基于改进 Vision Transformer 的局部光照一致性估计

王杨[1]，宋世佳[1]，王鹤琴[1]，袁振羽[1]，赵立军[2]，吴其林[1]
（1.安徽师范大学计算机与信息学院，安徽 芜湖 241000；2.长三角哈特机器人产业技术研究院，安徽 芜湖 241000）

摘　要：光照一致性是增强现实（AR）系统中实现虚实有机融合的关键因素之一。由于拍摄视角的局限性和场景光照的复杂性，开发者在估计全景照明信息时通常忽略局部光照一致性，从而影响最终的渲染效果。为解决这一问题，提出一种基于改进视觉 Transformer（ViT）结构的局部光照一致性估计框架（ViTLight）。首先利用 ViT 编码器提取特征向量并计算回归球面谐波（SH）系数，进而恢复光照信息；其次改进 ViT 编码器结构，引入多头自注意力交互机制，采用卷积运算引导注意力头之间相互联系，在此基础上增加局部感知模块，扫描每个图像分块并对局部像素进行加权求和，捕捉区域内的特定特征，有助于平衡全局上下文特征和局部光照信息，提高光照估计的精度。在公开数据集上对比主流特征提取网络和 4 种经典光照估计框架，实验和分析结果表明，ViTLight 在图像渲染准确率方面高于现有框架，其均方根误差（RMSE）和结构相异性（DSSIM）指标分别为 0.129 6 和 0.042 6，验证了该框架的有效性与正确性。

关键词：增强现实；光照估计；球面谐波系数；视觉 Transformer；多头自注意力
中图分类号：TP391.41　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0068905

Estimation of Local Illumination Consistency Based on Improved Vision Transformer

WANG Yang[1], SONG Shijia[1], WANG Heqin[1], YUAN Zhenyu[1], ZHAO Lijun[2], WU Qilin[1]
(1. School of Computer and Information, Anhui Normal University, Wuhu 241000, Anhui, China;
2. Yangtze River Delta Region Hart Robotics Industry Technology Research Institute, Wuhu 241000, Anhui, China)

【Abstract】 Illumination consistency is a key factor in achieving the organic fusion of virtual and real elements in Augmented Reality (AR) systems. Owing to the constraints of capture perspectives and the complexity of scene illumination, developers often overlook local illumination consistency when estimating panoramic lighting information, thereby affecting the final rendering quality. To address this issue, this study proposes a local illumination consistency estimation framework, ViTLight, based on an improved Vision Transformer (ViT) structure. First, the framework uses a ViT encoder to extract feature vectors and calculate regression Spherical Harmonic (SH) coefficients, then recovers illumination information. Second, the ViT encoder structure is enhanced by introducing a multi-head self-attention interaction mechanism. Convolution operation guides the interplay between attention heads. Additionally, a local perception module is integrated to actively scan each image block and perform weighted summation on local pixels to capture specific features within regions. This proactive approach balances global contextual features and local illumination information, ultimately improving the precision of illumination estimation. The mainstream feature extraction network and four classical illumination estimation frameworks are compared on public datasets. The experimental results and analysis indicate that ViTLight is superior to existing frameworks in terms of image rendering accuracy, and its Root Mean Square Error (RMSE) and Structural Dissimilarity (DSSIM) index reach 0.129 6 and 0.042 6, respectively, which verifies its effectiveness and correctness.

【Key words】 Augmented Reality (AR); illumination estimation; Spherical Harmonics (SH) coefficient; Vision Transformer (ViT); multi-head self-attention

0　引言

增强现实（AR）是一种在真实环境中叠加虚拟内容的技术，通过增强和补充真实环境的信息，实现虚拟物体与真实环境的融合[1]。AR 技术已被广泛应用在工业、医疗、教育和军事等领域[2-3]。其中，几何一致性和时间一致性这 2 个关键因素的研究工作已取得显著进展[4]，而光照一致性问题仍存在很多

收稿日期：2023-11-27　　**修回日期：**2024-02-20
基金项目：国家自然科学基金（61871412）；安徽省自然科学基金重点项目（KJ2019A0938，KJ2021A1314，KJ2019A0979）；安徽高校自然科学重点项目（2022AH052899，KJ2019A0979，KJ2019A0511，2023AH052757）；机器视觉检测安徽省重点实验室开放课题（KLMVI-2023-HIT-11）；安徽省高校学科（专业）拔尖人才学术项目（gxbjZD2022147）。
通信作者 E-mail：wycap@126.com

