红外与激光工程
INFRARED AND LASER ENGINEERING
EI, SCOPUS, CSCD等收录期刊

基于并行多轴自注意力的图像去高光算法

李鹏越　续欣莹　唐延东　张朝霞　韩晓霞　岳海峰

Image highlight removal method based on parallel multi-axis self-attention

Li Pengyue, Xu Xinying, Tang Yandong, Zhang Zhaoxia, Han Xiaoxia, Yue Haifeng

在线阅读 View online: https://doi.org/10.3788/IRLA20230538

基于并行多轴自注意力的图像去高光算法

李鹏越[1,2*]，续欣莹[1]，唐延东[3,4]，张朝霞[1]，韩晓霞[1]，岳海峰[2]

(1. 太原理工大学 电气与动力工程学院，山西 太原 030024；
2. 太原重型机械 (集团) 有限公司，山西 太原 030027；
3. 中国科学院沈阳自动化研究所 机器人学国家重点实验室，辽宁 沈阳 110016；
4. 中国科学院机器人与智能制造创新研究院，辽宁 沈阳 110016)

摘　要：图像高光层模型的模糊性和高光动态范围大的特点，使得图像去高光成为了一个挑战性的视觉任务。纯局部性方法容易导致图像高光区出现伪影，纯全局性方法容易使图像非高光区色彩失真。针对图像去高光中局部和全局特征不平衡导致的上述问题，以及高光层建模的模糊性，提出了基于并行多轴自注意力机制的门限融合 U 型深度网络图像去高光算法。该方法通过隐式建模避免了高光层模型模糊引入的问题，利用 U 型网络结构将上下文信息与低层信息融合对无高光图像进行估计，并在 U 型结构编码器和解码器之间引入门限融合结构进一步提升网络模型的特征表达能力。此外，U 型网络的单元结构通过融合局部和全局自注意力平衡了局部和全局特征的编码和解码。定性实验结果表明，文中方法可以更有效地去除图像中的高光，其他对比算法在高光处容易产生伪影和失真。定量实验结果表明，文中方法在 PSNR 和 SSIM 指标上优于其他五种典型的图像去高光方法，在三个数据集上，PSNR 值分别高于次优方法 4.10、7.09、6.58 dB，SSIM 值分别取得了 4%、9%和 3%的增量。

关键词：图像处理；　去高光；　多轴自注意力；　深度学习

中图分类号：TP391.4　　**文献标志码：**A　　**DOI：**10.3788/IRLA20230538

0　引　言

高光是生活中常见的一种光学物理现象，通常表现为光照作用下有光泽材料表面的高亮点。场景成像后图像中的高亮点会对背景信息形成不同程度的遮挡，特别是在一些文字图像中高光容易引起图像关键信息的丢失。因此，图像去高光一直是计算机视觉和图像处理领域的基本问题。通过去除图像中的高光不仅可以复原图像中丢失的关键信息，而且可以提高许多计算机视觉任务的性能，如图像分类[1-2]、本征图像分解[3-4]、图像分割[5-7]、目标检测[8-9]、行为识别[10-11]和目标跟踪[12-15]。

早期的图像去高光算法通常是基于不同先验信息约束对无高光图像进行最优估计，进而对高光进行去除，例如双色反射模型[16-17]、稀疏先验[18]、暗通道先验[19]、色度分布先验[20-22]和滤波[17, 23]等约束条件。这类方法由于先验信息的局部性，无法去除大范围的高光，且它们过度依赖先验信息，容易将图像中白色像素误认为高光区域，将白色像素误去除，导致这些方法的准确性降低。自然光照图像中存在丰富的纹理、复杂的材料表面和阴影，这就导致图像高光层和非高光层模型的模糊性，这种模糊性容易给经典的去高光算法引入模型误差，所以图像去高光仍然是一个具有挑战性的图像复原任务。近年来，随着深度学习在视觉领域的迅速发展，出现了一些基于深度学习的图像去高光方法。尽管目前基于深度学习的图像去高光方法已经取得了显著的进展[24-30]，但它们仍然存在一些局限性。首先，这些方法通常是在合成数据或少量的真实数据上进行训练，训练数据和测试数据之间的域差异，可能导致这些方法在真实高光图像上去

收稿日期：2023-09-19；　**修订日期：**2023-11-13

基金项目：国家自然科学基金项目 (62203319)；山西省自然科学基金项目 (202203021212220；202103021224056)；山西省科技合作交流专项 (202104041101030)

通讯作者：李鹏越，男，讲师，硕士生导师，博士，主要从事计算机视觉方面的研究。

高光的泛化性不强。为此,文中在较大的真实高光图像数据库上对深度网络进行了训练,以提升深度网络模型的泛化性。其次,现有的去高光深度网络主要是通过设计新的卷积拓扑结构来提升算法性能。这类网络模型无法避免纯卷积结构的缺陷(归纳偏置对模型的局限、固定卷积核对感受野的局限、特征局部性对语义的局限等),并且结构越来越复杂。引入新的卷积结构对基于深度网络的图像去高光方法在理论意义和实际效果上都出现了瓶颈。为此,文中采用卷积和自注意力机制相融合的复合结构避免纯卷积结构的固有缺陷,提升基于深度网络的图像去高光算法的性能。

综上,文中开发了一种基于多轴自注意力机制的U型深度去高光网络模型(UMAVTB)。采用了较大规模的真实高光图像数据集对图像去高光深度网络算法进行了训练和综合评估。同时,为了避免图像高光层和非高光层模型的模糊性给经典的去高光算法引入的误差,采用了变量归一化的非线性高光图像模型。此外,为了避免纯卷积网络的局限性,开发了一种基于多轴自注意力的U型深度去高光网络模型。该深度网络通过并行多轴自注意力机制的U型门限结构提取高光图像高低层的局部和全局特征,以准确地去除不同范围和特征的高亮区域。在自然图像和文字图像数据集上的实验结果表明,文中的深度网络模型在高光去除方面优于目前典型的方法。

