基于语义信息引导的图像协调化

杨紫媛[1,2],李鹏程[1,2],刘芳岑[1,2],高陈强[1,2]
(1. 重庆邮电大学通信与信息工程学院,重庆 400065;2. 信号与信息处理重庆市重点实验室,重庆 400065)

摘 要: 图像协调化在图像处理中占据着一个重要的地位,它旨在调整前景外观(如光照、颜色、纹理等)使其与背景在视觉上保持一致. 然而,现有的基于深度学习方法通常将图像整体背景的特征分布作为线索来调整前景,没有注重语义信息对前景调整的关键作用,导致前景的局部区域与背景在视觉上出现差异. 为此,本文基于多分辨率选择融合模块(Multi-Resolution Selective Fusion Module,MRSFM)和轻量级的卷积块注意力模块(Convolutional Block Attention Module,CBAM),设计了一个基于双注意力机制的多分辨率选择融合模块(Multi-Resolution Selective Fusion module based on Dual Attention Mechanism,MRSF-DAM),使得最后输出的特征图具有丰富的语义信息,从而引导网络更好地理解图像前景与它周围场景之间的相关性,使网络更加充分地从背景中获取协调前景所需的各种信息,最终缩小图像前景区域和背景区域在视觉上的外观差异. 此外,本文设计了一个新的网络架构来选择融合浅层和深层的特征信息,通过对解码器前6层网络层与MRSF-DAM的输出特征图进行多尺度融合和增强,将产生的增强特征图送入解码器的最后层,能够缓解由跳跃连接引入的与前景内容的特征不相关的问题,且减少了由于解码器经过多次下采样带来的空间特征信息损失,进一步提高生成协调图像的真实性. 在广泛使用的iHarmony4基准数据集上进行了大量的实验验证了本文方法的有效性. 相比于目前最新的方法SCS-Co(Self-Consistent Style Contrastive learning for image harmonization),本文方法在整个数据集的均方误差(Mean Squared Error,MSE)、前景均方误差(foreground Mean Squared Error,fMSE)和峰值信噪比(Peak Signal-to-Noise Ratio,PSNR)上分别提升了4.28,61.97和1 dB.

关键词: 图像协调化;图像处理;语义信息;局部背景信息;多分辨率选择融合;空间特征信息
基金项目: 国家自然科学基金(No.62176035)
中图分类号: TN911.73;TP391 **文献标识码:** A **文章编号:** 0372-2112(2023)07-1826-09
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20221322

Image Harmonization Guided by Semantic Information

YANG Zi-yuan[1,2], LI Peng-cheng[1,2], LIU Fang-cen[1,2], GAO Chen-qiang[1,2]
(1. *School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*; 2. *Chongqing Key Laboratory of Signal and Information Processing, Chongqing* 400065, *China*)

Abstract: Image harmonization occupies an important position in image processing. It aims to adjust the foreground appearance, e.g., illumination, color, texture, etc., to be visually consistent with the background. However, existing deep learning-based methods usually use the feature distribution of the overall image background as a cue to adjust the foreground, without focusing on the critical role of semantic information for foreground alignment, resulting in local areas in the foreground to appear visually different from the background. To this end, based on the multi-resolution selective fusion module (MRSFM) and the lightweight convolutional block attention module (CBAM), this paper designs a multi-resolution selective fusion module based on dual attention mechanism (MRSF-DAM), which makes the final output feature map rich in semantic information, thus guiding the network to better understand the correlation between the foreground of an image and its surrounding scene, more enabling the network to fully obtain the various information needed to coordinate the foreground from the background, and eventually reducing the visual discrepancy between the foreground and background regions of an image. In addition, this article designs a new network architecture to selectively fuse the shallow and deep feature information. By multi-scale fusion and enhancement of the output feature maps of the first six network layers of the decoder and MRSF-DAM, the generated enhanced feature maps are fed into the final layer of the decoder, which can alleviate the problem introduced by skip connections of the unrelated features to the foreground,and besides, it reduces the loss of

收稿日期:2022-11-18;修回日期:2023-03-13;责任编辑:覃怀银

spatial feature information caused by multiple downsampling of the decoder, further improving the authenticity of the generated harmonized images. A large number of experiments were conducted on the widely used iHarmony4 benchmark dataset to verify the effectiveness of our method. Compared to the latest method SCS Co (Self Consistent Style Comparative learning for image harmonization), this proposed method improves the mean squared error (MSE), foreground mean squared error (fMSE) and peak signal to noise ratio (PSNR) of the entire dataset by 4.28, 61.97, and 1 dB, respectively.

Key words: image harmonization; image processing; semantic information; local background information; multi-resolution selective fusion; spatial feature information

Foundation Item(s): National Natural Science Foundation of China (No.62176035)

1 引言

图像协调化旨在调整图像前景区域的外观使其与背景区域的外观在视觉上保持一致. 该技术可以广泛应用于电影场景抠图[1]和修图[2]等视觉任务. 同时,也常用于图像合成[3]、图像增强[4]、图像恢复[5,6]、图像超分重建[7~9]和风格迁移[10~12]等原始图像本身外观会发生变化的领域. 通常情况下,需要协调化的前景和背景来源于不同拍摄位置、角度、时间和设备的图像,存在明显的差异和复杂性. 这使得图像协调化是一项具有挑战性的视觉任务,近年来受到广泛关注.

