利用语义损失约束分割网络。融合网络在内容损失和语义损失的指导下,通过特征提取、特征融合和图像重构实现融合图像的生成,图 2 展示了融合网络的结构。

为了使生成的融合图像呈现更加丰富的纹理细节,本文引入了一种渐进式融合网络,以融合互补信息和公共信息,从而增加特征的复用[16]。首先,采用特征编码器 E_F 分别对红外图像和可见光图像进行深层特征的提取,如式(1)所示。

$$\{\boldsymbol{F}_{ir},\boldsymbol{F}_{vi}\} = \{E_F(\boldsymbol{I}_{ir}),E_F(\boldsymbol{I}_{vi})\} \tag{1}$$

式中:$\boldsymbol{F}_{ir}$ 和 $\boldsymbol{F}_{vi}$ 分别表示红外特征和可见光特征。此外,$\boldsymbol{F}_{ir}^{i}$ 和 $\boldsymbol{F}_{vi}^{i}$ 被定义为第 i 层卷积层从红外图像和可见光图像中提取到的特征。

随后,在特征编码器中引入跨模态差分特征补偿(cross-modality differential feature compensation, CMDFC)

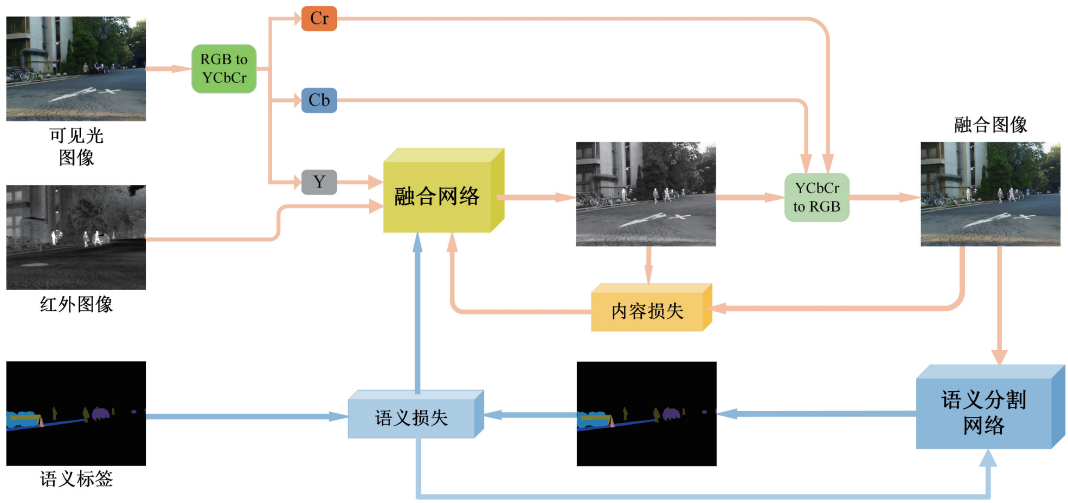


图 1 结合语义分割和跨模态差分特征补偿的红外和可见光图像融合框架

Fig. 1 An infrared and visible image fusion framework combining semantic segmentation and cross-modality differential feature compensation

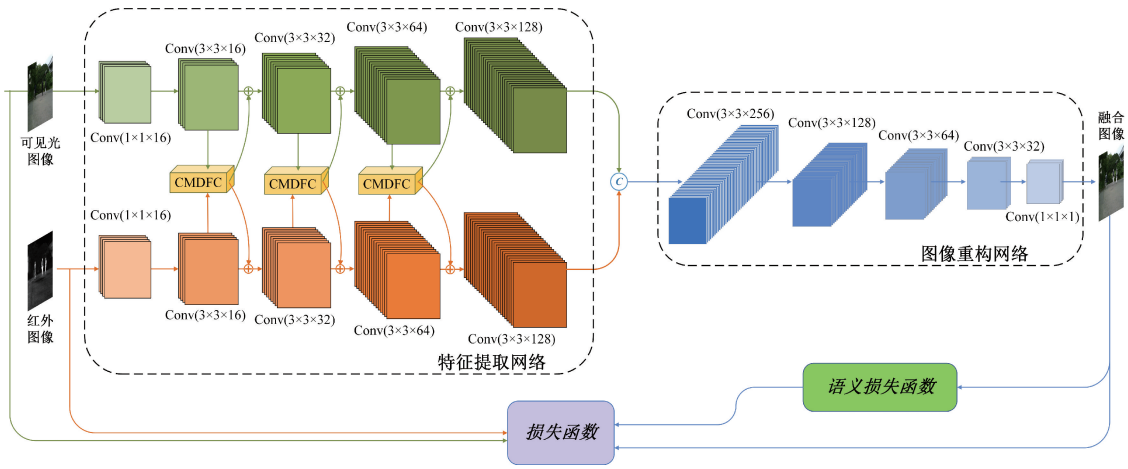


图 2 融合网络结构

Fig. 2 Fusion network structure

模块,将逐层提取到的特征 $\boldsymbol{F}_{ir}^{i}$ 与 $\boldsymbol{F}_{vi}^{i}$ 以渐进的形式进行差分信息的补偿。$\boldsymbol{F}_{ir}^{i}$ 与 $\boldsymbol{F}_{vi}^{i}$ 可以用公共部分和互补部分表示,如式(2)和(3)所示。

$$\boldsymbol{F}_{ir}^{i}=\frac{\boldsymbol{F}_{ir}^{i}+\boldsymbol{F}_{ir}^{i}}{2}+\frac{\boldsymbol{F}_{vi}^{i}-\boldsymbol{F}_{vi}^{i}}{2}=\frac{\boldsymbol{F}_{ir}^{i}+\boldsymbol{F}_{vi}^{i}}{2}+\frac{\boldsymbol{F}_{ir}^{i}-\boldsymbol{F}_{vi}^{i}}{2} \tag{2}$$

$$\boldsymbol{F}_{vi}^{i}=\frac{\boldsymbol{F}_{vi}^{i}+\boldsymbol{F}_{vi}^{i}}{2}+\frac{\boldsymbol{F}_{ir}^{i}-\boldsymbol{F}_{ir}^{i}}{2}=\frac{\boldsymbol{F}_{vi}^{i}+\boldsymbol{F}_{ir}^{i}}{2}+\frac{\boldsymbol{F}_{vi}^{i}-\boldsymbol{F}_{ir}^{i}}{2} \tag{3}$$