1 高光去除算法

1.1 高光图像模型

高光去除领域有三种被广泛使用的典型模型,一种是双色反射模型,该模型将彩色图像I分解为漫反射图像和高光图像的线性组合:

$$I = D + H \tag{1}$$

式中:D为漫反射图像(无高光层);H为高光层。第二种流行的高光图像模型是在双色反射模型上增加了区分高光区域和非高光区域掩膜的非线性模型:

$$I = D + M \otimes H \tag{2}$$

式中:M为高光掩膜;$\otimes$为逐元素相乘。通过引入掩膜可以避免出现饱和模糊问题,即饱和度较低的像素被误认为高光去除,导致非高光区域颜色失真,如白色材料表面。另一种被广泛采用的高光图像模型是在无高光图像进行建模的基础上形成的一种基于本征图像的非线性高光图像模型:

$$I = R \otimes S + H \tag{3}$$

式中:R为反射系数(本征图像);S为阴影;基于上述三种模型,传统的去高光算法将高光去除视为信号分离问题,即给定高光图像I,依据高光层和无高光层的先验信息,在基于先验信息约束下获得无高光图像的最佳估计。该信号分离问题在数学上属于病态问题。现实场景中的高光亮度值范围通常很广,且空间分布不同,所以传统的基于优化的方法利用平滑先验,并不能有效地对高光层建模以产生令人满意的结果。这些方法的有效性很大程度上依赖于先验信息的准确性。从本质上讲,造成上述问题的主要原因在于无高光层和高光层建模固有的模糊性。

为了正确分解高光图像的同时能保留更多的图像细节信息,避免显式模型优化方法的局限性,文中对高光图像进行了隐式建模,并利用深度学习强大的特征提取能力和非线性拟合能力,开发了基于多轴自注意力机制的图像高光去除深度网络算法。将高光图像建模为:

$$I = f(R,S,H) \tag{4}$$

式中:$f()$为从不同物理量特征层到高光图像的映射函数,并设$f()$的反操作是$g()$。若采用不同的子网络对漫反射层、阴影和高光掩膜等变量分别进行估计,每个子网络的估计结果都存在一定的误差,如果根据这些带有误差的估计代入显式模型求解高光图像,可能会进一步放大一起使用时的误差,即将多个物理量进行非联合的估计可能会进一步放大一起使用时的误差,所以文中采用多变量归一化的分解模型方法,进行端到端的从高光图像到正常光照图像的映射学习,该深度网络学习模型求解过程可以表示为:

$$D = g(I) \tag{5}$$

1.2 去高光深度网络

目前主流的图像去高光深度网络结构几乎都是纯卷积,由于卷积结构自身的局限性,基于纯卷积结构的图像去高光方法在理论意义和实际效果上都出

现了瓶颈。为此，文中追随了目前一些主流方法[31-34]中将卷积和注意力融合的设计思路，进行了卷积和多轴自注意力的融合设计，采用卷积和自注意力机制相融合的复合结构避免纯卷积深度网络的缺陷。对于图像去高光任务，不仅需要低层的位置和光强信息，也需要上下文语义信息引导高光区域图像的复原。为了精准定位和提取丰富的上下文信息，利用 U 型网络框架的收缩路径捕获高光区域上下文信息和对称的扩展路径对高光区域精确定位。同时，为了避免图像去高光中局部和全局特征提取不平衡导致的图像整体光照不自然和非高光区色彩的失真，设计了一种新颖的多轴自注意力机制以自动调整通道特征的权重，进而平衡局部特征和全局特征的提取和解码。文中将该去高光深度网络简称为 UMAVTB，其整体结构如图 1 所示。