早期的图像协调化工作[13~20]侧重于传递手工制作的低级外观统计来匹配前景与背景区域的视觉外观. 例如,基于相关和非相关颜色空间的参数化方法[13,14]、金字塔直方图匹配技术[17]以及全局和局部颜色统计数据[19]的方法. 虽然这些方法学习了自然图像的颜色转换,对图像内容的适应性更强,但仍然没有将物体颜色和照明颜色的效果分开,导致大多数协调后的图像缺乏真实性. 目前,国内外研究者为了得到更加逼真的协调化图像,设计了各种基于深度学习的网络模型,包括域转换[21,22]、风格迁移[23]、对比学习[24]、反射率和照明[25,26]和自监督[27]等方案. Cong等人[21,22]提出基于域转换的方法[21]从前景域和背景域、真实图像域和组合图像域一致性的角度来提高生成协调图像的真实性,随后他又提出域代码提取器[22]来捕获背景域信息. Ling等人[23]首次将图像协调化任务视为风格转移问题,提出了一种区域感知自适应实例归一化模块来从背景特征中捕获样式信息. 接着,Hang等人[24]首次将对比学习的概念引入图像协调化领域,从前景自身风格和前景与背景风格一致性两个方面学习更多的失真知识,以生成更逼真的视觉效果. Guo等人[25,26]是首位将固有图像理论的模型应用于图像协调化的研究者. 他们将合成图像分解为反射率和照明的内在图像以进行单独协调. Jiang等人[27]提出了第一个基于自监督方法的协调框架,解决了昂贵的人力标注问题. 然而,这些方法忽视了不同局部区域背景的外观存在较大的差异,以及语义信息对网络挖掘背景特征的指导作用,导致网络不仅对背景中与前景相关的内容关注不够,还容易学习到一些与前景内容不相关的信息,使得生成的协调图像不真实感明显. 尽管DIH(Deep Image Harmonization)[28]设计了一个场景分割解码器来使用语义信息进行协调前景,但是该模型重构了整张组合图像作为预测图像,导致协调后图像和真实图像的背景在视觉上具有显著差异. 随着注意力机制在深度学习领域的发展,最近的图像协调化方法[21,29~31]通常将注意力机制加入解码过程中,以增强提取特征的能力. SENet(Squeeze-and-Excitation Networks)[32]结合了轻量级的门控机制,能自动获取到每个特征通道的重要程度. GENet(Gather-excite:Exploiting feature context in convolutional neural Networks)[33]结合收集和激发操作,通过在空间域中提供重新校准功能来捕获远程空间. 基于自注意力的方法Non-Local(Non-Local neural networks)[34]通过非局部操作捕获长期依赖关系,能够保持输入输出尺度不变. 考虑到所需注意力机制模块要满足结构简单、占用计算资源少以及图像协调化任务的特点,本文选择结合了通道域和空间域的卷积块注意力模块(Convolutional Block Attention Module, CBAM)[35].

2 本文方法

图像协调化旨在调整前景外观使其与背景外观在视觉上协调,具体来说,给定一个真实图像$\boldsymbol{I}$、前景图像$\boldsymbol{I}_f$和背景图像$\boldsymbol{I}_b$,前景区域由前景掩码$\boldsymbol{M}$表示,背景区域由背景掩码$1-\boldsymbol{M}$表示,组合图像可以表示为$\bar{\boldsymbol{I}}=\boldsymbol{M}\circ\boldsymbol{I}_f+(1-\boldsymbol{M})\circ\boldsymbol{I}_b$,其中,"$\circ$"为哈达玛积,通过图像协调化网络G将组合图像重建为协调图像$\hat{\boldsymbol{I}}=G\left(\bar{\boldsymbol{I}},\boldsymbol{M}\right)$,并通过对$\left\|\boldsymbol{I}-\hat{\boldsymbol{I}}\right\|_1$进行优化使$\hat{\boldsymbol{I}}$与$\boldsymbol{I}$尽可能地接近.

为了实现这一目标,本文利用基于双注意力机制的多分辨率选择融合模块(Multi-Resolution Selective Fusion module based on Dual Attention Mechanism, MRSF-DAM)来获取图像中丰富的语义信息以便更好地指导网络从背景中捕捉调整前景所需的特征. 其次,本文在编码器和解码器的卷积前引入了门控通道转换单元(Gated Channel Transformation units, GCT)[36]来自适应地模拟竞争和合作的通道关系,从而能提取出更

适合网络层任务的特征信息. 最后,本文还改进了以往研究者普遍使用的编解码器架构来进一步降低空间信息损失,提高了基础网络的性能. 本文协调化网络主体结构图如图1所示.

