引用格式：BAI Bendu, LI Junpeng. Multi-exposure Image Fusion Based on Attention Mechanism[J]. Acta Photonica Sinica, 2022, 51(4):0410004
白本督,李俊鹏.基于注意力机制的多曝光图像融合算法[J].光子学报,2022,51(4):0410004

基于注意力机制的多曝光图像融合算法

白本督,李俊鹏
(西安邮电大学 通信与信息工程学院, 西安 710121)

摘　要:针对多曝光图像任务缺乏真值图像,以及现有多曝光图像融合算法存在的边缘特征丢失、细节模糊等问题,本文提出了一种基于注意力机制的多曝光图像融合算法。该算法建立权重独立的双通道Unet网络,对目标场景不同曝光图像分别进行特征提取,生成不同曝光图像的高维多尺度特征图;通过视觉注意力机制凸显目标场景在不同曝光下对融合有利的特征,并叠加到高维多尺度特征图上;最后,将过滤的高维多尺度特征进行特征拼接,通过扩张残差稠密单元进行特征重建,得到最终的高动态范围图像。以端到端网络作为设计基础,利用内容损失和结构损失计算策略对神经网络进行约束,实现无监督学习。实验结果表明,所设计的多曝光图像融合网络在保留过曝和欠曝图像纹理特征的同时降低了欠饱和与过饱和区域信息干扰,具有良好的融合效果和鲁棒性。

关键词:多曝光;注意力机制;卷积神经网络;高动态范围图像;无监督
中图分类号:TP391.41　　**文献标识码**:A　　**doi**:10.3788/gzxb20225104.0410004

0　引言

自然场景有宽广的动态范围,如微弱星光亮度约为10^{-4} cd/m^2,恒星强光亮度范围为10^5～10^9 cd/m^2[1],通过数码单反相机拍摄记录时,往往因为数码相机动态范围受限,导致拍摄的照片出现过曝光和欠曝光[2]。高动态范围(High Dynamic Range, HDR)成像技术旨在扩大图像动态范围,解决由数码相机动态范围受限无法捕获高动态范围图像而产生的问题[3]。目前,生成高动态范围图像的方法大致可分为两类:基于辐照域的多曝光HDR图像生成方法和基于图像域的多曝光HDR图像融合(Multi Exposure Fusion, MEF)方法。基于辐照域的多曝光HDR图像生成方法连续捕获一组不同曝光量的低动态范围(Low Dynamic Range, LDR)图像集合,实现目标场景动态范围的涵盖;然后通过求解相机响应函数(Camera Response Function, CRF)将该组不同曝光的LDR图像集合在辐照域合成一幅HDR图像;最后将HDR图像直接在HDR显示器上进行显示或者将HDR图像经过色调映射显示在LDR显示器[4]。基于图像域的多曝光HDR图像融合方法使用多张涵盖目标场景动态范围的低动态范围图像集合,利用图像融合的方式将具有互补信息多张不同曝光图像直接融合为包含最多信息的单张高动态范围图像[5]。

传统多曝光图像融合算法通过三个步骤获取HDR图像:首先采用拉普拉斯金字塔分解、小波变换或者稀疏表示等图像变换方法将输入图像转换为特征图;其次,根据定义的融合策略进行特征图融合;最后,通过色调映射得到融合更多信息的HDR图像[6]。然而,该类算法使用人工设计生成的特征图进行图像融合,对不同的输入图像不具有鲁棒性,因此随着输入图像的改变其融合质量也受到相应的限制。

近年来,众多学者利用卷积神经网络(Convolutional Neural Network, CNN)解决传统多曝光图像融合方法不能自适应学习图像特征的不足。基于卷积神经网络的MEF通过构建神经网络层将图像特征分层表示,从而有效提取过曝和欠曝图像局部和全局的特征信息。基于卷积神经网络图像融合主要理论研究可大

基金项目:国家自然科学基金(No. 41874173)
第一作者:白本督(1972—),男,副教授,博士,主要研究方向为图形图像处理,图像融合。Email: baibendu@163.com
通讯作者:李俊鹏(1996—),男,硕士研究生,主要研究方向为图像融合。Email: 793348945@qq.com
收稿日期:2021-11-03;**录用日期**:2022-01-15

致分为三类[7]：1)将传统方法与卷积神经网络相结合。例如，2016年LIU Y等[8]提出基于卷积稀疏表示的图像融合算法，该算法的利用提出的卷积稀疏表示模块对源图像序列分解的基础层和细节层分别进行处理，然后将处理后的基础层和细节层相结合得到融合图像。该类方法利用CNN自动提取源图像序列的特征信息，实现高低频特征信号的剥离并且对所分离的特征信号进行增强，但最终的融合规则仍需要人为设计，因此，仍然存在传统方法的局限性。2)将卷积神经网络视为一种生成权重图的方式，权重图表示了每个像素的有效性。例如，2018年LI H等[9]提出基于预训练的CNN模型算法，通过引入预训练的VGGNet网络解决图像特征提取的问题，并根据提取的特征计算源图像序列的融合权重值，以构建融合图像。2020年MA K等[10]提出MEF-Net模型，MEF-Net的核心思想是将任意空间分辨率和曝光次数的静态图像序列进行下采样，通过卷积神经网络提取图像，再将所提取的图像特征通过引导滤波和联合上采样映射成原图大小的融合权重图，通过加权融合的方式得到融合图像。该类方法本质为像素加权融合，像素级MEF对权重图直接融合往往会引入噪声，导致融合图像失真[11-12]。3)基于CNN的MEF端到端学习的方法。2017年PRABHAKAR K R等[13]使用MEF-SSIM无参考质量评价指标作为损失函数设计了第一个基于卷积神经网络的无监督多曝光图像融合模型DeepFuse，实现基于神经网络的MEF端到端的学习。2020年XU H等[14]在FusionGAN[15]的基础上提出了基于MEF-GAN的端到端多曝光图像融合，MEF-GAN将生成对抗网络应用于多曝光图像融合，并首次引入注意力机制，通过判别器区分输入的多曝光图像和融合图像之间的差异，迫使生成器生成的融合图像具有更多过曝和欠曝图像的细节。以上方法实现了MEF问题的端到端学习，但最终融合结果仍不够理想。此外，尽管基于学习的方法实现了MEF自适应特征学习，但是其主要缺点在于，目前为止尚未发现一个完美的损失函数可用于融合图像的无监督学习[12]。因此，如何充分利用不同曝光图像欠饱和与过饱和特征信息，设计一个有效的损失函数仍是一个挑战。