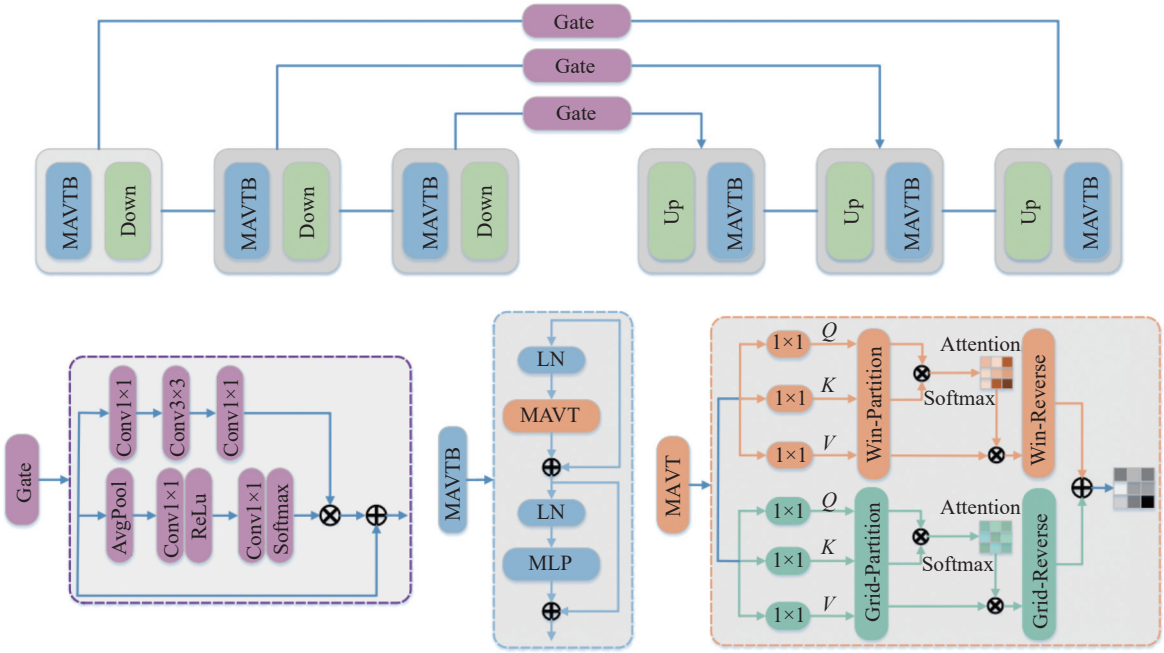


图 1 网络整体结构图

Fig.1 The overall structure of the proposed network

1.2.1 U 型去高光网络

特征提取是语义理解的基础，有效的特征提取结构能够提升网络的隐式表达能力。典型的深度网络结构是将多层卷积层顺序地串联起来并最终通过全连接层输出学习结果。这种结构随着层数的递进增加，网络在后端层提取的是高层语义信息，但丢失了图像的一些低层信息，并且随着网络层数的增加容易出现梯度爆炸。U 型网络采用一种收缩卷积的策略能够更快地提取上下文语义信息，降低了提取上下文语义信息的层数；同时，为了解决连续卷积深度网络丢失低层信息的问题，U 型网络在收缩路径的后面连接了一个对称的扩张路径，通过扩张的方式来逐步恢复图像的低层信息，并将收缩路径上各阶段的特征对应串联到扩张路径上；最终扩张路径输出的是包含高层上下文语义信息和低层信息的综合信息，可有效解决连续型深度网络丢失低层信息的问题。U 型网络既可以充分利用每一卷积层的特征，又能避免层数增加时对低层信息的丢失。具体地，U 型网络第 i 阶段收缩路径的输出可以描述为：

$$U_c^i = f_c(U_c^{i-1}), i = 1,2,3 \tag{6}$$

式中：U_c^0为网络的输入特征图；$f_c(\cdot)$为两个连续操作的复合函数：多轴自注意力块 (MAVT) 和下采样。对称地，扩展路径第 j 阶段的输出可描述为：

$$U_e^j = f_e([U_e^{j-1}, U_c^{7-j}]), j = 4,5,6 \tag{7}$$

式中：U_e^3为收缩路径的输出；$[U_e^{j-1}, U_c^{7-j}]$为串联操作；

$f_e(\cdot)$代表了两个连续操作的复合函数：上采样和多轴自注意力块 (MAVT)。

1.2.2 门限特征融合结构

在编码器和解码器之间采用门限机制来调节编码器各通道的信息流，这使的编码器尽可能在通道层面关注高光特征。通道门限结构首先对输入特征进行高低频解耦和特征提取，然后通过逐点相乘将两类特征进行融合，最后采用残差模式对低层特征进行补充学习。门限结构特征提取和融合过程可以表示为：

$$\begin{aligned} O_1 &= f_{\text{conv1}}(f_{\text{conv3}}(f_{\text{conv1}}(X))) \\ O_2 &= f_{\text{softmax}}(f_{\text{conv1}}(f_{\text{conv1}}(f_{\text{AvgPool}}(X)))) \\ O &= X + O_1 \otimes O_2 \end{aligned} \tag{8}$$

式中：$f_{\text{conv}m}(\cdot)$为卷积核大小为 m 的卷积操作；$f_{\text{softmax}}(\cdot)$为 softmax 激活函数；$f_{\text{AvgPool}}(\cdot)$为全局最大池化操作。

1.2.3 并行多轴自注意力模块

Transformer 最近在计算机视觉领域获得了极大的关注。但是 Transformer中的自注意力机制缺乏对图像尺寸的自适应性，这种缺陷限制了它们在视觉任务中的广泛应用。为此，文中采用了一种由尺度自适应的多轴自注意力机制构成的 Transformer，其包括局部和全局两种自注意力模块，将其简称为 MAVTB。这种多轴自注意力机制允许在任意尺寸图像上进行全局和局部空间的交互关注，并且计算复杂度是线性的。与局部卷积相比，全局相互作用是自注意力的关键优势之一。然而，标准的自注意力算子需要二次计算复杂度，直接沿整个空间施加自注意力会消耗大量的计算资源。为了降低自注意力的计算复杂度，文中将全尺寸注意力分解为不同维度下的两种稀疏形式的局部和全局自注意力模块，对特征进行不同感受野下的自注意力交互关注。此外，标准的自注意力机制是不具有平移不变性的，这是卷积神经网络在视觉任务中起作用的关键，因此，文中通过有效地将多轴自注意力模块与卷积融合来提升网络的特征提取和表达能力。

局部注意力：设 $X \in R^{n\times h\times w\times c}$ 为输入特征映射，局部注意力不是将注意力集中在平坦的空间维度 $H\times W$ 上，而是将特征图划分为不重叠的窗口，即形状为 $(H/P\times W/P, P\times P, C)$ 的张量，每个窗口的大小为 $P\times P$。将自注意力集中在局部空间维度 $P\times P$ 的一个小窗口内进行局部交互形成局部注意力。该窗口划分过程可以表示为如下过程：

$$\begin{aligned} F_{LB} : (H,W,C) &\to (H/P\times P, W/P\times P, C) \to \\ &\left(HW/P^2, P^2, C\right) \end{aligned} \tag{9}$$