挑战，是增强现实系统面临的主要难题之一。

由于相机视角的局限性，从单一图像中估计完整光照信息是一个颇具挑战的目标任务[5]。近年来，卷积神经网络(CNN)被广泛用于计算机视觉任务[6-9]。然而，从单一图像中估计空间位置变化的场景光照通常依赖已知额外线索的输入，计算成本高且照度估计精度低。目前，光照估计问题常常采用分解子任务方式解决[8]。但由于合成数据和真实数据之间的材质差异，真实数据上的内在分解结果不如合成数据上的结果准确。因此，从单一有限视场(LFOV)图像中恢复场景照明仍具有挑战性。

为了实现虚拟物体与真实环境之间的光照一致性，增强光照估计框架的局部信息感知能力同时提升模型鲁棒性，本文提出了一种新的光照估计框架 ViTLight(Vision Transformer Light)，主要工作如下：

1)采用 ViT 编码器解决光照估计任务；引入多头自注意力交互机制(MSAI)实现注意力头之间的信息联系，以获取更丰富的场景光照细节。

2)提出一种局部感知卷积模块(LPConv)，该模块能够捕获相邻像素的信息相关性，强化局部信息提取能力，使所提框架在全局和局部光照细节提取方面实现平衡。

1 相关工作

现有的光照估计方法主要分为两类，一类是基于生成环境图直接恢复光照，另一类是基于回归光照参数还原光照。基于生成全景图的方法通常将光照估计分为从有限视场的图像到全景图的补全、低动态范围(LDR)全景图生成高动态范围(HDR)全景图等多个子阶段[10-11]。文献[12]训练了一个照明分类器来自动标注 LDR 图中的光源位置，再使用训练后的神经网络模型微调光照强度，最后预测整体光照分布。文献[13]基于像素聚类先对图像进行反射光分解，得到漫反射图和镜面反射图后，进一步分解本征图像得到反照率图和阴影图，最后结合分解结果和场景深度对输入图像的光照信息进行计算。文献[14]基于生成对抗网络，在移动设备上实时估计 HDR 环境地图。文献[15]通过基于物理的天空模型，从一般场景的单个室外 LDR 图中自动预测 HDR 全景照明，HDR 全景环境地图被广泛用于照亮虚拟物体，但在实际应用中难以捕获。DLNet[16]基于球形多尺度动态卷积的 CNN，用于解码球形域的特征，以预测全景环境地图。文献[17]综合考虑光源类型和照明效果的对应关系，设计深度学习模型解析复杂 LDR 全景图，通过光源检测、分类和相机逆成像机制实现照度估计。从 HDR 全景图获取光照信息的方法通常对设备算力和渲染效率要求较高，实用性并不高。

基于回归光照参数的方法通常使用 LDR LFOV 图像作为输入，将场景中的有效光照信息表示为光照回归参数[18]，如光照方向和强度、球面谐波(SH)系数[19-20]和小波变换等。文献[21]利用视频序列来改善照明预测，在室内场景的任意位置进行时空一致的光照估计。文献[22]设计深度 CNN，将场景信息分为几个组件，分别为照明、法线和双向反射分布函数，引入可微屏幕空间渲染，使用 SH 系数和主要方向照明恢复真实光照。然而，当一个场景中包含镜片、金属等反射材料时，回归光照参数方法对场景中高频信息难以获取，因此不能提供较好的渲染效果，神经网络的训练往往不稳定。为此，文献[23]提出了镜面和透明物体的光照参数以及材质的联合全局优化方法，模拟焦散和估计折射率以逼真插入虚拟物体。文献[24]提出一种端到端的方法来预测用高斯函数表示的室内照明，以保留更多的光照细节。文献[25]提出 EMLight 光照估计框架，将环境图分解为球面光分布、光强和环境项以进行照度回归。

与先前研究相比，本文的光照估计框架不依赖任何场景几何、材质属性等先验知识，而是从全局和局部信息融合的角度，利用改进 ViT 编码器提取场景光照信息，并通过估计 SH 系数来实现光照估计。