针对以上问题，本文提出一种基于注意力机制的多曝光图像融合算法(Attention Multi Exposure Fusion Network，AMEFNet)。该算法针对MEF数据集无真值图像而无法监督学习的问题，提出一种新的端到端卷积神经网络MEF无监督学习算法；利用权重分离的双通道特征提取模块对目标场景欠曝光和过曝光图像进行特征提取，获得纹理信息表征能力强的特征图；将注意力机制引入到多曝光图像融合任务中，从局部到全局对欠曝光和过曝光图像的局部细节和全局特征进行聚焦，突出对融合有利的图像特征；为了更精确重建融合图像，以L_2范数和结构相似性SSIM作为神经网络的约束准则设计损失函数，获得源图像序列和融合图像之间更小的相似性差异，实现神经网络模型更精准的收敛。AMEFNet充分利用并融合了源图像序列亮度信息与欠饱和与饱和区域图像纹理细节，改善了融合图像细节丢失、失真等问题。

1 多曝光图像融合深度学习模型

本文设计的基于注意力机制的多曝光图像融合算法网络框架由特征提取模块(Feature Extraction Module，FEM)，注意力机制模块(Attention Module，AM)以及特征重建模块(Feature Reconstruction Module，FRM)三个核心计算模块组成，AMEFNet网络结构如图1所示。首先，将目标场景欠曝光和过曝光图像作为两个独立参数分别输入到结构相同的特征提取模块中，获得欠曝光和过曝光的高维特征图。随

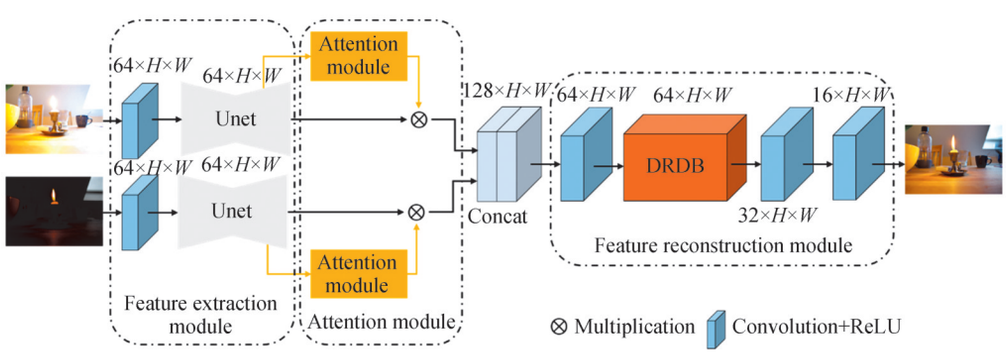

图1 AMEFNet网络结构示意图

Fig. 1 Network structure of AMEFNet

后通过注意力机制模块突出融合有利的图像特征，抑制欠饱和、过饱和等低质量区域的特征，得到重建融合图像所需的纯净高维特征。FRM将AM输出的不同曝光图像的高维特征重建为高动态范围图像。最后利用所设计的损失函数约束神经网络，提高模型的泛化能力。

1.1 AMEFNet网络框架

1.1.1 特征提取模块

在进行多曝光图像融合之前，需要对欠曝光和过曝光图像进行特征提取。以图2所示的Unet网络作为AMEFNet特征提取基础网络架构。

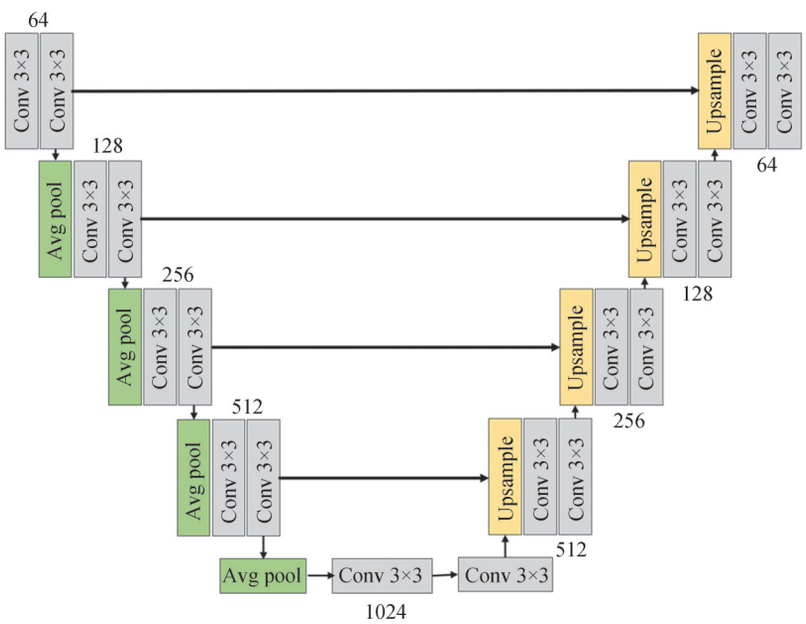

图2 AMEFNet特征提取网络结构

Fig. 2 Feature extraction network structure of AMEFNet

特征提取模块由独立的卷积层和Unet网络构成，其中Unet网络包含卷积，下采样，池化，上采样以及拼接操作。首先使用3×3大小的卷积核在提取低层图像特征的同时，将256×256尺寸的输入图像转化为64通道的高维特征，并利用Unet网络实现图像特征的多尺度特征提取，将浅层图像特征以及深层语义特征通过特征拼接方式堆叠，为保留图像结构和纹理特征提供了有效的解决方案。Unet网络完成特征的精细提取后，输出64通道的高维多尺度特征图，该特征图作为后续注意力机制模块的输入源。