1) 局部自注意力的相关度计算过程可以描述为：给定输入特征映射 $X \in R^{n\times h\times w\times c}$，首先将 X 通过 1*1 的卷积分别映射为 $Q,K,V \in R^{n\times h\times w\times c}$，然后通过公式 (10) 进行局部自注意力计算：

$$F_{LA}(Q,K,V) = softmax(QK^{\mathrm{T}}/\sqrt{d} + B)V \tag{10}$$

式中：Q、K、V 分别代表查询、键和值矩阵；B 表示相对位置偏移矩阵，d 表示特征的维度。在计算完窗口相关性后，通过窗口分割操作的反操作将窗口图像映射为与输入图像维度相同的输出图像。

2) 全局自注意力：全局自注意力是将输入特征划分为感受野更大的 $G\times G$ 网格，每个网格为计算相关性的一个单元，每个网格具有自适应大小的窗口空间，将自注意力集中在网格空间维度 $G\times G$ 上。与局部自注意力操作不同，全局注意力应用一个额外的转置将网格维度放置在假定的空间轴上，且计算相关性的窗口感受野更大，便于提取全局性的语义信息。全局自注意力窗口划分和注意力计算过程可以通过公式 (11) 和公式 (12) 进行描述。

$$\begin{aligned} G_{GB} : (H,W,C) &\to (H/G\times G, W/G\times G, C) \to \\ &\left(G^2, HW/G^2, C\right) \end{aligned} \tag{11}$$

$$G_{LA}(Q,K,V) = F_{LA}(Q,K,V) = softmax(QK^{\mathrm{T}}/\sqrt{d} + B)V \tag{12}$$

3) 局部全局注意力：通过自适应加权融合局部注意力和全局注意力，并通过残差学习的方式将注意力和输入特征进行融合，融合后的特征再经过层归一化处理和多层感知机特征映射，最后通过残差融合的方式输出。局部全局注意力特征提取过程可以表示为：

$$\begin{cases} F_{\text{lg}}(X) = \alpha(F_{unblock}(F_{LA}(F_{LB}(F_{LN}(X))))) + \\ \beta(F_{unblock}(F_{GA}(F_{GB}(F_{LN}(X))))) \\ X = X + F_{\text{lg}}(X) \\ X = X + F_{MLP}(F_{LN}(X)) \end{cases} \tag{13}$$

式中：F_{LN} 表示层归一化处理；F_{MLP} 是一个标准的 MLP 网络，它由两个线性变换层和一个非线性激活函数组成。

在图像复原估计中，平方损失和平均绝对误差损失是使用比较广泛的损失函数，但是平方惩罚会放大大小误差之间的差异，即它惩罚大误差但容忍小误差。这通常会导致平方损失约束下生成过度平滑的图像复原结果。因此，文中采用平均绝对误差损失作为损失函数，其计算公式如下：

$$L(\Theta)=\frac{1}{N}\sum_{i=1}^{N}(\|\widehat{I}_i-I_i\|_1) \tag{14}$$

式中：$\widehat{I}_i$和I_i表示预测的无高光图像和相应的真实值；N代表训练数据的总数。

2 实验结果分析

2.1 实验设置与评价指标

在 RD、SD1 和 SHIQ 三个基准数据集上进行了大量的实验验证文中方法在图像去高光任务上的有效性，三个数据库的训练集和测试集数据分布如表 1 所示。实验在单个 Nvidia RTX 3090 上进行训练和测试，训练期间采用 Adam 优化器对网络进行了优化学习，输入图像裁剪为 128 ×128 大小，初始学习率设置为 1e-3，批量大小设置为 60，训练总迭代次数设置为 200 次，训练过程中学习率在初始学习率基础上逐步递减进行动态调整直至网络收敛。

表 1 三个公开高光数据集上的数据分布情况

Tab.1 Data distribution of three public highlight datasets

Dataset	RD	SD1	SHIQ
Training dataset	1 800	12 000	9 825
Testing dataset	225	2 000	1 000

为了定量评估各算法在数据集上的性能，文中使用全参考指标 PSNR 和 SSIM 作为评估指标。PSNR 主要是从信噪比的角度评估预测图像与真实标签图像之间的差异。SSIM 主要从人类视觉系统感知上评估预测图像和真实标签图像之间的差异。一般来说，PSNR 和 SSIM 值越大，表明去除效果越好。此外，也从视觉上直观地评估了各对比方法的定性性能。

2.2 定性与定量实验结果对比分析

为了客观地评价算法在图像去高光上的有效性，文中在真实高光图像数据集上对算法做了定量和定性分析实验。与五种典型的图像去高光方法进行比较，五种方法包括：DCSR[35]、SHRBF[23]、NMF[36]、SCS[37]、SHRRI[18]。表 2 列出了上述方法在三个基准数据库上定量的性能指标，从表中可以看出，文中的方法在 PSNR 和 SSIM 指标上都优于其他方法，PSNR 值在数据集 SD1、RD 和 SHIQ 上高于次优方法的增量分别为 4.10 dB、7.09 dB 和 6.58 dB。SSIM 值在三个数据集上分别获得了 4%、9%和 3%的增量。

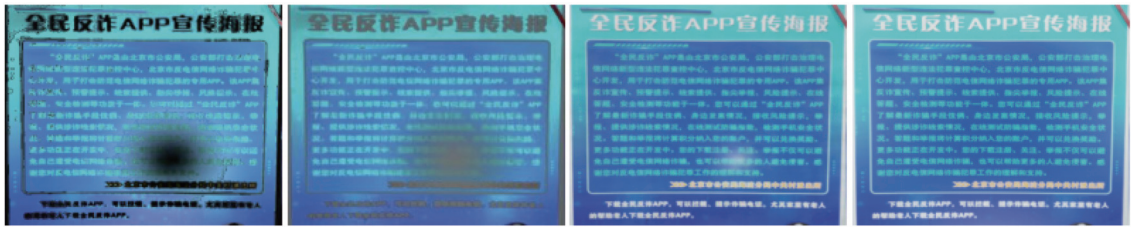
表 2 不同方法在三个公开数据集上的定量结果对比

Tab.2 Comparison of quantitative results of different methods on three public datasets