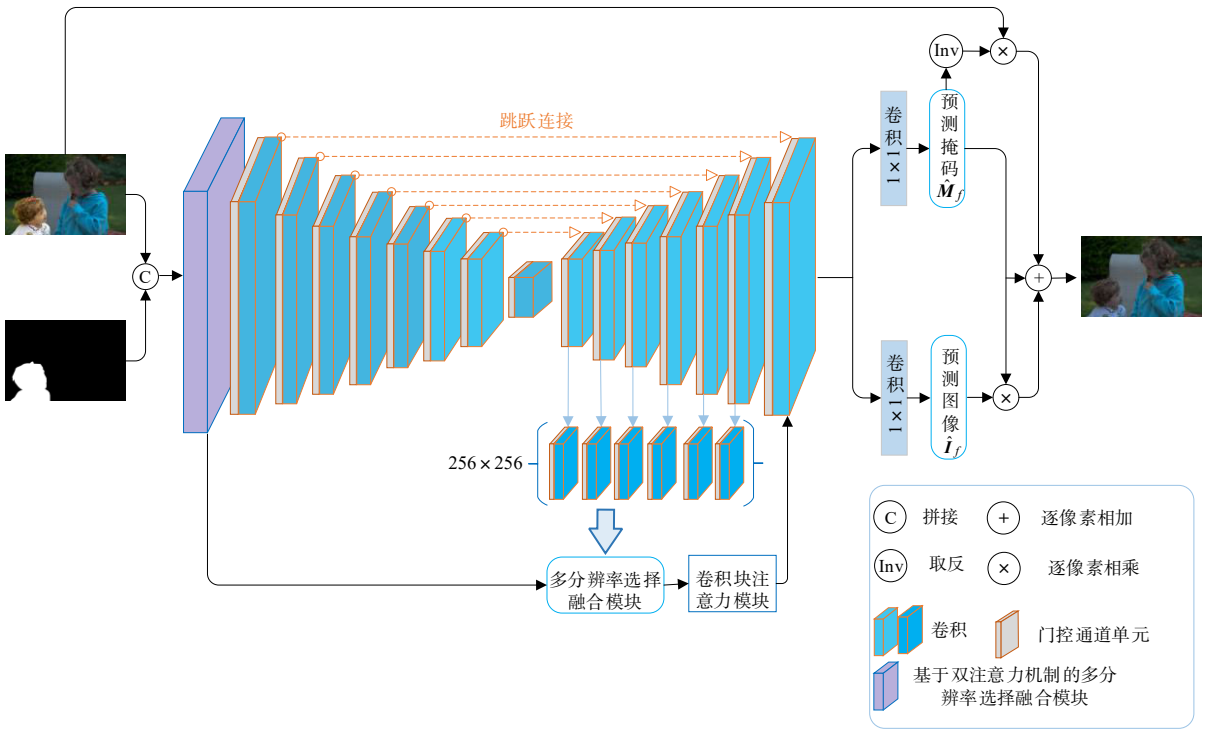


图1 本文的图像协调化网络结构图

2.1 基于双注意力机制的多分辨率选择融合模块

如图2所示,MRSF-DAM模块放置于编解码结构之前,它采用自上而下、自下而上的并行提取特征框架作为主干结构,在网络运行过程中保持着高分辨率的表征. 它实现了两种分辨率图像信息在不同尺度上的相互融合和交换,其中一层保持全分辨率处理,并行分支图像的分辨率保持为原分辨率的一半来增大图像的感受野,通过多分辨率选择融合模块来自适应地将不同感受野的特征信息选择性的进行聚合,如图3所示. 具体来说,本文用$\boldsymbol{X}\in\mathbb{R}^{H\times W\times C}$($H$,$W$和$C$分别表示输出特征图的高、宽和通道数)来表示输入特征图,通过基于残差结构的上采样和下采样方式来改变特征图的分辨率,分别用upsample和downsample来表示,给定特征图$\boldsymbol{X}\in\mathbb{R}^{H\times W\times C}$经过上采样后尺寸变大一倍,通道数变为原来的一半,下采样过程与上采样相反,得到新的特征图$\text{upsample}(\boldsymbol{X})\in\mathbb{R}^{2H\times 2W\times\frac{1}{2}C}$和$\text{downsample}(\boldsymbol{X})\in\mathbb{R}^{\frac{1}{2}H\times\frac{1}{2}W\times 2C}$. 值得注意的是,本文在上采样和下采样过程中分别聚合了双线性插值上采样和抗锯齿下采样(Anti-aliasing Downsampling, AD)[37],以防采样过程中引起的信号混叠效应. 激活函数选择高斯误差线性单元(Gaussian Error Linear Unit, GELU)[38],它引入了随机正则的思想,直观上更符合自然的认识. 他们的具体处理步骤如式(1)、式(2):

$$\begin{aligned}\text{upsample}(\boldsymbol{X})=&\text{CG}_{\text{RF}=1}(\text{CT}_{\text{RF}=3}(\text{CG}_{\text{RF}=3}(\text{CG}_{\text{RF}=1}(\boldsymbol{X})))\\&+\text{CG}_{\text{RF}=1}(\text{Bilnear}(\boldsymbol{X})))\end{aligned}\tag{1}$$

$$\begin{aligned}\text{downsample}(\boldsymbol{X})=&\text{CG}_{\text{RF}=1}(\text{AD}(\text{CG}_{\text{RF}=3}(\text{CG}_{\text{RF}=1}(\boldsymbol{X})))\\&+\text{CG}_{\text{RF}=1}(\text{AD}(\boldsymbol{X})))\end{aligned}\tag{2}$$

式中,CG表示Conv-GELU操作,CT表示反卷积操作,Bilnear表示双线性插值上采样,RF表示卷积核大小.

其次,本文采用多分辨率选择融合模块(Multi-Resolution Selective Fusion Module, MRSFM)来融合特征信息,先将两个不同分辨率的并行卷积流信息相加后采用全局平均池化(Global Average Pooling, GAP)来获取全局信息,如式(3)、式(4):

$$\hat{\boldsymbol{X}}=\boldsymbol{X}_1+\boldsymbol{X}_2\tag{3}$$

$$\boldsymbol{s}_c=\text{GAP}(\hat{\boldsymbol{X}}_c)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\hat{\boldsymbol{X}}_c(i,j)\tag{4}$$

式中,$\boldsymbol{X}_i$表示第i个分支的输出. 然后将输出的矩阵$\boldsymbol{s}\in\mathbb{R}^{1\times 1\times C}$送入全连接层(Fully Connected Layer, FCL)来融合不同层间的特征,得到特征表示$\boldsymbol{z}\in\mathbb{R}^{1\times 1\times d}$,计算