将等式化作两个部分相加,其中左侧部分表示公共部分,即不同模态反映的相同特征,右侧部分为互补部分,即不同模态反映的互补特征。跨模态差分特征补偿的目标是充分融合互补的特征信息和信道权重。具体地,其定义如式(4)和(5)所示。

$$\hat{\boldsymbol{F}}_{ir}^{i}=\boldsymbol{F}_{ir}^{i}\oplus A(\delta(GAP(\boldsymbol{F}_{vi}^{i}-\boldsymbol{F}_{ir}^{i}))\odot(\boldsymbol{F}_{vi}^{i}-\boldsymbol{F}_{ir}^{i})) \tag{4}$$

$$\hat{\boldsymbol{F}}_{vi}^{i}=\boldsymbol{F}_{vi}^{i}\oplus A(\delta(GAP(\boldsymbol{F}_{ir}^{i}-\boldsymbol{F}_{vi}^{i}))\odot(\boldsymbol{F}_{ir}^{i}-\boldsymbol{F}_{vi}^{i})) \tag{5}$$

式中:$\oplus$ 表示元素加法,$\odot$ 表示通道相乘,$\delta(\cdot)$ 和 $GAP(\cdot)$ 分别表示 Sigmoid 激活函数和全局平均池化,$A(\cdot)$ 表示卷积注意力机制。

首先利用全局平均池化将互补特征压缩为一个向量,再通过 Sigmoid 函数将向量归一化到 [0,1] 区间并生成信道权重。然后将互补特征与信道权值相乘,最终将结果作为模态补充信息添加到原始特征中。因此,特

征编码器可以利用 CMDFC 模块提取和预集成不同模态的互补特征。将补偿过差分信息的红外图像与可见光图像特征，通过通道维拼接互相融合，如式(6)所示。

$$\boldsymbol{F}_f = C(\boldsymbol{F}_{ir}, \boldsymbol{F}_{vi}) \tag{6}$$

式中：$C(\cdot)$ 表示通道维度上的级联。

最后，利用图像重构器 R_1 对融合图像特征信息 $\boldsymbol{F}_f$ 进行重构，恢复融合后的图像 $\boldsymbol{I}_f$，如式（7）所示。

$$\boldsymbol{I}_f = R_1(\boldsymbol{F}_f) \tag{7}$$

1.2　特征提取编码器

本文的特征编码器主要分为特征提取部分和跨模态差分特征补偿部分。特征提取部分包括 5 个卷积层，旨在提取红外图像和可见光图像的深层特征。第 1 层采用 1×1 卷积核的目的是增加特征图的通道数，为接下来的特征提取过程提供更多的特征表达空间，减小红外图像与可见光图像的模态差异。在后 4 层使用 3×3 卷积核进行特征提取，并在第 2、3、4 层的输出后接入 CMDFC 模块，这样，CMDFC 模块在提取红外和可见光图像的公共特征和互补特征的同时，能够逐层、渐进式地交换两者的互补信息。特征提取编码器的各卷积层的详细信息如表 1 所示，各卷积层均使用 Leaky ReLU 作为激活函数，缓解梯度消失问题，避免神经元“死亡”。CMDFC 模块如图 3 所示。

表 1　渐进式融合网络中各卷积层的相关参数

Table 1　Relevant parameters of each convolutional layer in progressive fusion network

卷积层	特征提取编码器			图像重构器		
	卷积核	输出通道数	激活函数	卷积核	输出通道数	激活函数
第一层	1×1	16	Leaky ReLU	3×3	256	Leaky ReLU
第二层	3×3	16	Leaky ReLU	3×3	128	Leaky ReLU
第三层	3×3	32	Leaky ReLU	3×3	64	Leaky ReLU
第四层	3×3	64	Leaky ReLU	3×3	32	Leaky ReLU
第五层	3×3	128	Leaky ReLU	1×1	1	Tanh

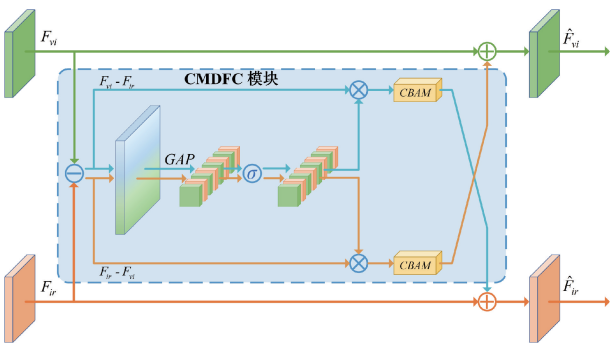


图 3　跨模态差分特征补偿模块

Fig. 3　Cross-Modal differential feature compensation module

在跨模态差分特征补偿模块中，引入了注意力机制[17]（convolutional block attention module，CBAM），如图 4 所示。

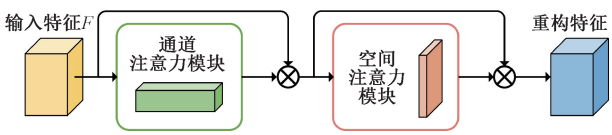


图 4　CBAM 注意力机制模块

Fig. 4　Convolutional block attention module

CBAM 是一种结合了通道注意力和空间注意力的注意力机制，可以自适应地调整特征图的通道和空间维度的权重从而减少图像纹理细节的丢失。

在获取特征编码器输出的红外与可见光特征图像后，对其进行特征级别的融合。利用通道维度的级联操作，实现不同模态的特征信息最大程度的保留，生成融合特征图像以便于后续的图像重构操作。

1.3　图像重构器

图像重构器包含 5 层卷积层，通过对融合特征图像进行操作，实现深层特征的整合和融合图像的生成，如图 5 所示。前 4 层采用 3×3 的卷积核和 Leaky ReLU 激活函数，有助于在特征映射通道数逐步减少的过程中保留更多的图像细节。最后一层使用 1×1 卷积核和 Tanh 激活函数，将特征映射转换为具有相同维度范围的像素值，使得生成的融合图像更加自然。

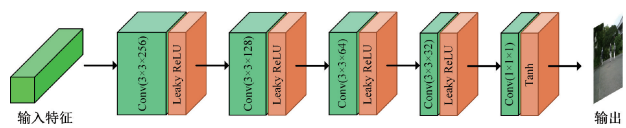


图 5　图像重构器

Fig. 5　Image reconstructor

为确保生成图像与原始图像尺寸相同，整个网络没有使用下采样操作，进而避免特征信息丢失，同时更好地保留原始图像的结构和细节。图像重构器各卷积层的详细信息如表 1 所示。

1.4　语义分割模型

语义分割模型引用自 PENG 提出的双边注意力解码器[18]，一种用于实时语义分割的轻量级解码器。从通道注意力层次和空间注意力层次分别对图像的高级特征映射与低级特征映射进行细化，再通过高效的池化融合块，充分利用高、低层特征图的优势实现两者的融合，得到同