Dataset	SD1		RD		SHIQ	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DCSR	18.48	0.89	18.60	0.80	27.64	0.92
SHRBF	8.01	0.31	9.66	0.35	21.66	0.75
NMF	18.41	0.58	16.11	0.59	22.82	0.62
SCS	9.77	0.30	12.96	0.28	13.47	0.49
SHRRI	12.92	0.66	12.35	0.61	16.34	0.69
Ours	22.58	0.93	25.69	0.89	34.22	0.95

为了直观地评估各算法在高光数据集上的去高光结果，分别从三个公开高光数据集中随机选择了两个样本图像对各方法的去高光效果进行了对比。图 2~图 4 分别给出了各数据库上图像去高光直观的视觉结果。从图中可以看出，文中方法在真实高光图像上去高光效果较好，其他对比算法通常都无法准确有效地去除高光，且在高光处容易产生伪影和失真。Saint (DCSR) [35] 的方法容易有高光残留并在高光处无法恢

(a) Highlight (b) DCSR (c) SHRBF (d) NMF



(e) SCS (f) SHRRI (g) Proposed method (h) GT

图 2 各算法在 SD1 数据库上的视觉结果

Fig.2 The visual results of different methods on SD1 dataset

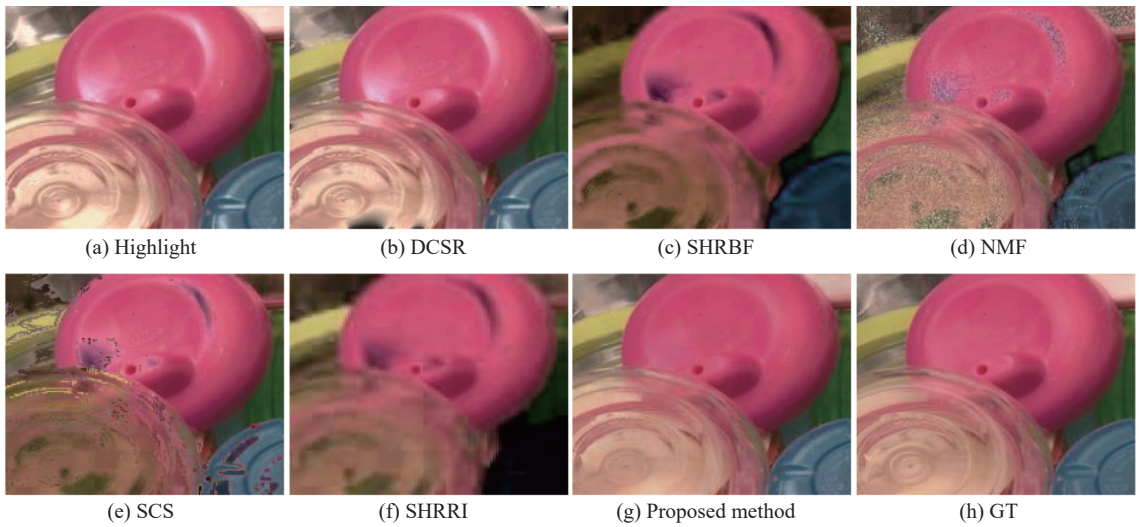
复背景信息；Yang (SHRBF) [23] 的方法使图像整体出现了严重的失真；Yamamoto (SCS) [37] 和 Akashi (NMF) [36] 的方法容易导致高光区域出现失真和严重的伪影，且将白色背景错误地视为高光去除；Fu