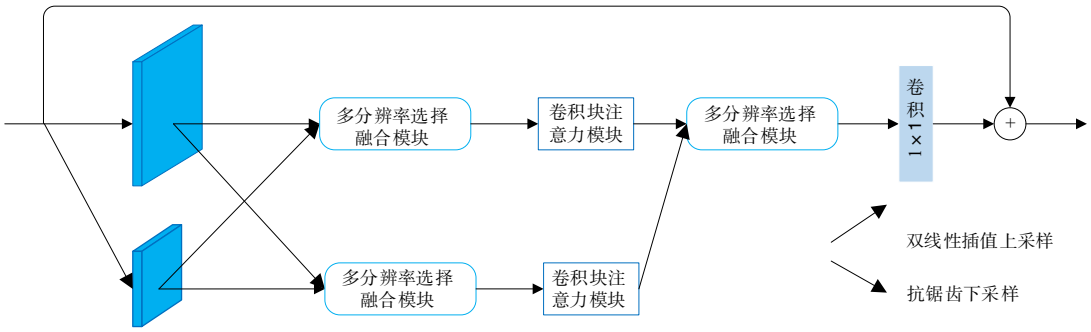


图2　基于双注意力机制的多分辨率选择融合模块结构图

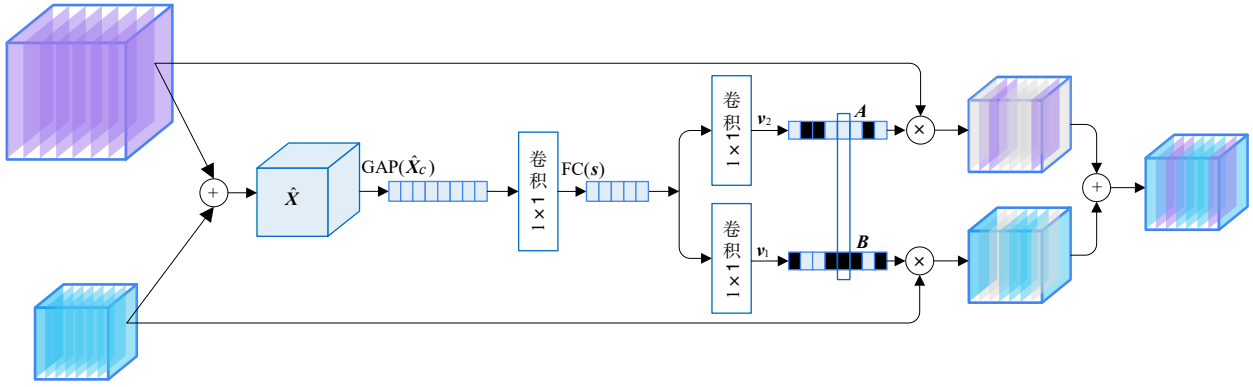


图3　多分辨率选择融合模块结构图

步骤如式(5):

$$z=\mathrm{FC}(s)=\sigma(\mathrm{Conv}_{\mathrm{RF}=1}(s)) \tag{5}$$

式中,σ为GELU激活函数.

接着,特征向量z通过并行的通道放大卷积层还原通道数,产生维度为$\mathbb{R}^{1\times1\times C}$的特征表示$v_1$和$v_2$,最后通过softmax激活函数计算不同感受野信息各层的权重矩阵A和B,通过它们来自适应地重新校准之前输入的不同分辨率的特征图,最后将校准后的特征图加权相加得到聚合后的特征图$\bar{X}\in\mathbb{R}^{H\times W\times C}$,从而实现聚合并行分支图像特征信息的同时又保留了它们独特的互补特征. 他们的具体处理步骤如式(6):

$$\bar{X}=A_c\cdot X_1+B_c\cdot X_2 \tag{6}$$

尽管MRSFM融合了跨多分辨率分支的信息,但本文还需要一种机制来共享空间和通道维度上特征张量内的信息. 基于上述需求,本文引入了双注意力机制来抑制不太有用的特征,只允许更多信息的特征进一步传递,从而减少融合特征过程中所带来的冗余信息. 因此,本文的MRSFM能够使得所有高分辨率到低分辨率的表征都有丰富的语义信息和精细的空间信息,以指导网络从背景中获取协调前景所需的各种信息,从而减小图像整体在视觉上的差异.

2.2　门控通道转换单元

本文在网络模型中引入GCT单元,具体结构如图4所示. 首先,GCT单元结合了归一化方法和注意力机制可以高效准确地给各通道的全局上下文信息建模. 由于CNN具有局部相关性机制,因此,CNN只能捕捉到视野周围的上下文,而没有全局背景上下文,这不利于模型对图像前景和背景连接关系的理解. 其次,GCT单元可以自适应地根据网络深度来创建竞争和合作的通道关系. 通常,GCT会鼓励浅层网络通道间的合作,尽可能地学习更加丰富的低级属性以捕获诸如纹理、边缘和光照等底层特征,但在更深层次中,GCT更倾向于鼓励网络通道产生竞争机制,筛选出更具辨别力的特征信息. 因此GCT能够根据网络层的任务选择性地提取特征信息. 具体来说,GCT先利用简单的l_2归一化得到了一个全局上下文嵌入算子[36],给定嵌入门控权重$\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_c)$,具体计算步骤如式(7):

$$s_c=\alpha_c\|x_c\|_2=\alpha_c\left(\left(\sum_{i=1}^{H}\sum_{j=1}^{W}(x_c^{i,j})^2\right)+\varepsilon\right)^{\frac{1}{2}} \tag{7}$$

式中,ε是一个小常数,以避免在零点求导的问题.

其次,它设计了可训练的门控权重$\chi=(\chi_1,\chi_2,\cdots,\chi_c)$和门控偏差$\beta=(\beta_1,\beta_2,\cdots,\beta_c)$. 当通道的门限权重为正时,GCT取得通道竞争关系;当通道的门限权重为负时,GCT取得通道协同作用. 具体实现步骤如下:

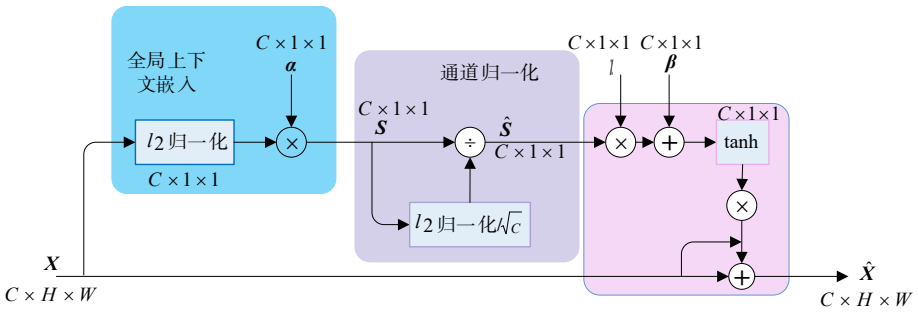


图4 门控通道转换单元结构图

$$\hat{\boldsymbol{s}}_c=\frac{\sqrt{C}\,\boldsymbol{s}_c}{\|\boldsymbol{s}\|_2}=\frac{\sqrt{C}\,\boldsymbol{s}_c}{\left(\left(\sum_{C=1}^{C}\boldsymbol{s}_c{}^2\right)+\xi\right)^{\frac{1}{2}}} \tag{8}$$

$$\hat{\boldsymbol{x}}_c=\boldsymbol{x}_c\left(1+\tanh\left(\boldsymbol{\chi}_c\hat{\boldsymbol{s}}_c+\boldsymbol{\beta}_c\right)\right) \tag{9}$$

式(8)中,ξ为一个小常数,以避免分母为零的问题.