时具有准确语义和空间信息的融合特征图。语义分割模型具体结构如图 6 所示。

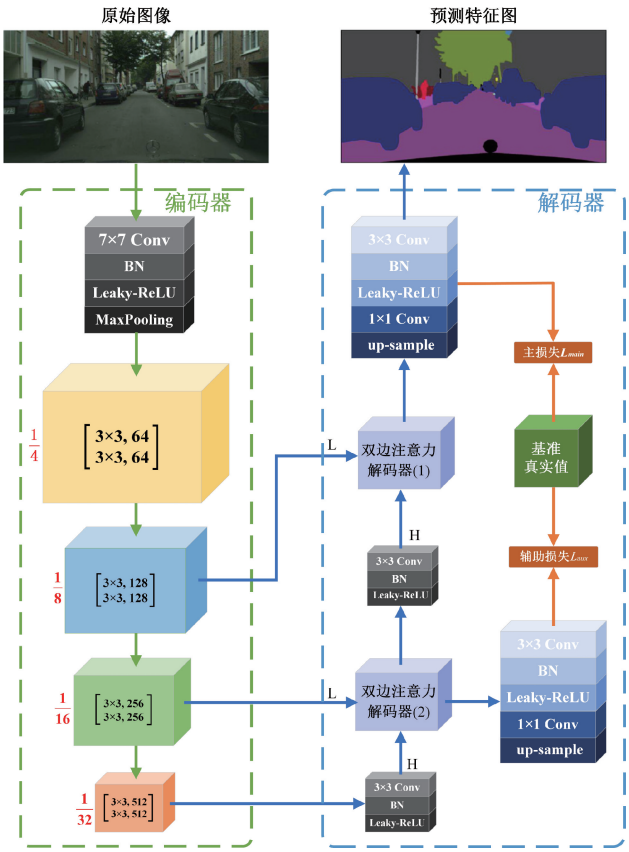


图 6 语义分割模型

Fig. 6 Semantic segmentation model

图 6 中 H 表示高级特征图，L 表示低级特征图，BN 表示批量归一化，Up-sample 表示上采样操作。输入图像送入编码器提取特征，每一块表示对输入图像进行卷积操作，经过 4 个块后输出的高级特征映射分辨降低为原来的 1/32。使用包括批量归一化和 Leaky-ReLU 激活函数的卷积层对第 4 块输出特征进行处理，通过双边注意力解码器（bilateral attention decoder）将其与第 3 块的输出相融合。将改进的融合结果再以相同方式与第 2 块输出特征结合得到原输入大小为 1/8 的输出特征图。进一步对输出的特征图进行细化处理，再经 1×1 卷积核和上采样操作，得到输出通道数等于类别数的输出特征图。此外，为了保证网络的鲁棒性，构造了损失函数监督解码器的输出，如式(8)所示。

$$L_{total} = L_{main} + \lambda \times L_{aux} \tag{8}$$

式中：L_{main} 和 L_{aux} 分别表示双边注意力解码器(1)和双边注意力解码器(2)的输出与基准真实值的交叉熵，λ 表示平衡两交叉熵的权重。

双边注意力解码器如图 7 所示。将低阶和高阶特征映射输入进双边注意力解码器，分别通过空间和通道注意力分支进行细化。随后，将细化后的低阶特征映射进行卷积操作控制其通道数，对细化后的高阶特征映射使用上采样操作还原分辨率。最后将处理后的高、低阶特征映射输入进池化融合块，得到具有准确语义和空间信息的输出特征映射。使用平均池化对上采样后的高阶特征映射进行平滑处理，再与低阶特征映射进行通道级联，得到最终输出结果。

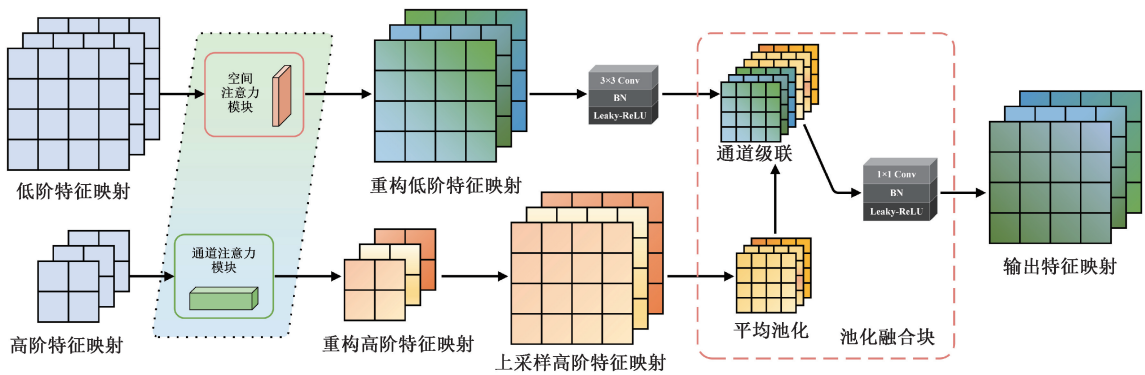


图 7 双边注意力解码器结构

Fig. 7 The bilateral attention decoder structure

1.5 损失函数

为了提高生成融合图像的视觉质量和评估指标，本文从两方面设计损失函数，共同对渐进融合网络进行约束。一方面，为了保证融合图像的视觉保真度，设计了内容损失；另一方面，构建语义损失，使用语义分割网络评估融合图像分割效果从而约束融合模型。

1)内容损失函数

内容损失函数的确切定义如式(9)所示。

$$L_{content} = \lambda_1 L_{int} + \lambda_2 L_{texture} + \lambda_3 L_{ssim} \tag{9}$$

式中：L_{int} 表示强度损失，负责约束融合图像的整体强度；$L_{texture}$ 表示纹理损失，使融合图像包含更多纹理细节；L_{ssim} 表示结构相似度损失，使融合图像尽可能与源图像相似。

λ_1、λ_2 和 λ_3 为超参数,用于平衡 3 种损失值的权重。

为了保持融合图像的最优强度分布,引入强度损失测量融合图像与源图像之间的差异,具体如式(10)所示。

$$L_{int} = \frac{1}{HW} \| \boldsymbol{I}_f - \max(\boldsymbol{I}_{ir}, \boldsymbol{I}_{vi}) \|_1 \tag{10}$$

式中:H 和 W 分别表示输入图像的高和宽,$\boldsymbol{I}_f$、$\boldsymbol{I}_{ir}$ 和 $\boldsymbol{I}_{vi}$ 分别表示融合图像、红外图像和可见光图像的特征,$\| \cdot \|_1$ 表示 L1 范数,$\max(\cdot)$ 表示逐个对元素进行最大选择。利用最大选择策略,实现红外和可见光图像最大像素强度的整合。

优秀的模型在确保融合图像具有足够强度分布的同时,还会保留源图像丰富的纹理细节,然而单一的强度损失只能提供粗略的分布约束,无法满足对纹理细节的需求。若想使融合图像包含更多细粒度的纹理信息,则需要引入纹理损失对模型进行约束。纹理损失的定义如式(11)所示。

$$L_{texture} = \frac{1}{HW} \| \, | \nabla \boldsymbol{I}_f | - \max(| \nabla \boldsymbol{I}_{ir} |, | \nabla \boldsymbol{I}_{vi} |) \|_1 \tag{11}$$