由于欠曝光和过曝光图像在曝光时间上存在差异，同一场景不同曝光图像中的物体具有信息互补，以及亮度，色度，结构对应关系复杂的特点对。若将不同曝光图像直接融合，经过同一网络进行特征提取，生成的共享权值会破坏目标场景不同曝光图像的固有特征。因此，本文在特征提取模块上采用双通道架构，使用结构相同但不共享任何学习参数的特征提取模块，对欠曝光和过曝光图像同时训练。

1.1.2 注意力机制模块

不同曝光图像含有目标场景不同的局部细节特征，为保留多曝光图像丰富的细节信息，凸显融合有利的兴趣特征，抑制非兴趣特征，以校正融合图像局部失真和信息丢失，采用图3所示的注意力机制模块针对不同的曝光图像生成相应的有益特征图。

注意力机制模块分别对Unet网络输出的不同曝光图像的特征图进行Squeeze操作，采用全局平均池化方式将通道空间特征编码为全局特征，计算公式为

$$z_c = F_{sq}(x_c) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i,j) \tag{1}$$

式中，i，j表示像素点坐标，$F_{sq}(\cdot)$为Squeeze操作，x_c为C维度特征输入。之后对全局特征z_c采用Excitation操作，为了降低模型复杂度以及提升泛化能力，采用两个全连接操作，全连接之间使用ReLU激活函数进行

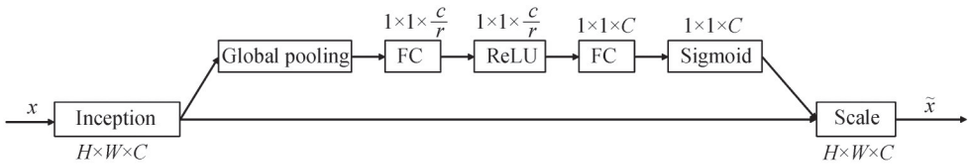

图 3 AMEFNet注意力机制网络结构

Fig. 3 Attention network structure of AMEFNet

非线性处理，最后通过归一化函数Sigmoid输出权值向量，Excitation操作使得网络学习各通道间的关系，也得到不同通道的权值，Excitation操作计算方式为

$$s_c = F_{ex}(z_c, W) = \sigma(g(z_c, W)) = \sigma(W_2\delta(W_1 z_c)) \tag{2}$$

式中，$W_1 \in R^{\frac{C}{r}\times C}$表示维度为$\frac{C}{r}\times C$，$W_2 \in R^{C\times\frac{C}{r}}$表示维度为$C\times\frac{C}{r}$，$r$为缩放因子，$\sigma$、$\delta$分别表示Sigmoid和ReLU激活函数。最后将通道权值s_c与Unet输出的图像特征x_c相乘得到最终特征。

$$\tilde{x} = F_{scale}(x_c, s_c) = x_c \cdot s_c \tag{3}$$

$F_{scale}(\cdot)$表示通道维度对原始特征的重标定，整个操作可以看成学习到了各个通道的权重系数，从而使得模型对各个通道的特征更有辨别能力，注意力机制模块能够在突出融合有利的通道特征的同时抑制非兴趣区域的通道特征。

1.1.3 特征重建模块

完成不同曝光图像的高维特征过滤后，通过拼接操作保留过曝和欠曝图像的高维图像特征得到特征图F_0，再以图4所示的特征重建模块将拼接的特征重建为高动态范围图像。

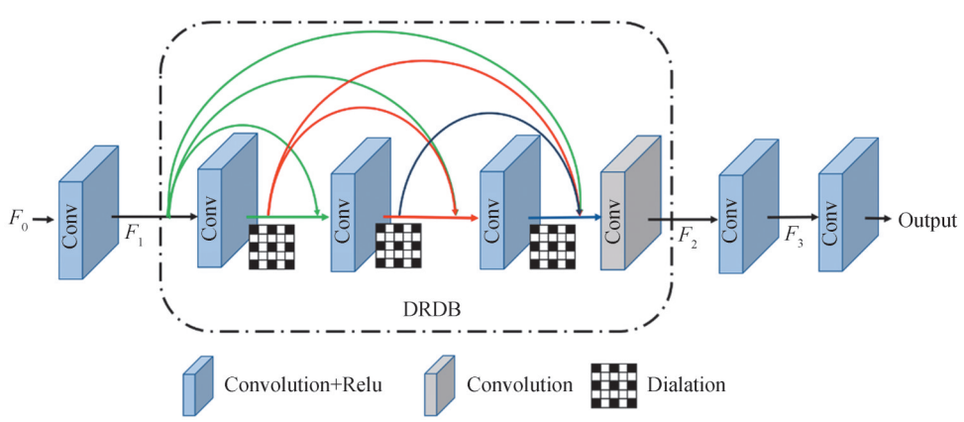

图 4 Unet特征重建网络结构

Fig. 4 Feature reconstruction network structure of Unet

特征重建模块由3个独立的卷积层与1个扩张残差稠密单元(Dilation Residual Dense Block，DRDB)组成。首先第1个卷积层将拼接的特征图F_0转化成64通道的特征图F_1，其次将特征图F_1提供给DRDB单元输出特征图F_2，其中DRDB单元是基于扩张卷积改进残差稠密单元(Residual Dense Block，RDB)得到的，所使用的DRDB充分利用不同网络层级的图像特征，在保留LDR图像细节信息的同时利用更大的感受野去推测饱和区域丢失的细节[16]。此时F_2已有足够的信息去重建高动态范围图像，最后利用2个卷积层依次卷积特征图F_2得到特征图F_3和高动态范围图像。