(a) Highlight (b) DCSR (c) SHRBF (d) NMF

(e) SCS (f) SHRRI (g) Proposed method (h) GT

图 3 各算法在 RD 数据库上的视觉结果

Fig.3 The visual results of different methods on RD dataset



(a) Highlight (b) DCSR (c) SHRBF (d) NMF

(e) SCS (f) SHRRI (g) Proposed method (h) GT

图 4 各算法在 SHIQ 数据库上的视觉结果

Fig.4 The visual results of different methods on SHIQ dataset

(SHRRI) [18] 的方法容易在高光区域出现失真，并使图像整体亮度变暗和出现模糊现象。

2.3 消融实验分析

为了证明网络中各模块的有效性，文中进一步对网络结构进行了消融实验研究，包括 UMAVTB 总体结构有效性验证和 MAVTB 结构有效性验证。

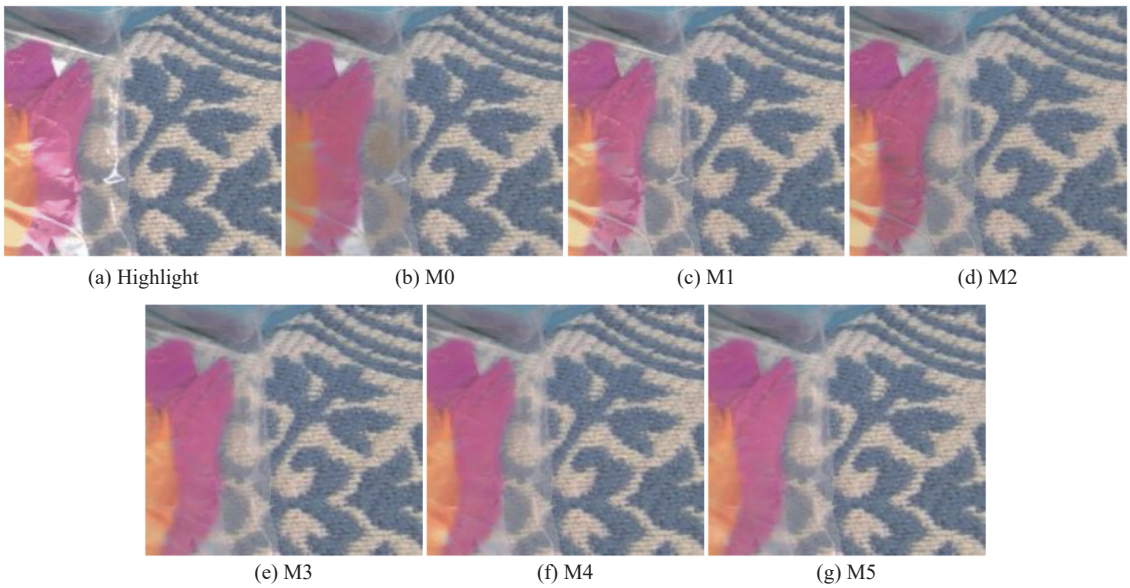
UMAVTB 总体结构有效性的验证。UMAVTB 网络包含两个关键的功能模块门限融合模块和并行多轴自注意力模块。为了证明这两个模块的有效性，文中在高光数据集 SHID 上进行了五个消融实验。首先建立了一个基准网络 M0，该网络由普通的卷积层串联构成，层数与 U 型网络相同。第二个基准网络为 M1，该基准网络由三阶段的普通卷积模块构成 U 型网络，但不含门限融合结构。然后，在基准网络 M1 基础上通过增加 U 型网络编码器和解码器之间的门限融合结构形成网络模型 M2。进一步地，通过将网络模型 M2 中的普通卷积模块替换为局部自注意力模块形成网络模型 M3，将网络模型 M2 中的普通卷积模块替换为全局自注意力模块形成网络模型 M4。为了检验多轴自注意力模块的有效性，通过替换 M3 中普通局部自注意力模块为多轴自注意力模块，形成更有效的去高光深度网络模型 M5。表 3 显示了消融实验的结果。通过对比 M0 和 M1 的结果可以看出，U 型网络结构性能更优越。基准网络 M1 使 PSNR 和 SSIM 值分别达到 30.78 dB 和 0.92。M2 在 M1 中增加了门限模块，将 PSNR 提高 0.68 dB，SSIM 提高 1%。模型 M5 将模型 M3 的标准局部自注意力模块改进为多轴自注意力模块。通过比较 M5 和 M3 的结果可以看出，多轴自注意力模块可以使 PSNR 平均值增加 0.55 dB，SSIM 提升 1%。通过比较 M5 和 M4 的结果可以看出，多轴自注意力模块可以使 PSNR 平均值增加 1.07 dB，SSIM 提升 1%。这两组对比实验性能的提升说明了多轴自注意力模块的有效性。

表 3 门限融合模块和并行多轴自注意力模块功效验证消融实验

Tab.3 Ablation experiments of the contribution of the threshold fusion structure and parallel multi-axis self-attention mechanism

Model	M0	M1	M2	M3	M4	M5
PSNR	30.15	30.78	31.46	33.67	33.15	34.22
SSIM	0.91	0.92	0.93	0.94	0.94	0.95

各消融模型的直观视觉对比如图 5 所示，可以看出，随着网络结构的逐步优化，去高光的效果在视觉上都得到了改善。基于纯卷积的深度网络模型 MI 和 M2 的输出结果有较多的高光残留，并在高光处



(a) Highlight (b) M0 (c) M1 (d) M2

(e) M3 (f) M4 (g) M5

图 5 消融模型去高光视觉结果对比

Fig.5 Comparison of visual results by ablation models for image highlight removal

会产生失真。基于卷积和自注意力融合的深度网络模型 M3、M4 和 M5 在视觉上都取得了较好的结果，但模型 M3 和 M4 在部分局部区域也存在少量的高光残留和失真。与此相比，文中采用的基于多轴自注意力模块的深度网络模型 M5 能够使网络可以同时学习全局和局部特征，使算法在准确去高光的同时保留了更多的背景细节信息，并且几乎无失真和伪影。

3 结 论

图像高光层和非高光层模型的模糊性导致了典型的基于先验信息的优化算法在去高光任务上的局限性，因此，文中利用深度学习强大的特征提取和非线性拟合能力，融合了卷积和自注意力机制在特征提取和表征上的优势，建立了基于复合型深度网络结构的图像去高光方法。通过门限融合 U 型网络将上下文信息与低层信息更好地融合提高像素级估计的准确性。通过并行多轴自注意力机制融合局部和全局稀疏型自注意力平衡了局部和全局特征的提取和解码。在真实高光数据集上的定量和定性实验结果表明，文中方能够获得较好的高光去除视觉效果，并在量化评价指标上优于其它主流的方法。虽然文中方法在公开的数据库上取得了可观的效果，但因这些数据库没有给出高光区域大小和高光强弱等信息，所以没有对这些因素与去高光性能之间的关系做定量的分析，这方面需要进一步优化。

参考文献：

[1] Zhang Z D, Xue Z Y, Chen Y, et al. Boosting verified training for robust image classifications via abstraction[C]//Proc of the IEEE Conference on Computer Vision & Pattern Recognition, 2023: 16251-16260.

[2] Muhammad F N, Muhammad G Z A K, Xian Y Q, et al. I2mvformer: large language model generated multi-view document supervision for zero-shot image classification[C]// Proc of the IEEE Conference on Computer Vision & Pattern Recognition, 2023: 15169-15179.

[3] Zhu Y, Tang J, Li S, et al. Derendernet: intrinsic image decomposition of urban scenes with shape-(in)dependent shading rendering[C]//Proc of the 2021 IEEE International Conference on Computational Photography (ICCP), 2021: 1-11.