2.3 编码器解码器网络

本文采用图像协调化领域常用的编解码网络作为本文主干结构. 为了避免图像模糊和纹理细节丢失,本文在编码器和解码器之间采用了跳过连接,这样可以同时考虑到高层特征和低层特征. 尽管编解码结构采用端到端的方式不仅能够提取到广泛的上下文信息,同时还能够滤除部分噪声的干扰,但下采样会丢失空间细节信息容易造成空间上不太准确的输出. 为缓解这个问题,本文提出了一种基于编解码器的网络结构,如图1所示. MRSF-DAM输出特征图的尺寸大小为256×256,本文通过反卷积操作将解码器中除了最后一层外的每层网络层的输出特征图尺寸放大到256×256,然后通过MRSFM将她们从多个角度有选择性地聚合,再将输出送入卷积块注意力模块[35]中进一步细化,最后将细化后的输出叠加解码器最后一层的输出作为编解码网络的总输出. 采用这种结构不仅可以将不同感受野的特征信息融合,提供更好的上下文信息,还可以减少空间上的信息损失,弥补了编解码结构的缺陷. 本文在网络架构命名前加上“i”代表本文设计的编解码结构,因此本文将总体网络命名为iMRSF-DAM-GCT.

2.4 网络细节

本文的协调化网络结构中,第一层是基于双注意力机制的多分辨率选择融合模块,编码器和解码器一共7层,均以2为因子进行卷积上采样和反卷积下采样,卷积核大小均为4×4. 其中编码器前面两层是带有激活函数的标准卷积层,后面接有5个网络层,包含卷积层、批量归一化(Batch Normalization,BN)[38]层和指数线性单元. 解码器有7层网络层,包含反卷积层、BN层和GELU激活函数. 解码器前面6层通过反卷积操作将特征图尺寸放大到256×256. 对于解码器最后一层输出的特征图,本文分别通过两个1×1的卷积层预测出前景掩码$\hat{\boldsymbol{M}}_f$和RGB三通道的预测协调图像. 接着,本文利用前景掩码$\hat{\boldsymbol{M}}_f$提取出预测协调图像的前景,并将预测的$\hat{\boldsymbol{M}}_f$取反后与组合图像相乘得到组合图像的背景,然后将提取到的前景和背景进行组合得到最终的网络协调图像.

2.5 损失函数

由于图像协调化旨在调整前景区域外观,而前景区域面积的变化对图像协调化的性能影响较大,因此本文认为使用前景归一化均方误差(Foreground Normalized Mean Square Error,FN-MSE)[29]作为本文的损失函数更加合理,如式(10)所示:

$$L_{\text{res}}=\frac{\sum_{h,w}\left\|\hat{\boldsymbol{I}}_{h,w}-\boldsymbol{I}_{h,w}\right\|_2^2}{\max\{A_{\min},\sum_{h,w}\boldsymbol{M}_{h,w}\}} \tag{10}$$

式中,$\boldsymbol{M}_{h,w}$表示前景掩码,$\hat{\boldsymbol{I}}_{h,w}$表示网络输出的协调图像,$\boldsymbol{I}_{h,w}$表示对应的真实图像,$A_{\min}$是用于防止训练期间不稳定的超参数,本文按照文献[29]中的建议将其设置为100.

3 实验结果与分析

3.1 数据集与评估指标

为了评估本文方法在图像协调化方面的性能,本文在广泛使用的iHarmony4公开数据集[21]上进行了实验. 该数据集由4个子数据集组成,包括HCOCO、HAdobe5k、HFlickr和Hday2night,共包含73 147组前景掩码图像、合成图像和对应的真实图像. 在这项工作中,本文遵循与DoveNet[21]相同数量的训练集测试集划分,65 742组用于训练,7 404组用于测试.

本文用均方误差(Mean Square Error,MSE)、前景均方误差[35](foreground MSE,fMSE)和峰值信噪比(Peak Signal to Noise Ratio,PSNR)来评估本文模型性能. MSE本质上测量了整个数据集中所有像素的平均误差,而

fMSE仅计算前景区域MSE,更加适合在背景区域像素不变的情况下对图像协调化的结果真实性进行衡量.其中,MSE和fMSE值越小,PSNR值越大,代表模型性能越好.

3.2　网络训练细节

本文以Adam优化器来训练模型,参数为$\beta_1=0.9$,$\beta_2=0.999$,总共120个epoch. 设置初始学习率为1e−3,并在第55、81、105和115个epoch进行自动更新,衰减系数为$\gamma=0.5$. 本文的训练样本通过水平翻转和随机大小裁剪来增强模型的泛化能力,并将输入图像尺寸调整到256×256. 此外,本文的实验环境为PyTorch=3.8.2、CUDA=11.2和4张Nvidia GTX 1080Ti GPU.

3.3　与已有方法对比结果

3.3.1　定量结果对比

表1显示了本文方法与最新图像协调化方法,包括DoveNet[21]、RainNet[23]、SCS-Co[24]和内在图像分解算法D-HT[25]、IIH[26],DIH[28]、以及S^{2}AM[31]的定量结果. 可以从表1中观察到,本文的方法除在HFlickr子数据集上的PSNR值之外,在其余子数据集以及整个数据集的平均评估分数都已经超过目前最新的方法. 在整个数据集的平均评估分数PSNR中实现了1 dB、MSE中4.28和fMSE中61.97的平均性能增益. 此外,本文根据前景区域占整幅图像的比例来对其进行划分,分别是0~5%,5%~15%和15%~100%,评估结果如表2所示.