1.2 损失函数

损失函数决定了所提取的图像特征类型以及不同类型的图像特征之间的比例关系[12]。为了满足融合图像在包含不同曝光图像细节和结构信息的同时，又符合人眼的视觉感知特性的要求。本文设计了基于L_2范数的内容损失和基于SSIM的结构损失的多损失函数。

结构相似性度量指标SSIM从亮度，对比度和结构三方面衡量源图像与融合图像相似性程度。设x为输入图像，y为输出图像，其数学表达式为

$$\mathrm{SSIM}_{x,y}=\frac{2\mu_x\mu_y+C_1}{\mu_x^2+\mu_y^2+C_1}\cdot\frac{2\sigma_x\sigma_y+C_2}{\sigma_x^2+\sigma_y^2+C_2}\cdot\frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3} \tag{4}$$

式中，μ和σ分别表示均值和标准差，σ_{xy}表示x，y的协方差，C_1，C_2和C_3为常数系数。以结构相似性SSIM为基础，针对多曝光图像融合任务设计子结构损失L_{SSIM}，O，U和F分别表示过曝图像，欠曝图像以及融合图像，则L_{SSIM}的数学表达式为

$$L_{\mathrm{SSIM}}=\alpha_O\mathrm{SSIM}_{O,F}+\alpha_U\mathrm{SSIM}_{U,F} \tag{5}$$

$\mathrm{SSIM}_{O,F}$，$\mathrm{SSIM}_{U,F}$分别表示过曝图像O和欠曝图像U与融合图像F的结构相似性。在多曝光图像融合任务中，过曝和欠曝图像的具有相同的纹理细节，但其亮度强度过大或过小。所以对权重系数α_O和α_U设置相同的权重进行平衡，以获得适当大小的亮度强度和纹理细节，可表示为

$$\alpha_O=\alpha_U \tag{6}$$

内容损失L_{content}在保证多曝光图像序列和融合图像的纹理细节信息失真最小的同时避免了噪声的干扰，内容损失的计算表示为

$$L_{x,y}=\left\| x-y \right\|_2 \tag{7}$$

式中，计算输入图像x与输出图像y像素点之间的欧式距离，其中$\left\| \cdot \right\|_2$为L_2范数。内容损失可以定义为

$$L_{\mathrm{content}}=\beta_O L_{O,F}+\beta_U L_{U,F} \tag{8}$$

与结构损失相似的，β_O和β_U具有相同的权重系数。为实现结构损失函数与内容损失函数之间权值平衡，通过赋予结构损失相应的超参数λ提高模型的泛化能力。综上，AMEFNet整体损失函数可表示为

$$\mathrm{Loss}=\lambda L_{\mathrm{SSIM}}+L_{\mathrm{content}} \tag{9}$$

2 实验及结果分析

2.1 实验环境及相关参数

本文基于CAI J等[17]提供的公共可用数据集SICE进行无监督学习的训练。其中训练数据集包含589组不同曝光的LDR图像集合。训练硬件平台为配置为Inter(R) Core(TM) i5-9600k 3.7GHz CPU和NVIDIA Geforce RTX 2080ti GPU的PC，配置环境为Unbuntu18.0.4，网络模型的编写语言为python3.7，配合Pytorch1.5与Opencv3.2作为辅助高级API，并且选用Adam优化器以参数β_1=0.9，β_2=0.999，初始学习率为10^{-4}，学习率每迭代50次便以0.5倍进行衰减，进行模型优化。

通过图5可以看出，AMEFNet在训练过程损失函数具有较好的收敛，未出现梯度消失或梯度爆炸情况，验证了所设计的多曝光图像融合模型的合理性与可行性。完成多曝光图像融合模型训练后，为验证本文算法模型的有效性，本文选取ZHANG X[12]提供的基准数据集MEFB部分图像作为测试集，其中包含室内、室外、白天和黑夜等静态场景，涵盖了广泛的真实环境，更能展现出真实场景信息。为比较所提算法和其他算法的性能，选取三种传统算法：GFF算法[18]，DSIFT算法[19]，SPD-MEF算法[11]和两种基于深度学习

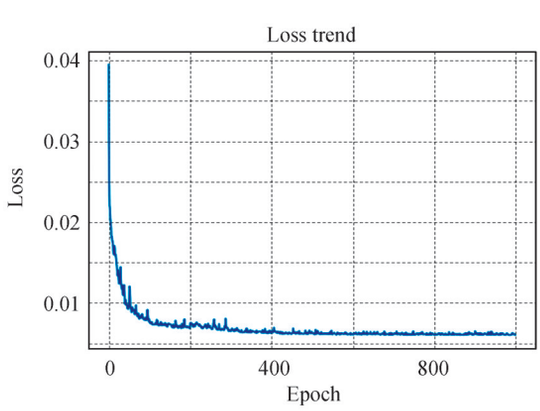

图5 AMEFNet损失函数曲线

Fig. 5 Loss function curve of AMEFNet

的算法：MEFNet算法[10]，MEF-GAN算法[14]从主观和客观两种不同维度进行比较。上述算法的融合图像均由公共可实现代码生成，并且基于深度学习的训练模型由原作者提供。

2.2 主观评价

首先从视觉感知的角度对不同融合算法进行主观对比分析，House多曝光图像序列，TableLamp多曝光图像序列的融合结果效果图如图6、图7所示。

图6为House图像序列融合结果，图6(a)、图6(b)为源图像，图6(c)、图6(h)为不同算法融合结果。由图6(c)、图6(d)可知，与原图像序列相比，GFF和DSIFT处理后的融合图像，虽然能有效提高图像质量，但仍然存在部分区域图像失真的现象，如图6(c)、图6(d)中红色框区域，过曝图像中亮度最强区域融合图像具有明显的暗区；SPD-MEF算法整体过于明亮，如图6(e)红色框区域可以看出，在窗外部分可视信息模糊，视觉效果欠佳；通过图6(f)红色框区域可以发现，MEFNet存在亮度不均匀的问题，导致融合图像细节不清晰，亮度失真；而图6(g)中，MEF-GAN算法可以较好均衡亮度信息，但图像局部出现了失真现象，如图6(g)红色框区域，远景树木和门框轮廓都出现伪影现象。而本文算法(h)在保持高清晰度和对比度的同时，有效避免了亮度不均匀和局部区域细节失真的问题。