式中:∇ 表示 Sobel 梯度算子,用于测量图像的细粒度纹理信息,$|\cdot|$ 表示绝对操作。

为了尽可能使融合图像与源图像更加相似,使用结构相似指标度量(structural similarity index measure, SSIM),根据光线、对比度和结构信息的相似性对模型进行约束,结构相似度损失如式(12)所示。

$$L_{ssim} = \lambda \cdot \frac{1 - SSIM(\boldsymbol{I}_f, \boldsymbol{I}_{ir})}{HW} + (1 - \lambda) \cdot \frac{1 - SSIM(\boldsymbol{I}_f, \boldsymbol{I}_{vi})}{HW} \tag{12}$$

式中:$SSIM(\boldsymbol{I}_x, \boldsymbol{I}_y)$ 表示 $\boldsymbol{I}_x$ 与 $\boldsymbol{I}_y$ 之间的结构相似度,λ 为超参数,负责维持红外与可见光图像结构的平衡。

综上所述,基于跨模态差分特征补偿的融合网络可以在内容损失的约束下,在保留输入图像最优强度分布的同时,在生成图像中融入更多的纹理细节,使得融合后图像与源图像更加相似。因此,内容损失可以指导模型训练实现融合图像视觉质量的提高。

2)语义损失函数

语义分割将图像中的每个像素按照预定义的语义类别进行分类,为每个像素都赋予一个语义标签。通过引入实时语义分割模型[18]对融合后的图像进行分割,分割网络输出分割结果 $\boldsymbol{I}_s \in \mathbb{R}^{H \times W \times C}$ 和辅助分割结果 $\boldsymbol{I}_{sa} \in \mathbb{R}^{H \times W \times C}$。主要语义损失和辅助语义损失的定义如式(13)和(14)所示。

$$L_{main} = \frac{-1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} \boldsymbol{L}_{so}^{(h,w,c)} \log(\boldsymbol{I}_s^{(h,w,c)}) \tag{13}$$

$$L_{aux} = \frac{-1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} \boldsymbol{L}_{so}^{(h,w,c)} \log(\boldsymbol{I}_{sa}^{(h,w,c)}) \tag{14}$$

式中:$\boldsymbol{L}_{so} \in \mathbb{R}^{H \times W \times C}$ 表示从分割标签 $\boldsymbol{L}_s \in (1, \boldsymbol{C})^{H \times W}$ 变换过来的独热编码。

主语义损失和辅助语义损失从不同角度反映了融合图像包含的语义信息,最终的语义损失定义如式(15)所示。

$$L_{semantic} = L_{main} + \alpha L_{aux} \tag{15}$$

式中:α 为超参数,用于平衡主语义损失和辅助语义损失,参考语义分割模型[18]设为 0.1。

最后,利用内容损失和语义损失构造联合损失函数来约束融合模型训练,具体定义如式(16)所示。

$$L_{fusion} = L_{content} + \beta L_{semantic} \tag{16}$$

式中:β 为表征语义损失重要性的超参数。同时,β 并不是一个固定的值,随着训练的进行,语义分割网络会逐渐适应融合网络并呈现出更好的拟合效果。

为了更好的维持两者性能的平衡,通过联合训练策略训练融合网络。具体来说,对融合网络和分割网络进行迭代训练,并设置迭代为 m,利用 Adam 优化器对融合网络的所有参数进行优化,同时根据迭代次数动态调整 β 超参数,具体表示如式(17)所示。

$$\beta = \gamma \times (m - 1) \tag{17}$$

式中:m 表示第 m 次迭代,超参数 γ 负责平衡内容损失与语义损失。随着迭代次数的增加,语义分割网络对融合模型的拟合效果会越来越好,语义损失也随之能够更好的指导融合网络的训练,β 也相应的会逐渐增大。根据每次迭代后生成的融合结果,通过优化语义损失来更新分割模型的参数,实现模型性能的最优化。

2　实验与结果分析

2.1　实验数据集介绍

1)实验训练集

本实验采用的训练集由 MSRS 数据集挑选出的 187 张所组成。多光谱道路场景(multi-spectral road scenarios, MSRS)数据集是在 MFNet 数据集基础上,由 TANG 发布的一个新的数据集[19],通过剔除未对齐的图像对,并对其中低信噪比和低对比度的图像对使用图像增强算法,从而得到高质量的红外和可见光对齐图像。同时,MFNet 数据集提供了 9 个对象的语义标签,即汽车、人、自行车、弯道、站牌、护栏、锥形路标和背景,因此,MSRS 数据集同样包含这 9 个对象的语义标签。

本文在 MSRS 数据集基础上,均匀的选取了包括白天和黑夜的 187 对分辨率为 480×640 的图像对,将这些图像裁剪尺寸为 64×64 共 13 090 张图像对,并利用这些图像来训练渐进融合模型的参数。同样的,选取并裁剪

出 13 090 个语义标签用于语义分割网络的训练。所有图像在输入网络之前都归一化为 [0,1]。

2)实验测试集

为了进一步评估模型在不同场景下的综合表现,本文将实验测试集分为“均衡图像”、扩展组和综合测试数据集共 3 个组。“均衡图像”的红外和可见光图像均包含较丰富的细节信息,能够实现一组图表现各模型不同特征的目的;扩展组具有更复杂的场景数据,包括夜间较大光强的数据、阴影较多的数据、语义标签复杂的数据等;综合测试数据集由 MSRS 测试数据集中的 64 组典型图片组成,用来衡量模型的平均性能。

2.2 参数选择

本实验依次对渐进融合网络和语义分割网络进行迭代训练。在训练第 1 轮融合网络时不引入语义损失,只使用内容损失来约束网络模型。在生成第 1 轮融合图像后,通过对生成图像使用语义分割模型得到语义损失,约束后续融合网络的迭代训练。迭代次数 $m = 4$,每次迭代融合网络 $epoch = 10$,批处理量 $batch\ size = 128$,模型参数使用 Adam 优化器更新,学习率初始化为 0.001,然后呈指数衰减。对于式(9)内容损失函数的超参数,设置为 $\lambda_1 = 8$,$\lambda_2 = 55$,$\lambda_3 = 15$。语义分割训练方面,每轮迭代次数 $iter_num = 10\ 000$,最大迭代次数 $max_iter = 40\ 000$,利用小批量随机梯度下降法(mini-batch stochastic gradient descent),令批处理量 $batchsize = 16$ 来优化语义分割网络。将初始学习率设为 0.01,使用学习率预热和学习率衰减调整策略,提高训练的稳定性与性能。平衡内容损失和语义损失的超参数 $\gamma = 1$。