表1　不同方法在iHarmony4测试集上的定量对比

Dataset	HCOCO			HAdobe5k			HFlickr			Hday2night			Average		
Metric	PSNR	MSE	fMSE	PSNR	MSE	fMSE	PSNR	MSE	fMSE	PSNR	MSE	fMSE	PSNR	MSE	fMSE
DoveNet[21]	35.83	36.72	—	34.34	52.32	—	30.21	133.14	—	35.18	54.05	—	34.75	52.36	—
RainNet[23]	37.08	—	—	36.22	—	—	31.64	—	—	34.83	—	—	36.12	—	—
SCS-Co[24]	39.88	13.58	245.54	38.29	21.01	165.48	**34.22**	55.83	393.72	37.83	41.75	606.80	38.75	21.33	248.86
D-HT[25]	38.76	16.89	299.30	36.88	38.53	265.11	33.13	74.51	515.45	37.10	53.01	704.42	37.55	30.30	320.78
IIH[26]	37.16	24.92	416.38	35.20	43.02	284.21	31.34	105.13	716.60	35.96	55.53	797.04	35.90	38.71	400.29
DIH[21,28]	34.69	51.85	—	32.28	92.65	—	29.55	163.38	—	34.62	82.34	—	33.41	76.77	—
S^{2}AM[21,31]	35.47	41.07	—	33.77	63.40	—	30.03	143.45	—	34.50	76.61	—	34.35	59.67	—
本文方法	**40.74**	**11.16**	**201.96**	**40.08**	**12.74**	**92.56**	33.92	**55.36**	**307.97**	**39.15**	**38.13**	**479.70**	**39.75**	**17.05**	**186.89**

表2　不同方法在iHarmony4测试集上的不同前景比率范围的MSE和fMSE指标对比

前景比率	0~5%		5%~15%		15%~100%		0~100%	
Metric	MSE	fMSE	MSE	fMSE	MSE	fMSE	MSE	fMSE
DoveNet[21]	14.03	591.88	44.90	504.42	152.07	505.82	52.36	549.96
RainNet[23]	11.66	550.38	32.05	378.69	117.41	389.80	40.29	469.60
DIH[21,28]	18.92	799.17	64.23	725.86	228.86	768.89	76.77	773.18
S^{2}AM[21,31]	15.09	623.11	48.33	540.54	177.62	592.83	59.67	594.67
本文方法	**5.06**	**215.55**	**14.22**	**159.03**	**48.20**	**155.18**	**17.05**	**186.89**

注:由于DIH和S^{2}AM方法在原始论文中没有给出在iHarmony4数据集上的实验结果,本文引用了文献[21]中对应的数据结果放入表1和表2中.

从表2的结果可以看出本文方法在具有不同前景区域比例的数据上都达到了最优的性能. 以上实验结果验证了本文方法的有效性.

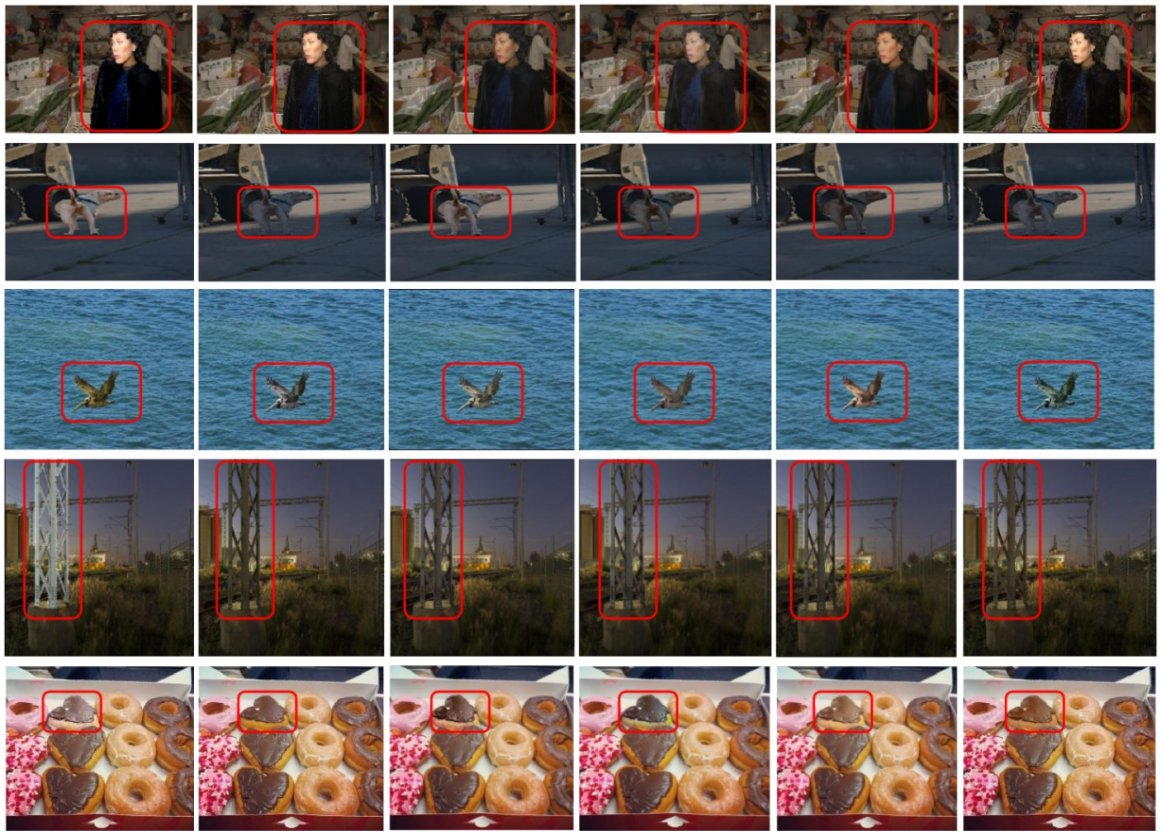
3.3.2　定性结果对比

本文进一步展示了iHarmony4的定性比较结果,如图5所示,其中红色实线框出部分为前景区域,且本文方法预测的前景在颜色和光照上明显比其他方法更加接近真实情况,能将前景对象更好地融合到背景图像中. 如图5前三行所示,前景和背景没有相同内容时,本文方法的视觉效果整体上都比较接近于真实图像,而其他方法的局部前景外观与真实图像有着不小的差距,包括前景对象的脸部,脖子和衣服等. 如图5最后一行所示,前景和背景存在相似内容的情况下,其他方法预测的前景蛋糕在颜色上偏绿色和白色更多,这可能是网络没有学习到正确语义导致的结果. 其次,与其他方法相比,本文能恢复更多的局部细节信息,实现了更逼真的输出. 如图5第四行所示,可以看出RainNet[23]和D-HT[25]输出图像的前景边缘存在明显白边,而本文方法协调出来的结果很少有这样的情况,这依赖于本文提出的MRSF-DAM模块,它通过聚合多尺度图像特征使得网络更能注重局部细节信息.