2 问题描述与系统模型

2.1 光照估计问题与方案描述

由于拍摄视角受限和场景光照的复杂性，用户无法观察到视野外的照明，因此完整提取环境光照信息较为困难。尽管 ViT 在捕获输入内容远程依赖关系方面表现出色，并在各种视觉任务中取得了成功，但与现有 CNN 相比性能仍然不高，主要原因是 ViT 缺乏局部信息提取能力，CNN 固有的局部归纳偏置导致其更倾向于提取全局信息，而对局部特征的感知能力相对较弱[26-28]。

针对标准 ViT 提取全局光照信息时易丢失局部照明细节的问题，本文构建了改进型 ViT 光照估计框架 ViTLight。ViT 是一种基于自注意力机制的视觉模型，受 ResT 网络多头自关注算法的启发，在保持多头多样性的同时，将交互作用投射到注意力头维度上，设计多头自注意力交互机制指导不同注意力头的行为[29]。本文通过引入编码器局部感知卷积模块和多头自注意力交互机制这 2 个关键组

件，实现对光照估计框架的改进，促进光照估计框架对全局和局部信息的共同关注。

2.2 ViTLight 结构

ViTLight 光照估计框架结构如图 1 所示（彩色效果见《计算机工程》官网 HTML 版，下同），主要分为特征提取网络和球面谐波系数计算 2 个部分。基于改进 ViT 的光照估计算法描述如算法 1 所示。

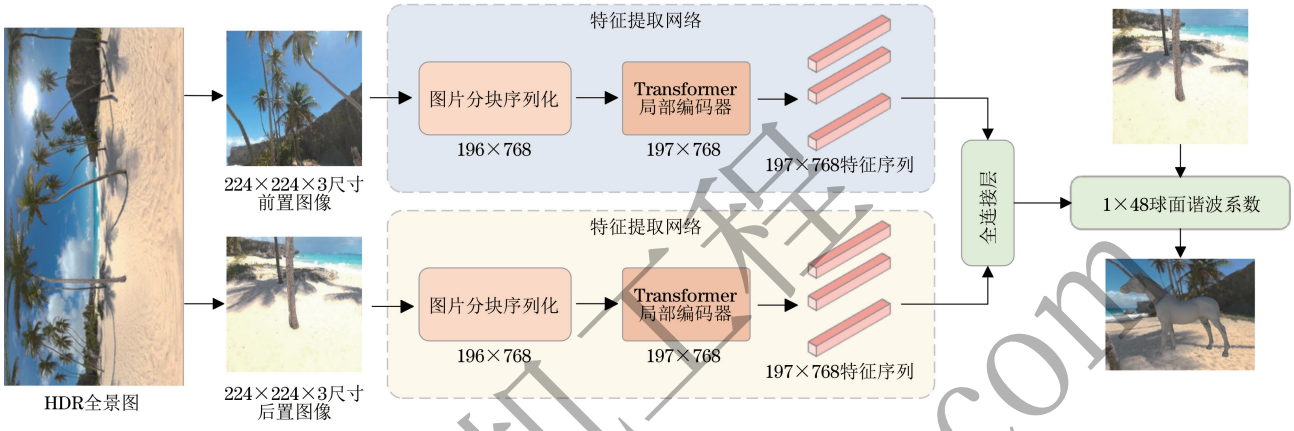


图 1 ViTLight 光照估计框架

Fig. 1 ViTLight illumination estimation framework

算法 1 ViTLight 算法

输入 移动设备前后拍摄的图像 input_image

输出 SH 光照参数

1. for epoch = 1,…,100 do//训练 100 轮
2. for x =1,…,total_image do//遍历输入图像
3. H,W,C←224,224,3;//适应 ViT 的输入分辨率
4. X_p^i←[X_p^1,X_p^2,…,X_p^N]←patch_embed (x); //图//片序列化
5. X_p^i sent to ViTLight Encoder;//特征序列至编//码器
6. update MSAI(Q,K,V) using Eq. (7) and Eq. (8);
7. Z_m←conv_block (Z_{m-1});
8. end for
9. fusion([feature_front, feature_rear]); //融合图//像特征
10. sh ← net. output(); //输出 SH 系数
11. loss ← sh_loss + render_loss using Eq. (13);
12. optimizer. step();//更新参数
13. end for