图6 House图像序列算法结果比较图

Fig. 6 Comparison of House sequence algorithm results

图7为Table Lamp图像序列融合结果，图7(a)、图7(b)为源图像，图7(c)、图7(h)为不同算法融合结果。通过图7(c)、图7(d)可以看出，GFF和DSIFT算法虽然能较好提升视觉效果，但局部区域仍存在亮度偏低的问题，如图7(c)、图7(d)红色框区域中台灯后的墙体具有明显的暗区；SPD-MEF算法虽然获得更好

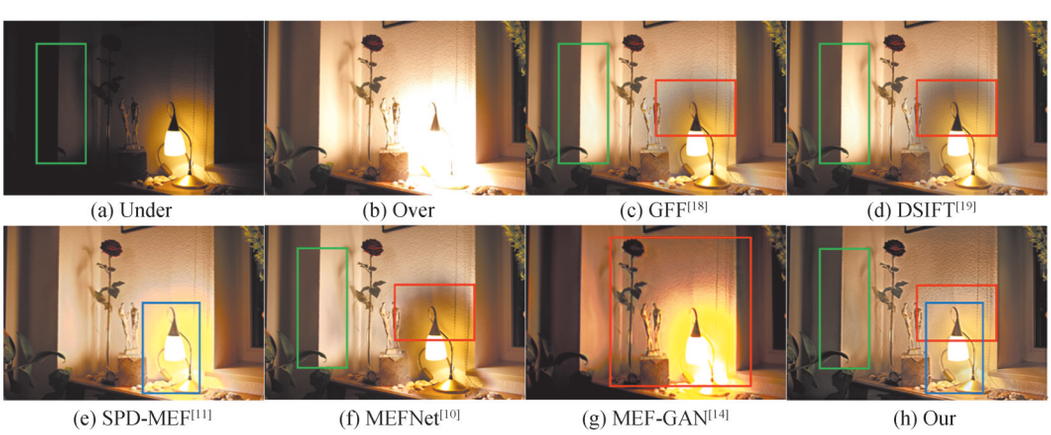

图7 TableLamp图像序列算法结果比较图

Fig. 7 Comparison of TableLamp sequence algorithm results

的视觉增益，但部分区域原始结构信息有所丢失，如图7(e)蓝色框标注区域中台灯的轮廓出现伪影现象，导致融合图像缘结构信息丢失；MEFNet算法处理后的融合图像细节和色彩信息具有明显的改善，但如图7(f)红色框区域所示，MEFNet具有与GFF和DSIFT相同的问题，存在局部暗区，明暗过度不平滑；由图7(g)所示MEF-GAN算法的融合图像存在较为严重的失真现象，如图7(g)红色框区域中花朵和台灯区域出现伪影，导致融合图像不够生动自然，视觉效果欠佳；由于过曝图像相比于欠曝图像具有更大的亮度数值，因此过曝图像在多曝光图像融合中具有更高的权重比，导致上述算法局部区域亮度过度不够自然，如绿色框所示，而本文算法7(h)在一定程度上解决了上述算法的缺点，使得融合图像更加清晰，亮度过渡更加自然。

2.3 客观评价

由于人眼的主观感受判断融合结果存在一定的误差，融合图像在保留欠曝光与过曝光图像的梯度信息和空间频率信息，同时融合图像要尽可能自然，使人眼能够快速、准确的从融合图像获取综合信息，因此本文选用峰值信噪比(Peak Signal Noise Ratio，PSNR)、平均梯度(Average Gradient，AG)、空间频率(Spatial Frequency，SF)、信息熵(Entropy，EN)以及视觉信息保真度(Visual Information Fidelity，VIF)进行客观评价。

1)峰值信噪比(PSNR)。PSNR表示融合图像峰值功率与噪声之间的比值，用于测量融合图像的畸变情况，PSNR定义为

$$\mathrm{PSNR}=10\log_{10}\frac{r^2}{\mathrm{MSE}} \tag{10}$$

式中，r表示融合图像的峰值，MSE表示均方误差。较大的PSNR意味着融合图像与源图像更加接近，拥有更小的失真，所以PSNR越大，融合性能越好。

2)平均梯度(AG)。AG用于量化融合图像的梯度信息，衡量融合图像的细节和纹理信息的保有量，AG定义为

$$\mathrm{AG}=\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\sqrt{\frac{\nabla F_x^2(i,j)+\nabla F_j^2(i,j)}{2}} \tag{11}$$

式中，$\nabla F_x^2(i,j)=F(i,j)-F(i+1,j)$和$\nabla F_y^2(i,j)=F(i,j)-F(i,j+1)$分别表示融合图像水平方向和垂直方向的梯度变化。AG值越大说明融合图像包含的梯度信息越多，融合效果越好。

3)空间频率(SF)。SF通过对融合图像的行频率和列频率求均方得到，反映了融合图像空间频率信息，SF定义为

$$\mathrm{SF}=\sqrt{\mathrm{RF}^2+\mathrm{CF}^2} \tag{12}$$

式中，$\mathrm{RF}=\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j)-F(i,j-1))^2}$和$\mathrm{CF}=\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j)-F(i-1,j))^2}$分别表示行频率和列频率。SF值越大表示融合图像细节信息越丰富，融合性能越好。

4)信息熵(EN)。信息熵作为图像量化标准，反映图像包含的平均信息量多少，定义为

$$\mathrm{EN}=-\sum_{l=0}^{L-1}P_l\log_2 P_l \tag{13}$$

式中，L表示灰度级数，P_l表示融合图像相应灰度的归一化直方图。信息熵大小表示融合图像携带信息量的多少。但是，信息熵很容易受到噪声的影响，因此通常用作辅助指标。

5)视觉信息保真度(VIF)。VIF是基于人眼视觉信息保真度的质量评价指标，主要通过建立视觉模型计算源图像和融合图像间的信息失真。计算融合图像基于VIF的总体度量，VIF值越大，融合性能越好。