本文实验环境在 NVIDIA GTX 1080 Ti GPU 上进行,编程语言为 Python3.8,深度学习框架为 Pytorch1.9,硬件配置环境为 Intel(R) Core(TM) i7-10700K CPU 3.80 GHz,对比实验采用的软硬件环境相同。

2.3 定性分析

为了直观地展示基于语义分割和注意力机制的融合算法的融合效果,分别对“均衡图像”和扩展组进行主观评价,并对本文提出的方法进行全面评估,选择了 6 种典型融合模型与本文提出的模型进行比较。其中包括密集连接网络(DenseFuse)[20]、基于多尺度特征集成与边缘注意引导的图像融合(MFEIF)[21]、基于光照感知的渐进式红外与可见光图像融合网络(PIAFusion)[19]、基于语义感知的实时红外和可见光图像融合网络(SeAFusion)[15]、双分支残差网络(Dualbranch)[22]、统一无监督图像融合网络(U2Fusion)[23]。本文同时对图像中的细节部分进行框选并放大,进一步分析各模型特征,“均衡图像”的可视化结果如图 8 所示。

由图 8 可知,DenseFuse、Dualbranch 和 U2Fusion 生成的融合图像,对比度较低,整体偏暗。MFEIF 在保留可见光图像方面效果不佳,图像左右两侧阴影处模糊不清,缺少纹理细节。框选中的垃圾桶放大到左下方可以看出,U2Fusion 和 MFEIF 中的轮廓不明显,且框选靠上部分的红外图像纹理细节没有被保留下来。由浅色框可得,PIAFusion 和 SeAFusion 在显著目标信息的突出显示方面表现良好,但 PIAFusion 存在显著信息边缘不够平滑的问题。相较于 SeAFusion,SSCM-Fusion 在边缘细节方面表现的更好。

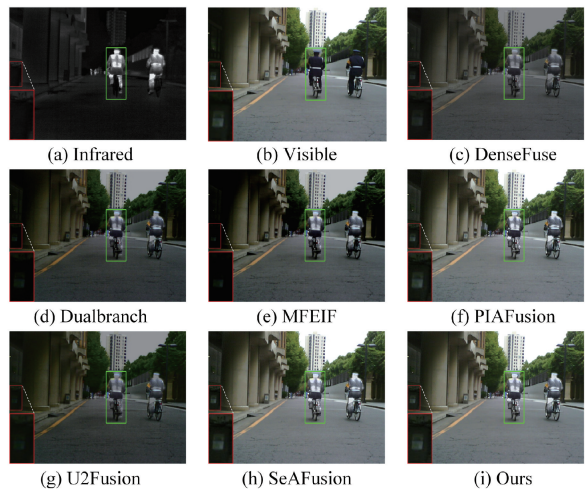


图 8 “均衡图像”的定性分析

Fig. 8 Qualitative analysis of “Balanced Image”

综合来看,SSCM-Fusion 模型能够针对不同环境的差异,更好的把握不同模态图像中纹理细节丰富的部分,使生成图像的边缘、轮廓等细节信息更明显,生成更符合人眼感观的图像。

为了更加直观的展示结合语义分割和跨模态差分特征补偿的红外和可见光图像融合方法的整体效果,本文继续对扩展组中 2 幅日间图像和 3 幅夜间图像进行可视化分析。如图 9~13 所示。

在日间场景下,由于可见光图像的纹理细节十分丰富,则红外图像主要作为补充信息参与融合,以增强阴影部分的细节特征。

对图 9 和 10 的框选内容进行放大,可以直观地看出,MFEIF 过分的增强了红外图像的权重,使阴影处模糊不清,无法保留可见光图像丰富的背景细节。DenseFuse、Dualbranch 和 U2Fusion 能够有效的结合可见光与红外图像的纹理细节,但是会减弱红外图像的显著目标信息,使红外特征变得模糊。而 PIAFusion 会在光照强度不足的情况引入额外的信息,SeAFusion 则在显著目标的边缘轮廓上表现的较为模糊,分别如图 9 地平线和 10 中人物所示。综上,只有本方法能够更好地提取并保留可见光图像中的纹理细节,平衡好融合过程中红外与可见光图像特征的关系。

图 9　SSCM-Fusion 在日间场景下的定性分析(00196D)

Fig. 9　Qualitative analysis of SSCM-Fusion on the daytime scene (00196D)

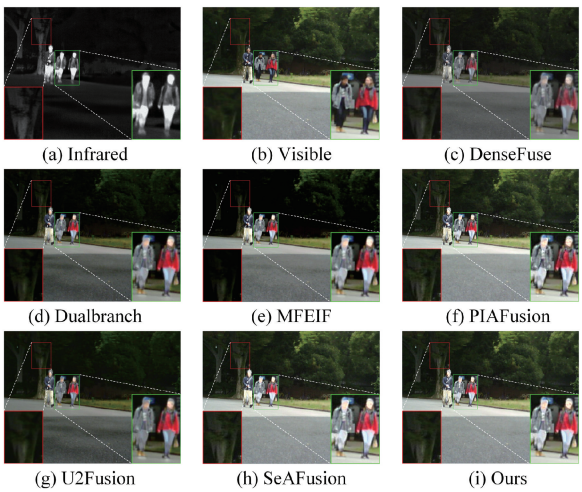


图 10　SSCM-Fusion 在日间场景下的定性分析(00581D)

Fig. 10　Qualitative analysis of SSCM-Fusion on the daytime scene (00581D)

在夜间场景下,随着光照强度的减弱,可见光图像包含的纹理细节逐渐减少。融合模型需要把握好夜间纹理细节丰富但分辨率较低的红外图像特征,使生成图像在保持较高对比度的同时,包含更多的局部特征信息。

由图 11 ~ 13 的框选内容可以观察到,DenseFuse、Dualbranch、MFEIF 和 U2Fusion 无法突出夜晚可见光图像中的背景细节,昏暗环境下的树枝轮廓和大楼外轮廓边缘十分模糊,整体亮度偏低,无法给人较好的视觉感受。而对于夜间图像较强的光源,如路灯、车灯,MFEIF、U2Fusion、SeAFusion、PIAFusion 会引入不同程度的光污染,PIAFusion 还会在融合过程中引入噪声,如图 11 ~ 13 中路灯与车灯所示。只有本方法可以在有效融合红外和可见光图像细节的同时,提高图像的整体亮度,突出显著目标信息,使生成图像更自然。