[4] Zhang F, Jiang X, Xia Z, et al. Non-local color compensation network for intrinsic image decomposition [J]. ***IEEE Transactions on Circuits and Systems for Video Technology***, 2023, 33(1): 132-145.

[5] Minaee S, Boykov Y, Porikli F, et al. Image segmentation using deep learning: a survey [J]. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 2020, 44(7): 3523-3542.

[6] Clough J R, Byrne N, Oksuz I, et al. A topological loss function for deep-learning based image segmentation using persistent homology [J]. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 2023, 44(12): 8766-8778.

[7] Wang Dongdong, Zhang Wei, Jin Guofeng, et al. Application of cusp catastrophic theory in image segmentation of infrared thermal waving inspection [J]. ***Infrared and Laser Engineering***, 2014, 43(3): 1009-1015. (in Chinese)

[8] Zou Z, Shi Z, Guo Y, et al. Object detection in 20 years: a survey[J]. Proceedings of the IEEE, 2023, 111(3): 257-276.

[9] Li X, Lv C, Wang W, et al. Generalized focal loss: towards efficient representation learning for dense object detection [J]. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 2023, 45(3): 3139-3153.

[10] Kong Y, Fu Y. Human action recognition and prediction: a survey [J]. ***International Journal of Computer Vision***, 2022, 130(5): 1366-1401.

[11] Sun Z, Ke Q, Rahmani H, et al. Human action recognition from various data modalities: A review [J]. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 2023, 45(3): 3200-3225.

[12] Seidenschwarz J, Brasó G, Elezi I, et al. Simple cues lead to a strong multi-object tracker[C]//Proc of the IEEE Conference on Computer Vision and Pattern Recognition, 2023: 13813-13823.

[13] Hu W, Wang Q, Zhang L, et al. Siammask: a framework for fast online object tracking and segmentation [J]. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 2023, 45(3): 3072-3089.

[14] Chen Faling, Ding Qinghai, Luo Haibo, et al. Anti-occlusion real time target tracking algorithm employing spatio-temporal context [J]. ***Infrared and Laser Engineering***, 2021, 50(1): 20200105. (in Chinese)

[15] Li Bo, Zhang Xinyu. Target tracking algorithm based on adaptive feature fusion in complex scenes [J]. ***Infrared and Laser Engineering***, 2022, 51(10): 20220013. (in Chinese)

[16] Shafer S A. Using color to separate reflection components [J]. ***Color Research & Application***, 1985, 10(4): 210-218.

[17] Yang Q, Tang J, Ahuja N. Efficient and robust specular highlight removal [J]. ***IEEE Transactions on Pattern Analysis***

and Machine Intelligence, 2015, 37(6): 1304-1311.

[18] Fu G, Zhang Q, Song C, et al. Specular highlight removal for real-world images [J]. *Computer Graphics Forum*, 2019, 38(7): 253-263.

[19] Kim H, Jin H, Hadap S, et al. Specular reflection separation using dark channel prior[C]//Proc of the IEEE Conference on Computer Vision & Pattern Recognition, 2013: 1460-1467.

[20] Tan R T, Ikeuchi K. Separating reflection components of textured surfaces using a single image [J]. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 2005, 27(2): 178-193.

[21] Liu Y, Yuan Z, Zheng N, et al. Saturation-preserving specular reflection separation[C]//Proc of the IEEE Conference on Computer Vision & Pattern Recognition, 2015: 3725-3733.

[22] Suo J, An D, Ji X, et al. Fast and high quality highlight removal from a single image [J]. *IEEE Transactions on Image Processing*, 2016, 25(11): 5441-5454.

[23] Yang Q X, Wang S N, Ahuja N. Real-time specular highlight removal using bilateral filtering[C]//Proc of the 11th European Conference on Computer Vision, 2010: 87-100.

[24] Fu G, Zhang Q, Lin Q, et al. Learning to detect specular highlights from real-world images[C]//Proc of the ACM International Conference on Multimedia, 2020: 1873-1881.

[25] Shi J, Dong Y, Su H, et al. Learning non-lambertian object intrinsics across shapenet categories[C]//Proc of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1685-1694.

[26] Yi R, Tan P, Lin S. Leveraging multi-view image sets for unsupervised intrinsic image decomposition and highlight separation[C]//Proc of the AAAI Conference on Artificial Intelligence, 2020: 12685-12692.

[27] Huang Z, Hu K, Wang X. M2-NET: multi-stages specular highlight detection and removal in multi-scenes [DB/OL]. (2022-07-20)[2024-02-20].https://arxiv.dosf.top/abs/2207.09965.

[28] Hou S, Wang C, Quan W, et al. Text-aware single image specular highlight removal[C]//Proc of the 4th Chinese Conference on Pattern Recognition and Computer Vision (PRCV), 2021: 115-127.

[29] Jimenez-Martin L, Perez D A V, Asteasuainzarra A S M, et al. Specular reflections removal in colposcopic images based on neural networks: Supervised training with no ground truth previous knowledge [DB/OL]. (2020-06-21) [2024-02-20].https://doi.org/10.48550/arXiv.2106.02221.

[30] Fu G, Zhang Q, Zhu L, et al. A multi-task network for joint specular highlight detection and removal[C]//Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 7748-7757.

[31] Li K, Wang Y, Zhang J, et al. Uniformer: unifying convolution and self-attention for visual recognition [J]. ***IEEE Transactions on Pattern Analysis & Machine Intelligence***, 2023(1): 1-18.

[32] Gulati A, Qin J, Chiu C C, et al. Conformer: convolutionaugmented transformer for speech recognition [DB/OL].(2020-05-16)[2024-02-20].https://doi.org/10.48550/arXiv.2005.08100.