基于以上论述,本文方法验证了语义信息能引导网络更加充分地挖掘出调整前景所需的背景特征,能预测出更加逼近真实地面的图像,更适合图像协调化任务.

3.4　消融实验

本文做了对应的消融实验来验证所提出模块和结构的有效性,如表3所示. 本文的baseline是普通的7层编解码结构,“MRSF-DAM”代表在baseline的基础上仅加入MRSF-DAM模块的情况,它比baseline在PSNR、MSE、fMSE中分别提升了1.64 dB、8.95和98.48. 然后,本文在“MRSF-DAM”的基础上再加入GCT模块,从表3第三列可以看出,它在整体数据集上的平均性能排到了第二,仅次于本文提出的最终模型“iMRSF-DAM-



(*a*) 组合图像 (*b*) 真实图像 (*c*) DoveNet[3] (*d*) RainNet[8] (*e*) D-HT[10] (*f*) 本文方法

图5 不同方法在iHarmony4测试集上的定性结果对比

GCT". 观察表3可以发现,加入GCT单元和改变网络架构都会对模型性能有所提升.

接着,本文进一步探索MRSF-DAM模块中支路对性能的影响,如表4所示. K=3代表三条不同分辨率卷积流结构的"iMRSF-DAM-GCT",K=1代表单分支模型,K=2代表本文方法. 从表4可以看出,三支路卷积流结构和单支路卷积流结构均会降低本文的网络模型精度,因此不是分辨率特征交互次数越多,效果就越好,它反而可能引起网络冗余信息增加,甚至让网络关注错误的信息.

通过这些消融实验验证了本文所做的工作在图像协调化领域是有意义的.

表3 不同模块在iHamony4测试集上的消融实验

Dataset	HCOCO			HAdobe5k			HFlickr			Hday2night			Average		
Metric	PSNR	MSE	fMSE	PSNR	MSE	fMSE	PSNR	MSE	fMSE	PSNR	MSE	fMSE	PSNR	MSE	fMSE
baseline	38.25	20.13	341.02	36.78	30.02	215.91	32.40	88.81	588.44	37.21	47.14	711.45	37.15	31.18	338.84
MRSF-DAM	39.59	14.71	239.32	38.75	19.49	145.35	33.93	62.90	406.06	38.13	**34.13**	565.34	38.69	21.84	236.41
MRSF-DAM-GCT	40.02	12.87	220.96	39.19	16.17	128.53	**34.24**	55.88	373.93	38.09	37.25	531.63	39.10	19.08	216.68
iMRSF-DAM-GCT	**40.74**	**11.16**	**201.96**	**40.08**	**12.74**	**92.56**	33.92	**55.36**	**307.97**	**39.15**	38.13	**479.70**	**39.75**	**17.05**	**186.89**

表4 不同并行支路在iHamony4测试集上的消融实验

Dataset	HCOCO			HAdobe5k			HFlickr			Hday2night			Average		
Metric	PSNR	MSE	fMSE	PSNR	MSE	fMSE	PSNR	MSE	fMSE	PSNR	MSE	fMSE	PSNR	MSE	fMSE
K=1	39.88	13.52	249.38	39.34	15.34	107.98	32.93	65.91	382.80	**39.55**	**33.66**	441.71	38.49	20.27	226.51
K=2	**40.74**	**11.16**	**201.96**	**40.08**	**12.74**	**92.56**	**33.92**	**55.36**	**307.97**	39.15	38.13	479.70	**39.75**	**17.05**	**186.89**
K=3	39.84	13.98	253.85	39.33	15.23	108.87	33.01	66.06	374.04	39.22	34.53	**439.74**	38.92	20.54	228.33

4 总结

在本文中,本文设计了一种新的基于双注意力的多分辨率选择融合模块,能充分利用与前景具有相关内容的背景信息来缩小不协调区域的外观差异. 此外,本文还设计了一种新的基于编解码结构的网络架构,它能减轻传统编解码下采样所带来的空间精度损失. 在未来工作中,将进一步提升图像协调化的真实感,加快运行速度,以便生成更高质量的电影场景图,提升相机美颜功能与人脸的贴合度以及对其他自动化图像处理领域产生更多的贡献.

参考文献

[1] CHU L T, LIU Y, WU Z W, et al. Pp-humanseg: Connectivity-aware portrait segmentation with a large-scale teleconferencing video dataset[C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). Piscataway: IEEE, 2022: 202-209.

[2] GAO Qi-fan, WU Xiao-lin. Real-time deep image retouching based on learnt semantics dependent global transforms[J]. IEEE Transactions on Image Processing, 2021, 30: 7378-7390.

[3] CHEN B C, KAE A. Toward realistic image compositing with adversarial learning[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 8407-8416.

[4] 徐少平, 陈孝国, 李芬, 等. 采用两阶段混合策略实现的低照度图像增强算法[J]. 电子学报, 2021, 49(11): 2166-2170.
XU Shao-ping, CHEN Xiao-guo, LI Fen, et al. A low-light image enhancement algorithm using two-stage hybrid strategy [J]. Acta Electronica Sinica, 2021, 49(11): 2166-2170. (in Chinese)

[5] IIZUKA S, SIMO-SERRA E, ISHIKAWA H. Globally and locally consistent image completion[J]. ACM Transactions on Graphics, 2017, 36(4): 1-14.

[6] ZHENG C, CHAM T J, CAI J, et al. Bridging global context interactions for high-fidelity image completion[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 11502-11512.

[7] BULAT A, YANG J, TZIMIROPOULOS G. To learn image super-resolution, use a GAN to learn how to do image degradation first[C]//Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 187-202.