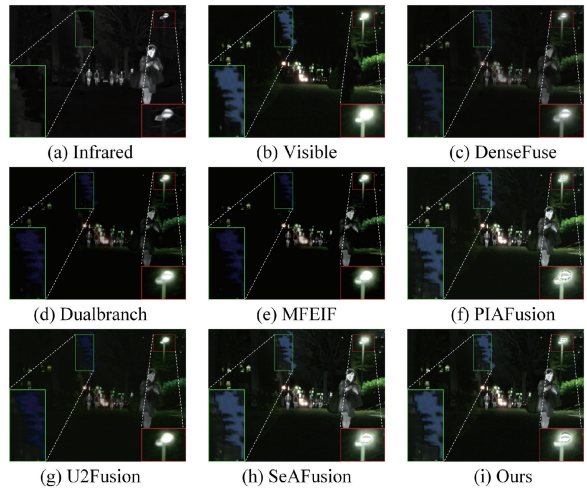


图 11　SSCM-Fusion 在夜间场景下的定性分析(00036N)

Fig. 11　Qualitative analysis of SSCM-Fusion on the nighttime scene (00036N)

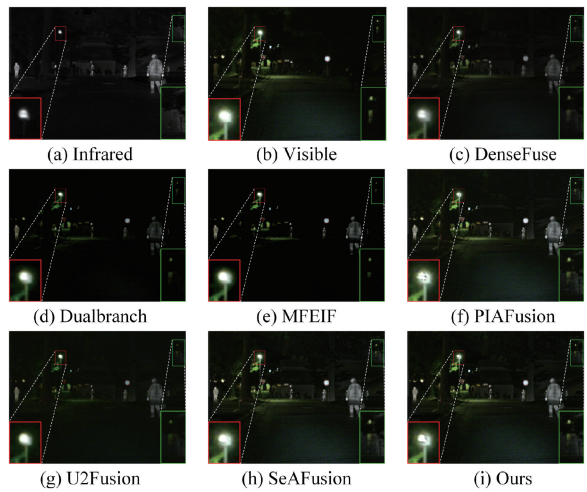


图 12　SSCM-Fusion 在夜间场景下的定性分析(00718N)

Fig. 12　Qualitative analysis of SSCM-Fusion on the nighttime scene (00718N)

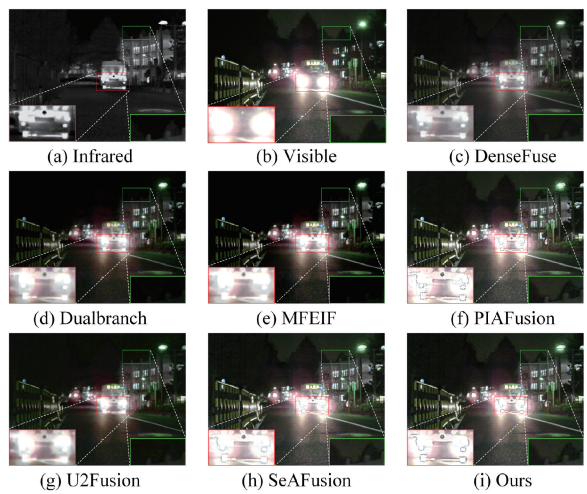


图 13　SSCM-Fusion 在夜间场景下的定性分析(00947N)

Fig. 13　Qualitative analysis of SSCM-Fusion on the nighttime scene (00947N)

2.4 定量分析

由于定性分析存在一定主观性,在评价融合图像效果时,不同的人得出的结论存在差异。因此,本文选择 5 种接受度比较广的客观评价指标,来对融合图像的质量进行分析和量化评价。互信息(mutual information, MI)度量融合图像获取源图像信息量的多少,空间频率(spatial frequency, SF)度量图像灰度的变化率,视觉信息保真度(visual information fidelity, VIF)从人类视觉感观的角度评价融合图像的信息保真度,基于梯度的融合性能(Q_{abf})度量从源图像传输到融合图像的边缘信息量,结构相似性指标(SSIM)衡量源图像与结构之间的结构相似性。MI、SF、VIF、Q_{abf}、SSIM 越大,表明融合模型能够提取并保留更多源图像的特征,说明融合效果越好。

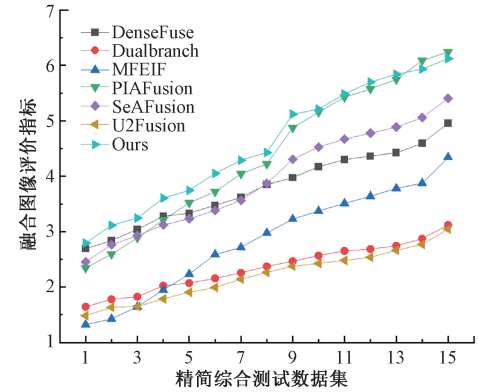
表 2 表示综合测试数据集中 64 组图像的 5 个客观评价指标的对比,其中黑体表示最优值,下划线表示次优值。如表 2 所示,本文所提出的模型在 5 项评价指标中,获得了 4 项最优和 1 项次优,其中次优值仅低于最优值不到 2%。MI 和 VIF 指标相较于对比模型中的最优指标分别提高了 4.41%和 4.25%。指标最佳的 MI 值表明该模型通过利用渐进融合网络,得以将源图像的大部分信息传输到生成的融合图像中。较高的 SF 值和 VIF 值表明生成的融合图像更清晰,具有更高的对比度和视觉效果。Q_{abf} 值则表明跨模态差分特征补偿模块的优异性和有效性,使生成的融合结果中包含了更多源图像的边缘信息。SSIM 值表明融合图像包含更多源图像的细节,与源图像更相似。

表 2 对综合测试数据集的 64 组图像的定量分析

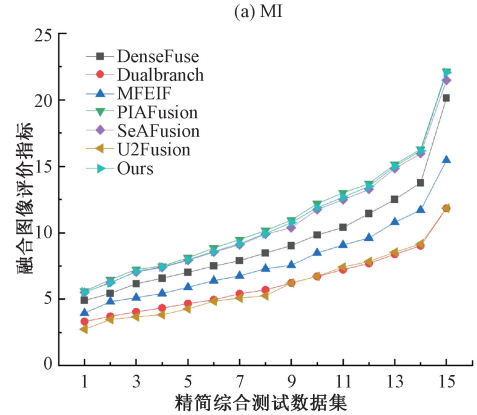
Table 2 Quantitative analysis of 64 sets of images from the comprehensive test dataset

模型	MI	SF	VIF	Q_{abf}	SSIM
DenseFuse	3.837	9.291	0.905	0.637	0.999
Dualbranch	2.377	6.244	0.666	0.375	0.896
MFEIF	2.903	7.838	0.738	0.534	0.640
PIAFusion	4.465	**11.083**	0.989	0.676	0.991
SeAFusion	3.989	10.711	0.987	0.668	0.994
U2Fusion	2.241	6.052	0.525	0.294	0.636
Ours	**4.662**	10.857	**1.031**	**0.686**	**1.013**