具体来说，该网络的输入是从 HDR 图像中提取的成对前后图像，同时调整输入图片大小为 224×224×3，送入改进后的 ViT 特征提取网络。首先，为了将输入图像转换为可供 ViT 使用的特征序列，利用图片分块序列化模块将输入图片分割成固定尺寸的图像块，再经过尺寸变换产生 196 个维度为 768 的视觉特征序列，代表输入图像的空间信息；之后，为了防止输入序列的顺序信息丢失，在视觉特征序列后面添加一个可学习的类别词符（class token），维度同为 768，同时对每个图像块进行位置编码（PE），以保留输入序列的位置信息；接着，将 197 个维度为 768 的特征向量送入编码器进行特征交互，得到更加语义化的特征表示。经过全连接层处理，得到表示光照的 SH 系数向量，再通过最小化 SH 损失和渲染损失进一步优化光照估计框架参数。

2.3 场景图像分块

机器视觉任务中的输入图像尺寸通常表示为 $X \in \mathbb{R}^{H\times W\times C}$，其中，$H$ 表示图像高度，W 表示图像宽度，C 表示图像通道数，本文输入图像尺寸为 224×224×3。图片分块序列化首先将图片切分为尺寸为 $P\times P\times C$ 的图像块，其中 P 为图片分块大小，本文取 $P=16$；接着展开每个块，得到一个一维向量，最终图片表示为 $X_p \in \mathbb{R}^{N\times(P^2\times C)}$，$N$ 为图像块的数量，即一张 224×224×3 的图片可切分成 196 个尺寸为 16×16×3 的图像块。图像分块后的特征序列称为视觉特征序列，线性映射生成的视觉特征序列 X_p^i 如式(1)所示：

$$X_p^i=[X_p^1,X_p^2,\cdots,X_p^N] \tag{1}$$

式中：i 表示当前特征序列维度。为保证各序列的顺序和位置信息，将序列与类别词符组合成向量 $\boldsymbol{Z}_0$，具体计算如式(2)所示：

$$\boldsymbol{Z}_0=[X_{\text{cls}},X_p^1\boldsymbol{E},X_p^2\boldsymbol{E},\cdots,X_p^N\boldsymbol{E}]+P_{\text{PE,local}} \tag{2}$$

式中：X_{cls} 是为了实现光照场景分类而在网络中加入的可学习的类别词符，其维度与视觉特征序列单个向量维度保持一致；$\boldsymbol{E}$ 是实现线性映射的矩阵；$P_{\text{PE,local}}$ 是位置编码。由于自注意力计算过程中忽略了位置信息，而图片分块之间的关系与位置信息有关，因此可以将图片分块序列化后的特征向量与一个

额外位置向量相加，来表示该图像小块在整个图像序列中的位置，最终得到 197 个维度为 768 的特征向量作为编码器的最终输入，计算过程如式(3)所示：

$$\begin{cases} P_{\mathrm{PE},(\mathrm{local},2i)}=\sin\left(\dfrac{L_{\mathrm{local}}}{10\ 000^{\frac{2i}{d}}}\right) \\ P_{\mathrm{PE},(\mathrm{local},2i+1)}=\cos\left(\dfrac{L_{\mathrm{local}}}{10\ 000^{\frac{2i}{d}}}\right) \end{cases} \tag{3}$$

式中：L_{local} 表示每个元素在序列中的位置；d 表示线性映射后的维度；i 表示位置编码的当前维度，i 的取值范围是 $\left[1,2,\cdots,\dfrac{d}{2}\right]$；sin 和 cos 分别代表正弦三角函数和余弦三角函数。

2.4 全局-局部特征平衡编码器

改进后的 ViT 模型结构如图 2 所示。模型将分块序列化、拼接类别词符、添加位置编码处理后的输入图片送至 Transformer 局部编码器。ViT 的特征提取模块由 M 个相同的特征编码器叠加而成，本文 M 取 12。ViT 的单个特征提取模块是由多头自注意力交互机制、局部感知卷积模块和多层感知模块(MLP)三部分组成。首先，输入图像经过层归一化(LN)后，经过本文设计的 MSAI，将输出结果与输入特征图进行残差连接；然后，将输出结果经过 LN 并进行局部卷积操作，送入 MLP 提取特征，同样适用残差连接，并且在图像的特征提取过程中，仍然是 197 个维度为 768 的视觉特征序列。第 m 个编码器的计算过程是输入视觉特征序列 Z_{m-1} 经过编码层得到 Z_m，计算过程如式(4)所示：

$$\begin{aligned} Z'_m&=\mathrm{MSAI}(\mathrm{LN}(Z_{m-1}))+Z_{m-1},\ m=1,\cdots,M \\ Z_m&=\mathrm{MLP}(\mathrm{LN}(Z'_m))+Z'_m,\ m=1,\cdots,M \end{aligned} \tag{4}$$