为了方便观察，本文将测试集的5个客观质量评价指标平均得分在表1中列出(粗体值代表了每列的最佳值)。从表1中数据可知，本文算法融合图像虽然在PSNR指标上获得第二高的得分，但最高得分的SPD-MEF算法融合图像具有肉眼可见的伪影现象。而在EN指标上获得第四高的得分，但得分最高的前三算法融合图像均出现由于融合不完全导致的不同程度黑影，并且本文算法与最高得分的差距仅为2.6%。而本文算法在AG，SF和VIF性能指标结果均优于其他对比算法，因此可以表明相比于其它算法本文算法的融合图像蕴含了更多原图像序列的场景细节与边缘信息，符合人眼的视觉感知特性的要求。

表1 测试集融合图像的5个客观评价指标的平均值

Table 1 The values of five quality metrics averaged over the fused images on test set

Fusion Method	PSNR	AG	SF	EN	VIF
GFF	58.224 0	5.623 4	18.826 9	7.393 9	0.825 3
DSIFT	**59.825 6**	5.105 1	17.103 8	7.354 3	0.759 1
SPD—MEF	58.754 5	5.877 3	21.033 9	7.039 9	0.794 9
MEFNet	58.380 4	6.029 3	20.596 6	**7.394 8**	0.844 9
MEF—GAN	58.773 6	4.756 5	13.554 8	6.982 3	0.621 1
Ours	59.115 0	**6.199 1**	**21.098 6**	7.203 7	**0.874 6**

2.4 算法复杂度与运行时间

复杂度是体现算法优劣的一个重要指标。因此，在这一节中，将对MEF算法的时间复杂度进行讨论。设m表示图片的行数，n表示图片的列数，N表示源图像序列中的图像数量，因此传统算法的复杂度均为$O(Nmn)$。对于深度学习算法将以具体的浮点运算数(Floating Point Operations，FLOPs)进行比较，FLOPs主要衡量模型的复杂度，FLOPs值越大，模型需要更多的运行时间，因此模型的复杂度越大，其中1GFLOPs表示10^9次浮点运算。不同算法复杂度将在表2列出。从表2中可以看出，本文的FLOPs值最大，这是由于特征重建层中使用了扩张残差稠密单元，扩张残差稠密单元的使用提高了AMEFNet网络图像重建的性能，但代价是算法复杂度的增加。

表2 不同算法的复杂度

Table 2 Time complexity of the different algorithm

Fusion method	Complexity	FLOPs($\times10^9$G)
GFF	$O(Nmn)$	—
DSIFT	$O(Nmn)$	—
SPD—MEF	$O(Nmn)$	—
MEFNet	—	13.2
MEF—GAN	—	71.1
Ours	—	165.6

为更直观的评估算法复杂度，将不同分辨率多曝光图像序列在不同算法运行时间在图8进行绘制，其运行时间测试平台为i5-8265u CPU环境。从图8可以看出，本文算法的运行速度快于SPD-MEF与DSIFT两种传统算法，但慢于MEF-Net与GFF两种算法，其中MEF-Net运行最快，而在低分辨率图像运行时间比较中，本文算法与MEF-GAN算法运行时间接近，但是本文算法的融合结果相比于MEF-GAN具有更好的鲁棒性，比MEF-Net与GFF算法具有更高的饱和度，更好的视觉效果。综上所述，本文算法的复杂度适中，能够较好融合不同曝光图像。

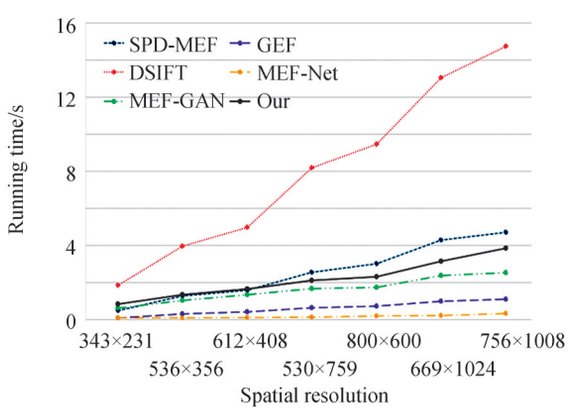

图8 不同算法的运行时间比较

Fig. 8 Comparison of running time

2.5 消融实验

为验证所提AMEFNet网络框架在多曝光图像融合任务中的有效性,从两个方面进行消融性分析:1)证明注意力机制模块的有效性;2)证明超参数λ的有效性。从上述两个方面,设计具有相同设置的基本网络用于探索不同模块的有效性。设计网络分别称之为:RAMEFNet(移除AMEFNet中的注意力模块)和AMEFNet-λ_x(网络结构与AMEFNet一致,不同之处在于超参数λ数值)。

2.5.1 注意力机制有效性实验

AMEFNet中引入注意力机制模块,从局部到全局的方式聚焦于目标场景不同曝光图像的细节特征,从而校正融合图像局部失真和图像畸变,得到更好的融合效果。为验证注意力机制的有效性,将AMEFNet的注意力机制模块移除,即将图1中的AM模块移除,称之为RAMEFNet,其余设置与AMEFNet保持一致。

在主观评价中,图9展示了Studio图像序列有无注意力模块的可视化结果。其中图9(a)、图9(b)为欠曝和过曝图像,图9(c)是MEF-Net融合结果,图9(d)是AMEFNet融合结果。从图9可以明显看出,RAMEFNet虽然能融合过曝和欠曝图像部分互补信息,但在部分区域存在细节模糊的问题,如图9(c)中红色框区域灯泡未显示出应有的轮廓信息以及远景树木存在一定程度的模糊现象。RAMEFNet与本文算法相比,本文算法融合结果不仅能够融合不同曝光图像的细节信息而且图像整体更加自然,这是因为注意力模块能够有效聚焦源图像序列的局部细节和图像特征避免细节丢失等现象发生。