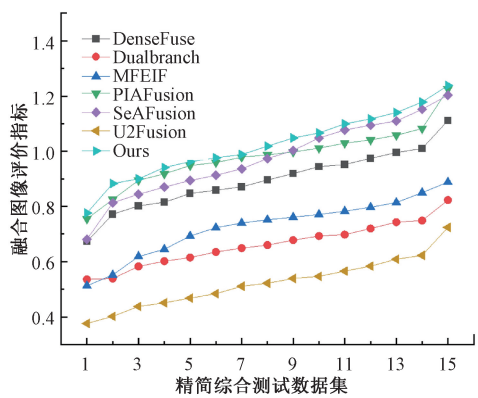
更进一步的,从综合测试数据集中再选取 15 张图片进行融合分析,精简后的数据集能够以折线图的形式将指标数据可视化,如图 14 所示。可以看出本文模型所求得的多个指标数据优于其他模型,其他指标数据则接近模型效果最好的其他模型。具体来说,SSCM-Fusion 模型在 VIF 指标上有优势,意味着这种方法可以使生成的融合图像具有更高的视觉保真度,能够得到满意的融合结果。在 Q_{abf}、SSIM 和 SF 值上,与 PIA 模型的表现相当,说明该模型从源图像中保留了足够的边缘信息,并与源图像足够相似。



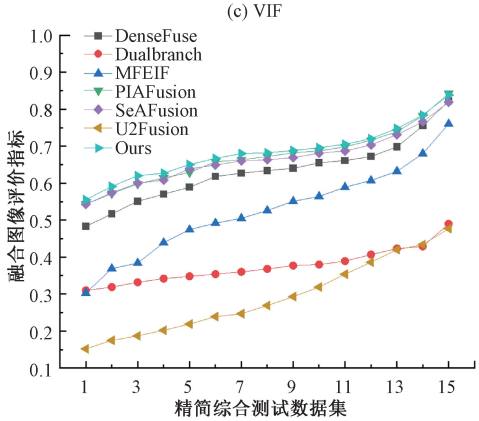
(a) MI



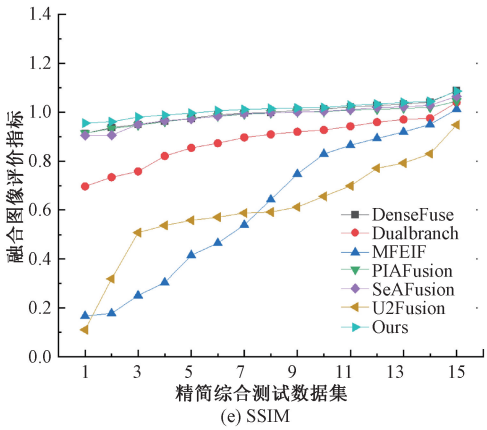
(b) SF



(c) VIF



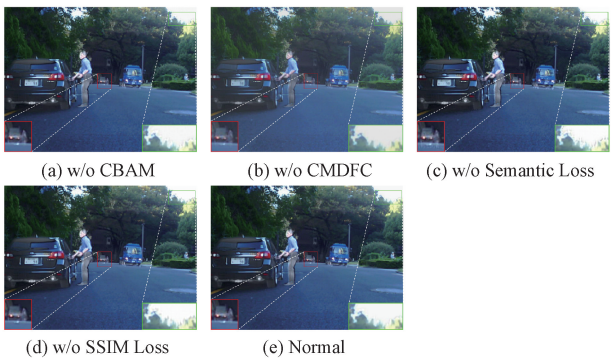
(d) Q_{abf}



(e) SSIM

图 14　不同模型在 MSRS 数据集下的指标分析

Fig. 14　Analysis of metrics for different models across MSRS Dataset



(a) w/o CBAM　(b) w/o CMDFC　(c) w/o Semantic Loss

(d) w/o SSIM Loss　(e) Normal

图 15　消融实验可视化结果

Fig. 15　Visualized results of ablation studies

综上所述,SSCM-Fusion 模型在多种复杂场景下的融合实验均展示了良好的性能,能够满足不同环境不同场景的融合需求,在保证较好的视觉效果的同时,增强了图像的质量,与主观评价一致。

2.5　消融实验和超参数选择

为了能够更好地提取红外图像与可见光图像的特征,本文采用跨模态差分特征补偿模块对互补特征进行了补偿,利用语义损失约束模型训练,并通过迭代的方式使融合网络和语义分割网络互相适应达到更好的效果。为了确认模型各模块的有效性,选择利用消融实验进行验证。具体来说,在保证其他配置相同的情况下,进行以下实验:1) w/o CBAM:不使用注意力机制模块进行图像融合实验;2) w/o CMDFC:不使用跨模态差分特征补偿模块进行图像融合实验;3) w/o Semantic Loss:不使用语义损失约束网络模型进行图像融合实验;4) w/o SSIM Loss:不使用结构相似性损失函数进行图像融合实验;5) Normal:正常的 SSCM-Fusion 模型。结果如表 3、图 15 所示,其中加粗字体表示最优值,下划线字体表示次优值。

表 3　消融实验定量分析

Table 3　Quantitative analysis of ablation experiments

实验	MI	VIF	SSIM
w/o CBAM	4.487	1.002	<u>1.012</u>
w/o CMDFC	4.142	0.823	0.832
w/o Semantic Loss	4.49	<u>1.014</u>	0.983
w/o SSIM Loss	<u>4.593</u>	0.965	0.986
Normal	**4.662**	**1.031**	**1.013**

由表 3 可知,各模块的加入均使模型在性能上获得了提升。其中,MI 指标相较于 w/o Semantic Loss 提升了 3.8%,说明语义损失函数的加入能够使融合图像从源图像中获得了更多的特征细节。VIF 指标相较于 w/o CMDFC 提升了 25.3%,说明 CMDFC 模块能够使融合图像拥有更高的视觉保真度。从图 15 的框选内容可以观察到,缺乏语义损失的约束导致生成图像失真,而缺少 SSIM 损失的约束则在图像中引入噪声。缺少 CBAM 模块会在目标边缘轮廓处引入噪声,而缺少整个 CMDFC 模块的特征补偿后,融合图像中树叶的边缘轮廓模糊,阴影中树干的细节也不明显。这表明 CMDFC 模块语义损失和 SSIM 损失在融合模型中都起到了关键作用,使融合图像的轮廓更加清晰、细节更加丰富,验证了该方法的有效性。

3　结　论

为了获得视觉效果更佳的红外与可见光融合图像,本文提出了一种结合语义分割和跨模态差分特征补偿的红外和可见光图像融合深度学习模型。通过编码器进行图像的特征提取并利用跨模态差分特征补偿模块进行互补信息的补偿,以保留更多的图像纹理细节;结合语义分割网络生成语义损失,使高级语义信息回流到图像融合模块;对融合网络和语义分割网络进行迭代训练,利用内容损失和语义损失约束模型,以产生最终视觉效果更佳的融合图像。通过多种复杂场景实验,证明本文提出的融合网络模型能够应对不同场景和较差环境造成的影响,实现良好的融合效果。对模型进行客观评价指标分析,其中互信息 (MI) 和视觉信息保真度 (VIF) 相比于对照模型中最优指标提升了 4.41%和 4.25%。在主观视觉方面,该模型在保持亮度和对比度的同时,能够突出显著目标信息并保留丰富纹理细节,综合视觉效果由于其他对比模型。

参考文献

[1] 黄月平,李小锋,卢瑞涛,等. 基于自适应标签和稀疏学习相关滤波的红外目标跟踪算法研究 [J]. 仪器仪表学报,2022,43 (12): 199-208.