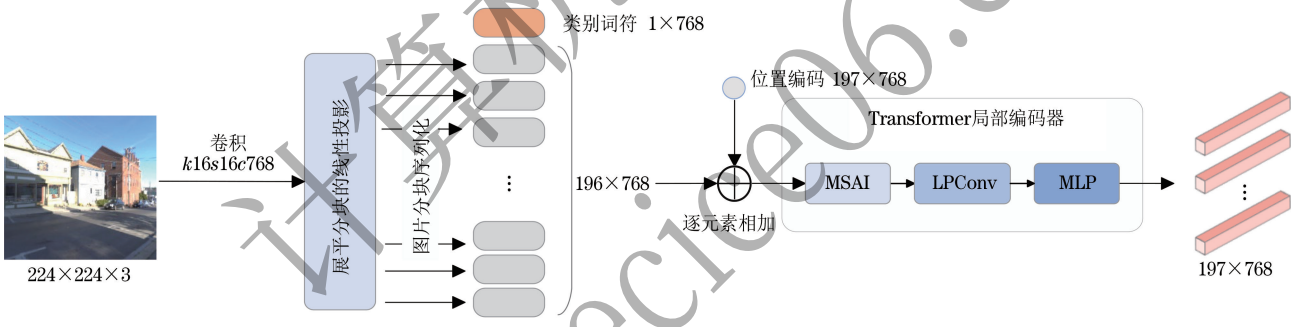


图 2 改进 ViT 的模型结构

Fig.2 Structure of improved ViT model

2.4.1 多头自注意力交互模块

为解决传统 CNN 中卷积核感受野受限的问题，自注意力机制被应用于建立图像空间上的长距离依赖关系。但多头自注意力中每个自注意力头仅计算输入的部分子集，缺乏多头之间的特征交互功能，导致头之间的信息不足，无法充分利用输入的信息。受 ResT 架构的启发[29]，本文设计了多头自注意力交互模块 MSAI 来解决标准 ViT 多头相互独立的问题。如图 3 所示，在保持多头能力多样性的同时，通过在多注意力头维度上投射交互，来弥补每个头的输入标记数量的限制，避免每个头只能观察有限信息；通过计算特征序列中每个向量之间的相互作用，来寻找图像块之间的关系。ViTLight 使用缩放点积注意力，用 $Z\in\mathbb{R}^{(N+1)\times d}$ 表示 $N+1$ 个特征序列，其中，d 表示每个序列的维度。自注意力机制包含 3 个可学习的权重矩阵，即查询矩阵 $\boldsymbol{W}_q$、键矩阵 $\boldsymbol{W}_k$ 以及值矩阵 $\boldsymbol{W}_v$。首先，将输入特征序列与 3 个权重矩阵做乘法得到 3 个不同的特征向量，分别为查询向量 $\boldsymbol{Q}$、键向量 $\boldsymbol{K}$ 以及值向量 $\boldsymbol{V}$，用于计算不同位置之间的相关性并生成最终的注意力权重，计算过程如式(5)所示：

$$\boldsymbol{Q}=Z\boldsymbol{W}_q,\ \boldsymbol{K}=Z\boldsymbol{W}_k,\ \boldsymbol{V}=Z\boldsymbol{W}_v \tag{5}$$

接着，计算每对向量的点积分数 $\boldsymbol{Q}\cdot\boldsymbol{K}^{\mathrm{T}}$，再进行归一化处理使梯度保持稳定，并将结果通过