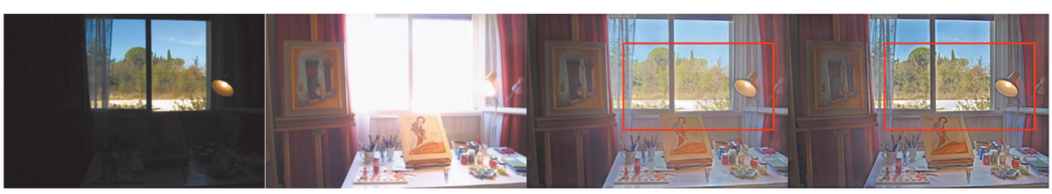

图9 Studio图像序列各融合结果

Fig. 9 Exposure fusion results of Studio sequence

在客观评价中,表3展示了有无注意力模块在PSNR,SF,AG,EN和VIF五种客观质量评价指标下的结果。从表3可以看出相较于RAMEFNet,对于测试集,AMEFNet均取得了最优结果。表明AMEFNet的融合图像具有更多图像细节和纹理信息,更符合人眼的主观视觉特性,这也与可视化结果相匹配。

表3 有无注意力模块的5个客观评价指标的平均值

Table 3 Evaluations of attention module on five quality metrics average

Fusion method	PSNR	AG	SF	EN	VIF
RAMEFNet	58.750 9	5.986 6	20.135 1	6.9819	0.833 9
AMEFNet	**59.115 0**	**6.199 1**	**21.098 6**	**7.203 7**	**0.874 6**

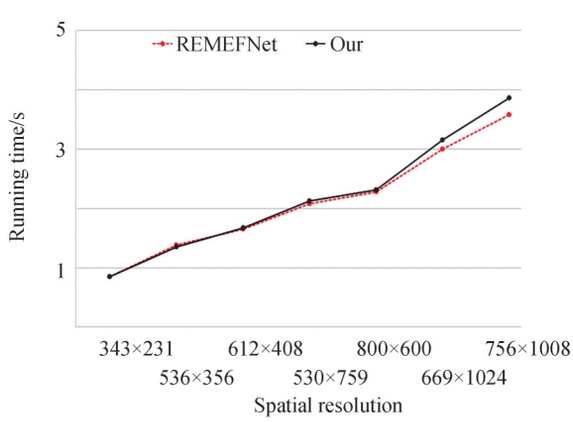

图10 不同算法的运行时间比较

Fig. 10 Comparison of running time

在运行时间比较中，图10展示了有无注意力机制在不同分辨率多曝光图像序列的运行时间。从图10中可以看出去除注意力机制模块后减少了相应网络的参数，增快了RAMEFNet运行时间，但是RMEFNet运行时间与本文算法相差不大，并且本文算法融合结果能够保留更多的图像细节，进一步说明了注意力机制在本文算法应用的有效性。

2.5.2 超参数λ有效性验证

在1.2小节介绍了损失函数的设计，除了使用经典的L_2范数作为内容函数之外还引入了基于SSIM的结构损失，为验证不同超参数λ对网络的影响，将AMEFNet中超参数λ设置为不同的数值，称之为AMEFNet-λ_x。

由于融合结果相近，图像视觉感知效果几乎一致，因此对超参数λ的有效性不做主观分析。在表3展示了不同超参数λ在PSNR，SF，AG，EN和VIF五个客观评价指标上的结果。从表4可以看出当超参数λ太高或太低时，融合图像在一定程度丢失了部分细节信息，导致融合图像指标偏低，而中间值λ=0.2的融合图像拥有更多梯度和细节信息在五个指标中占据了三个最优结果，因此本文将λ=0.2作为AMEFNet网络默认设置。

表4 不同超参数λ的5个客观评价指标的平均值

Table 4 Evaluations of different hyperparameter λ on five quality metrics average

Fusion method	PSNR	AG	SF	EN	VIF
AMEFNet−$\lambda_{0.02}$	**59.249 8**	6.035 1	20.113 4	7.1011	0.843 1
AMEFNet−$\lambda_{0.2}$	59.115 0	**6.199 1**	**21.098 6**	**7.203 7**	0.874 6
AMEFNet−λ_2	59.0990	6.174 3	21.035 9	7.1915	**0.879 5**

3 结论

本文提出了一种基于注意力机制的卷积神经网络(AMEFNet)用于多曝光图像融合。AMEFNet引入了注意力机制模仿人类视觉机制，从过曝和欠曝图像中突出对融合有利的图像特征；此外，为了捕获更多源图像序列的结构和内容信息构建基于L_2范数和SSIM的多损失函数。上述操作使得本文网络可以捕捉更多细节信息，生成质量更好的融合图像。对比实验和消融实验均表明本文所提的AMEFNet在多曝光图像融合任务中具有显著的优越性。

参考文献

[1] DONG G, YUAN C, ZHUN S, et al. Correcting over-exposure in photographs[C]. Proceedings of IEEE Computer Vision and Pattern Recognition, 2010: 515-512.

[2] RAMIREZ O R, MARTIN I, LOSCOS, et al. Full high dynamic range images for dynamic scenes[C]. Proceedings of SPIE, 2012: 843609-843625.

[3] SHEN R, CHENG I, SHI J, et al. Generalized random walks for fusion of multi-exposure images[J]. IEEE Transactions on Image Processing, 2011, 20(12): 3634-3646.

[4] ZHAO Jinbo, BAI Bendu, FAN Jiulun, et al. An acquisition method of minimal-bracketing sets based on optimal exposure [J]. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(10): 1890-1898.
赵金波, 白本督, 范九伦, 等. 基于最优曝光的最小包围曝光集合获取方法[J]. 计算机辅助设计与图形学学报, 2018, 30(10): 1890-1898.

[5] REINHARD E, STARKM, SHIRLEY P, et al. Photographic tone reproduction for digital images[C]. Proceedings of Computer Graphics and Interactive Techniques, 2002: 267-276.