HUANG Y P, LI X F, LU R T, et al. Research on infrared object tracking algorithm via adaptive label and sparse-learning correlation filter [J]. Chinese Journal of Scientific Instrument, 2022, 43 (12): 199-208.

[2] 罗娟,王立平. 基于非下采样 Contourlet 变换耦合特征选择机制的可见光与红外图像融合算法 [J]. 电子测量与仪器学报,2021,35 (7): 163-169.
LUO J, WANG L P. Infrared and visible image fusion algorithm based on nonsubsampled contourlet transform coupled with feature selection mechanism [J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(7): 163-169.

[3] 丁麒文,祁海军,赵金博,等. 基于目标增强的红外与可见光图像融合算法研究 [J]. 激光与红外,2023, 53 (3): 457-463.
DING Q W, QI H J, ZHAO J B, et al. Research on infrared and visible image fusion algorithm based on object enhancement [J]. Laser & Infrared, 2023, 53 (3): 457-463.

[4] LIU Y, LIU S P, WANG Z F. A general framework for image fusion based on multi-scale transform and sparse representation [J]. Information Fusion, 2015, 24: 147-64.

[5] BAVIRISETTI D P, DHULI R. Two-scale image fusion of visible and infrared images using saliency detection [J]. Infrared Physics & Technology, 2016, 76: 52-64.

[6] 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.

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

[8] 江兆银,王磊. 基于显著性检测与权重映射的可见光与红外图像融合算法 [J]. 电子测量技术,2021, 44 (1): 174-182.
JIANG ZH Y, WANG L. Visible and infrared image fusion algorithm based on significance detection and weight mapping [J]. Electronic Measurement Technology, 2021, 44 (1): 174-182.

[9] 王君尧,王志社,武圆圆,等. 红外与可见光图像多特征自适应融合方法 [J]. 红外技术,2022,44 (6): 571-579.
WANG J Y, WANG ZH SH, WU Y Y, et al. Multi-feature adaptive fusion method of infrared and visible image [J]. Infrared Technology, 2022, 44 (6): 571-579.

[10] 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.

[11] LI H, WU X J, KITTLER J. RFN-Nest: An end-to-end residual fusion network for infrared and visible images [J]. Information Fusion, 2021, 73: 72-86.

[12] ZHANG H, XU H, XIAO Y, et al. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34 (7): 12797-12804.

[13] MA J Y, LIANG P W, YU W, et al. Infrared and visible image fusion via detail preserving adversarial learning [J]. Information Fusion, 2020: 5485-98.

[14] XU H, ZHANG H, MA J Y. Classification saliency-based rule for visible and infrared image fusion [J]. IEEE Transactions on Computational Imaging, 2021: 7824-836.

[15] TANG L F, YUAN J T, MA J Y. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network [J]. Information Fusion, 2022, 82: 28-42.

[16] LU Y, WU Y, LIU B, et al. Cross-modality person re-identification with shared-specific feature transfer [J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020: 13376-13386.

[17] 徐健,陆珍,刘秀平,等. 注意力机制优化 RetinaNet 的密集工件检测方法研究 [J]. 电子测量与仪器学报, 2022,36 (1): 237-235.
XU J, LU ZH, LIU X P, et al. Research on intensive workpiece detection method by attention mechanism optimization RetinaNet [J]. Journal of Electronic Measurement and Instrumentation, 2022, 36 (1): 237-235.

[18] PENG C L, TIAN T, CHEN C, et al. Bilateral attention decoder: A lightweight decoder for real-time semantic segmentation [J]. Neural Networks, 2021, 137: 188-199.

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

[20] 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.

[21] LIU J Y, FAN X, JIANG J, et al. Learning a deep

multi-scale feature ensemble and an edge-attention guidance for image fusion [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32 (1): 105-119.

[22] CAO H C, LIU H, SONG E M, et al. Dual-branch residual network for lung nodule segmentation [J]. Applied Soft Computing, 2020, 86: 105934.

[23] 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, 2020, 44 (1): 502-518.

作者简介

李天放,2019 年于北京化工大学获得学士学位,现为北京信息科技大学仪器科学与光电工程学院在读硕士研究生,主要研究方向为红外成像。

E-mail: litianfang@bistu.edu.cn

Li Tianfang received his B. Sc. degree from Beijing University of Chemical Technology in 2019. Now he is a M. Sc. candidate at School of Instrument Science and Opto-Electronics Engineering in Beijing Information Science and Technology University. His main research interest includes infrared imaging.

孙一宸,2021 年于北京信息科技大学获得硕士学位,现为北京工业大学博士研究生,主要研究方向为红外成像处理算法和衍射深度神经网络增强算法。

E-mail: sunyichen0429@163.com

Sun Yichen received his M. Sc. degree from Beijing Information Science and Technology University in 2021. He is now a Ph. D. candidate at Beijing University of Technology. His main research interests include infrared imaging processing algorithms and diffractive deep neural network enhancement algorithms.

于明鑫,2010 年北京理工大学获得硕士学位,2015 年北京理工大学获得博士学位,现为北京信息科技大学副教授,主要研究方向为机器学习理论与应用、智能微系统轻量化算法、计算机视觉。

E-mail: yumingxin@bistu.edu.cn

Yu Mingxin received his M. Sc. degree in 2010 from Beijing Institute of Technology, and Ph. D. degree in 2015 from Beijing Institute of Technology. Now he is an associate professor in Beijing Information Science and Technology University. His main research interests include theory and application of machine learning, lightweight algorithms for intelligent microsystems, computer vision.

董明利(通信作者),1986 年和 1989 年于合肥工业大学分别获得学士和硕士学位,2009 年于北京理工大学获得博士学位,现为北京信息科技大学教授,北京信息科技大学仪器科学与光电工程学院院长,主要研究方向为光电与视觉检测。

E-mail: dongml@bistu.edu.cn

Dong Mingli (Corresponding author) received her B. Sc. and M. Sc. degree from Hefei University of Technology in 1986 and 1989, and Ph. D. degree in 2009 from Beijing Institute of Technology. Now she is a professor in Beijing Information Science and Technology University and the dean of School of Instrument Science and Opto-Electronics Engineering in Beijing Information Science and Technology University. Her main research interests include optoelectronic and visual inspection.