[8] 王相海, 赵晓阳, 王鑫莹, 等. 非抽取小波边缘学习深度残差网络的单幅图像超分辨率重建[J]. 电子学报, 2022, 50(7): 1753-1765.
WANG Xiang-hai, ZHAO Xiao-yang, WANG Xin-ying, et al. Single image super-resolution reconstruction using deep residual networks with non-decimated wavelet edge learning [J]. Acta Electronica Sinica, 2022, 50(7): 1753-1765. (in Chinese)

[9] 周登文, 李文斌, 李金新, 等. 一种轻量级的多尺度通道注意图像超分辨率重建网络[J]. 电子学报, 2022, 50(10): 2336-2346.
ZHOU Deng-wen, LI Wen-bin, LI Jin-xin, et al. Image super-resolution reconstruction based on lightweight multi-scale channel attention network[J]. Acta Electronica Sinica, 2022, 50(10): 2336-2346. (in Chinese)

[10] 李大锦, 高文冉, 高俊杰. 基于kuwahara滤波的视频风格化框架[J]. 电子学报, 2020, 48(3): 538-544.
LI Da-jin, GAO Wen-ran, GAO Jun-jie. Artistic video stylization based on kuwahara filter[J]. Acta Electronica Sinica, 2020, 48(3): 538-544. (in Chinese)

[11] AN J, HUANG S Y, SONG Y B, et al. ArtFlow: unbiased image style transfer via reversible neural flows[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 862-871.

[12] WU X, HU Z, SHENG L, et al. Styleformer: Real-time arbitrary style transfer via parametric style composition[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 14618-14627.

[13] REINHARD E, ADHIKHMIN M, GOOCH B, et al. Color transfer between images[J]. IEEE Computer Graphics and Applications, 2001, 21(5): 34-41.

[14] XIAO X Z, MA L Z. Color transfer in correlated color space [C]//Proceedings of the 2006 ACM International Conference on Virtual Reality Continuum and Its Applications. New York: ACM, 2006: 305-309.

[15] FECKER U, BARKOWSKY M, KAUP A. Histogram-based prefiltering for luminance and chrominance compensation of multiview video[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2008, 18(9): 1258-1267.

[16] PITIÉ F, KOKARAM A C, DAHYOT R. Automated colour grading using colour distribution transfer[J]. Computer Vision and Image Understanding, 2007, 107(1/2): 123-137.

[17] SUNKAVALLI K, JOHNSON M K, MATUSIK W, et al. Multi-scale image harmonization[J]. ACM Transactions on Graphics, 2010, 29(4): 1-10.

[18] SONG S B, ZHONG F, QIN X Y, et al. Illumination Harmonization with Gray Mean Scale[M]//Advances in Computer Graphics. Cham: Springer International Publishing, 2020: 193-205.

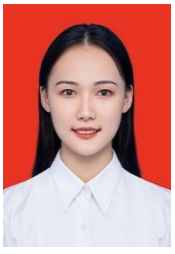
[19] LALONDE J F, EFROS A. Using color compatibility for assessing image realism[C]//2007 IEEE 11th International Conference on Computer Vision (CVPR). Piscataway:IEEE, 2007: 1-8.

[20] XUE S, AGARWALA A, DORSEY J, et al. Understanding and improving the realism of image composites[J]. ACM Transactions on Graphics, 2012, 31(4): 1-10.

[21] CONG W, ZHANG J, NIU L, et al. Dovenet: deep image harmonization via domain verification[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition

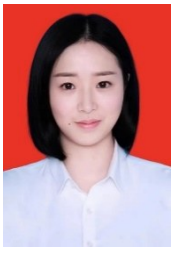
(CVPR). Piscataway: IEEE, 2020: 8391-8400.
[22] CONG W, NIU L, ZHANG J, et al. BargainNet: Background-guided domain translation for image harmonization[C]//2021 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE, 2021: 1-6.
[23] LING J, XUE H, SONG L, et al. Region-aware adaptive instance normalization for image harmonization[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 9357-9366.
[24] HANG Y, XIA B, YANG W, et al. Scs-co: Self-consistent style contrastive learning for image harmonization[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 19678-19687.
[25] GUO Z, GUO D, ZHENG H, et al. Image harmonization with transformer[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 14850-14859.
[26] GUO Z, ZHENG H, JIANG Y, et al. Intrinsic image harmonization[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 16362-16371.
[27] JIANG Y, ZHANG H, ZHANG J, et al. SSH: A self-supervised framework for image harmonization[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 4812-4821.
[28] TSAI Y H, SHEN X, LIN Z, et al. Deep image harmonization [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 2799-2807.
[29] SOFIIUK K, POPENOVA P, KONUSHIN A. Foreground-aware semantic representations for image harmonization[C]//2021 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2021: 1619-1628.
[30] CONG W, TAO X, NIU L, et al. High-resolution image harmonization via collaborative dual transformations[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 18449-18458.
[31] CUN X D, PUN C M. Improving the harmony of the composite image by spatial-separated attention module[J]. IEEE Transactions on Image Processing, 2020, 29: 4759-4771.
[32] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.
[33] HU J, SHEN L, ALBANIE S, et al. Gather-excite: Exploiting feature context in convolutional neural networks[C]. Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM, 2018: 9423-9433.
[34] WANG X L, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2018: 7794-7803.
[35] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.
[36] YANG Z X, ZHU L C, WU Y, et al. Gated channel transformation for visual recognition[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 11791-11800.
[37] ZHANG R. Making convolutional networks shift-invariant again[C]//Proceedings of International Conference on Machine Learning. New York: ACM, 2019: 7324-7334.
[38] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [C]// Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. New York: ACM, 2015: 448-456.

作者简介



杨紫媛 女,1998年5月出生于重庆市. 现为重庆邮电大学通信与信息工程学院研究生. 主要研究方向为图像协调化、计算机视觉和机器学习.
E-mail: 785459971@qq.com

李鹏程 男,1995年12月出生于重庆市,现为重庆邮电大学通信与信息工程学院博士生. 主要研究方向为智能医学影像分析、计算机视觉和机器学习.
E-mail: lipengchengme@163.com



刘芳岑 女,1995年出生于重庆市,现于重庆邮电大学攻读博士学位. 主要研究方向为红外小目标检测.
E-mail: liufc67@gmail.com

高陈强(通讯作者) 男,1981年8月生于重庆市,现为重庆邮电大学人才工作办公室副主任、教授、博士生导师. 主要研究方向为图像处理、视频分析和机器学习.
E-mail: gaocq@cqupt.edu.cn