[33] Jiang Z H, Yu W, Zhou D, et al. Convbert: improving bert with span-based dynamic convolution [J]. ***Advances in Neural Information Processing Systems***, 2020, 33: 12837-12848.

[34] Wu H, Xiao B, Codella N, et al. Cvt: introducing convolutions to vision transformers[C]//Proc of the IEEE/CVF International Conference on Computer Vision, 2021: 22-31.

[35] Saint-Pierre C A, Boisvert J, Grimard G, et al. Detection and correction of specular reflections for automatic surgical tool segmentation in thoracoscopic images [J]. ***Machine Vision & Applications***, 2011, 22(1): 171-180.

[36] Akashi Y, Okatani T. Separation of reflection components by sparse non-negative matrix factorization[C]//Proc of the Asian Conference on Computer Vision, 2015: 611-625.

[37] Yamamoto T, Nakazawa A. General improvement method of specular component separation using high-emphasis filter and similarity function [J]. *ITE Transactions on Media Technology and Applications*, 2019, 7(2): 92-102.

Image highlight removal method based on parallel multi-axis self-attention

Li Pengyue[1,2*], Xu Xinying[1], Tang Yandong[3,4], Zhang Zhaoxia[1], Han Xiaoxia[1], Yue Haifeng[2]

(1. College of Electrical and Power Engineering, Taiyuan University of Technology, Taiyuan 030024, China;
2. Taiyuan Heavy Machinery (Group) Company, Taiyuan 030027, China;
3. State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China;
4. Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110016, China)

Abstract:

Objective Highlights are manifested as high bright spots on the surface of glossy materials under the action of light. The highlights of the image can obscure background information with different degrees. The ambiguity of the image highlight layer model and the large dynamic range of highlights enable highlight removal to be still a challenging visual task. The purely local methods tend to result in artifacts in the highlight areas of the image, and the purely global methods tend to produce color distortion in highlight-free areas of the image. To address the issues caused by the imbalance of local and global features in image highlight removal and the ambiguity of highlight layer modeling, we propose a threshold fusion U-shaped deep network based on parallel multi-axis self-attention mechanism for image highlight removal.

Methods Our method avoids the ambiguity of highlight layer modeling by implicit modeling. It uses the U-shaped network structure to combine the contextual information with the low-level information to estimate the highlight-free image, and introduces a threshold fusion structure between the encoder and decoder of the U-shape structure to further enhance the feature representation capability of the network. The U-shaped network uses the contraction convolution strategy to extract the contextual semantic information faster. It gradually recovers the low-layer information of the image by expanding, and connects the features of the various stages of the contraction path in the corresponding stages of the expansion path. The threshold mechanism between the encoder and decoder is used to adjust the information flow in each channel of the encoder, which allows the encoder to extract features related to highlights as much as possible at channel level. The threshold structure first performs high- and low-frequency decoupling and feature extraction for the input features, then fuses the two types of features by pixel-wise multiplication, and finally uses the residual pattern to learn the low-level features complementary. In addition, the parallel multi-axis self-attention mechanism is used as the unit structure of the U-shaped network to balance the learning of local and global features, which eliminates the distortion and artifacts of the recovered highlight-free images caused by the imbalance extraction of local and global features. The local self-attention calculates local interactions within a small $P*P$ window to form local attention. After the correlation calculation of the small window, the window image is mapped to an output image with the same dimension as the input image by the inverse operation of the window segmentation operation. Similarly, the global self-attention divides the input features into $G*G$ grids with larger receptive fields. Each grid is a cell for calculating correlation, which has an adaptive size of the window space. The larger receptive field window of calculating correlation facilitates the extraction of global semantic information. For the loss function, the squared loss and the mean absolute error loss are the widely used loss functions in the image restoration field. The squared penalty magnifies the difference between large and small errors. It usually results in excessively smooth restored images.

Therefore, the mean absolute error loss is used as the loss function to train our network.

Results and Discussions Qualitative experiments on real highlight images show that our method can remove highlights from images more effectively, and other compared methods usually cannot remove highlights accurately and efficiently. They are prone to produce artifacts and distortion in highlight-free areas of the image. Quantitative experiments on real-world highlight image datasets show that our method outperforms five other typical image highlight removal methods in both PSNR and SSIM metrics. The PSNR values are higher than those of the second-best method by 4.10 dB, 7.09 dB, and 6.58 dB on the datasets of SD1, RD, and SHIQ, respectively. The SSIM values of our method also outperform those of the second-best method with gains of 4%, 9%, and 3% on three datasets. In addition, we also conduct ablation studies for the network structure, and the experiment verifies the effectiveness of the threshold fusion module and the parallel multi-axis self-attention module; The threshold fusion module can increase the PSNR by 0.68 dB and the SSIM by 1%, and the multi-axis self-attention module can increase the average PSNR value by 0.55 dB and the SSIM by 1%. It can also be seen from the visual results of each ablation experimental model that with the gradual optimization of the network structure, the results of image highlight removal are visually improved. The outputs of the pure convolution-based deep network models of MI and M2 have more highlight residuals and produce distortion in the highlight-free areas of the image. The models of M3, M4 and M5 combining CNN with the self-attention module visually achieve better results.

Conclusions The experimental results show that good visual results for highlight removal on both public natural and textual image datasets are achieved with our method, which outperforms other methods in terms of quantitative evaluation metrics.

Key words: image processing; highlight removal; multi-axis self-attention; deep learning

Funding projects: National Natural Science Foundation of China (62203319); Natural Science Foundation of Shanxi Province (202203021212220, 202103021224056); Shanxi Province Science and Technology Cooperation and Exchange Program (202104041101030)