[6] LI S, KANG X, FANG L, et al. Pixel-level image fusion: a survey of the state of the art[J]. Information Fusion, 2017, 33(1): 100-112.

[7] YANG Z, CHEN Y, LE Z, et al. GANFuse: a novel multi-exposure image fusion method based on generative adversarial networks[J]. Neural Computing and Applications, 2021, 33(11): 6133-6145.

[8] LIU Y, XUN C, WARD R K, et al. Image fusion with convolutional sparse representation[J]. IEEE Signal Processing Letters, 2016, 23(12): 1882-1886.

[9] LI H, WU X J, KITTLER J. Infrared and visible image fusion using a deep learning framework[C]. Proceedings of IEEE International Conference on Pattern Recognition, 2018: 2705-2710.

[10] MA K, DUANMU Z, ZHU H, et al. Deep guided learning for fast multi-exposure image fusion[J]. IEEE Transactions

on Image Processing, 2020, 29(11): 2808-2819.
[11] MA K, HUI L, YONG H, et al. Robust multi-exposure image fusion: a structural patch decomposition approach[J]. IEEE Transactions on Image Processing, 2017, 26(5):2519-2532.
[12] ZHANG X. Benchmarking and comparing multi-exposure image fusion algorithms[J]. Information Fusion, 2021, 74(10): 111-131.
[13] PRABHAKAR K R, SRIKAR V S, BABU R V. DeepFuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs[C]. Proceedings of IEEE International Conference on Computer Vision, 2017: 4724-4732.
[14] XU H, MA J, ZHANG X P. MEF-GAN: multi-exposure image fusion via generative adversarial networks[J]. IEEE Transactions on Image Processing, 2020, 29(6): 7203-7216.
[15] MA J, YU W, LIANG P, et al. FusionGAN: a generative adversarial network for infrared and visible image fusion[J]. Information Fusion, 2019, 48(8): 11-26.
[16] YAN Q, GONG D, SHI Q, et al. Attention-guided network for ghost-free high dynamic range imaging[C]. Proceedings of IEEE Computer Vision and Pattern Recognition, 2019: 1751-1760.
[17] CAI J, GU S, ZHANG L. Learning a deep single image contrast enhancer from multi-exposure images[J]. IEEE Transactions on Image Processing, 2018, 27(4): 2049-2062.
[18] LI S, KANG X, HU J. Image fusion with guided filtering[J]. IEEE Transactions on Image processing, 2013, 22(7): 2864-2875.
[19] LIU Y, WANG Z. Dense SIFT for ghost-free multi-exposure fusion[J]. Journal of Visual Communication and Image Representation, 2015, 31(8): 208-224.

Multi-exposure Image Fusion Based on Attention Mechanism

BAI Bendu, LI Junpeng
(*School of Communication and Information Engineering, Xi'an University of Posts & Telecommunications, Xi'an* 710121, *China*)

Abstract: Humans mainly perceive and understand the unknown world by obtaining effective information, the visual system has always been an important way to obtain external information. With the development of digital information technology and the demand for human vision, imaging equipment has greatly improved in items of image resolution and dynamic response range. In recent years, imaging technology and its processing technology have played a vital role in various fields. Due to images captured by traditional cameras can only record a limited dynamic range, and the scene is unrepeatable and transient, the interested target cannot be captured again, and we can only process existing images, therefore, reconstructing the high dynamic range image from low-quality images and improving the visual quality of scenes is a key issue in computer vision and has very important research value. In this dissertation, we focus on the dynamic range image reconstruction method in improving image quality of static scenes. The lack of ground-truth fused images for supervised learning, and exiting multi-exposure image fusion suffer from loss of edge features and blurred detail. To address these problems, we propose an attention guided network for multi-exposure image fusion. First, a dual channel Unet network with independent weights is established, extract feature from under-exposure and over-exposure images of the target scene, and a multi-scale and high-dimensional feature maps with strong texture information feature expression ability is obtained. Then, through visual attention mechanism focus local details and global features of under- and over-exposure images, generated the logical mask of the target region of interest area and superimposed on the high-dimensional multiscale feature maps to highlight the target features and suppress the non target area. Finally, during the reconstruction process, we concatenate the filtered high-dimensional multiscale features, the dilation residual dense block is used, the dilation residual dense block makes full use of the features of different levels, retains more detailed information from low dynamic range image, and increases the image receptive filed to predict the details of the saturation region. Based on end-to-end network, in order to reconstruct the fused image more accurately, in which the L2 norm is used as the constraint criterion of the content loss and the SSIM is used as the constraint criterion of the structural loss to design multiple loss functions constrain the neural network, so as to obtain a small similarity difference between

the source image sequence and the fused image, realize more accurate convergence of the neural network model, and unsupervised learning. To verify the effectiveness of the proposed algorithm, some images selected from the MEFB benchmark dataset as the test set. The test set including indoor, outdoor, day, night and other static scenes, covering a wide range of real scenes, which can better show the real scene information. Combined three traditional algorithms and two deep learning algorithms for subjective analysis, and used five quality evaluation indicators of fusion image and average running time for objective evaluation. Ablation study were carried out from the effectiveness of both the attention mechanism module and the loss function λ hyper parameters, the experimental results show that the proposed algorithm can capture more detailed information and structural information from the source image sequences under static scenes, obtain fused images with clear scenes and salient features, and the fused image is more in line with human visual characteristics. Comparing with the other typical algorithms, the proposed algorithm not only overcomes the shortcomings of traditional algorithms that cannot adaptively learn features and the fusion rules need to be hand-crafted, but also introduces attention mechanism and dilation residual dense block, which make easier to predict the details and structural information of saturated areas and under-exposure areas, so as to obtain more comprehensive, reliable, abundant scene information with stronger robustness.

Key words: Multi-exposure; Attention mechanism; Convolution Neural Network; High-dynamic range image; Unsupervised

OCIS Codes: 100.2000; 150.0155; 350.2666

Foundation item: National Natural Science Foundation of China (No.41874173)