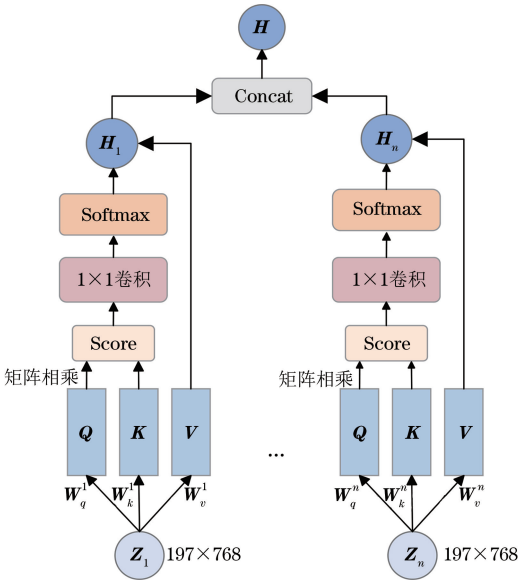


图 3 多头自注意力交互机制

Fig.3 Multi-head self-attention interaction mechanism

Softmax 函数进行映射。之后，与向量 $\boldsymbol{V}$ 做矩阵乘法，得到加权后每个输入向量的得分[27]。具体的标准多头自注意力计算过程如式(6)所示：

$$A_{\text{Atten}}=\text{Softmax}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\times\boldsymbol{V} \tag{6}$$

式中：$\boldsymbol{Q}\cdot\boldsymbol{K}^{\text{T}}$ 为注意力分数；d_k 为向量 $\boldsymbol{Q}$、$\boldsymbol{K}$ 的维度；$\sqrt{d_k}$ 是比例因子。改进后多头自注意力交互计算如式(7)所示：

$$A'_{\text{Atten}}=\text{Softmax}\left(\text{Conv}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\right)\times\boldsymbol{V} \tag{7}$$

式中：Conv(•)表示卷积核为 1×1 的计算操作，模拟不同头部之间的相互作用。

每个独立的自注意力头都有各自的权重矩阵 $\boldsymbol{W}_q^i$、$\boldsymbol{W}_k^i$、$\boldsymbol{W}_v^i$，多头自注意力的计算过程如式(8)所示：

$$\begin{aligned}&\boldsymbol{h}_{\text{head}_i}=A'_{\text{Atten}}(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i)\\&\text{MSAI}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(\boldsymbol{h}_{\text{head}_0},\cdots,\boldsymbol{h}_{\text{head}_n})\boldsymbol{\theta}\end{aligned} \tag{8}$$

式中：$\boldsymbol{h}_{\text{head}_i}$ 表示每个注意力头的输出向量；$\boldsymbol{\theta}$ 表示多头输出的权重矩阵。

与标准多头自注意力机制不同，本文在向量 $\boldsymbol{Q}$、$\boldsymbol{K}$ 进行点积后执行一次卷积操作来增加非线性度，从多个角度提取特征后，将多头信息进行聚合以得到更全面的特征。

2.4.2 局部感知模块

ViT 编码器中的多层感知机由输入层、输出层和至少一层的隐藏层构成。通过全局自注意力机制，ViTLight 光照估计框架能够捕捉到图像中的整体信息。然而，该机制对于纹理、边缘等一些局部信息的捕捉效果并不理想。为了解决 ViTLight 在局部归纳偏置方面的不足，本文引入局部特征信息来增强模型对局部特征的感知能力。在特征编码器的 MLP 部分之前引入局部感知卷积(LPConv)，使用小的卷积核(kernel 为 3×3，padding 为 1)，在每个位置使用局部感受野，以确保模型对局部相关性的建模。ViTLight 局部特征编码器结构如图 4 所示。

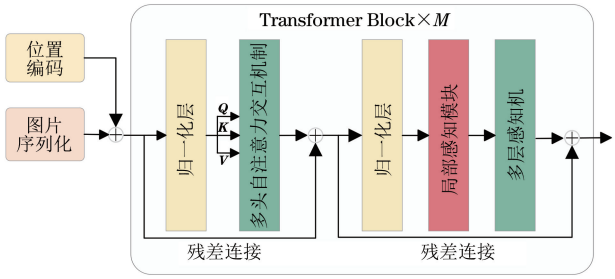


图 4 局部特征编码器结构

Fig.4 Structure of local feature encoder

2.5 基于球面谐波的照明表示

为使得提取特征中包含更多的高频信息，选用由 48 个系数组成的 SH 系数。环境光照的计算如式(9)所示：

$$L(p,w_0)=\int L(p,w_i)nw_i\,\text{d}w_i \tag{9}$$

该公式描述了空间上某个着色点 p 在整个球面空间中收到的光照总和，其中，w_i 为入射光方向，w_0 为观察方向，n 为着色点 p 的法线。使用 SH 参数编码关照信息的最大好处是节省了空间和时间，适合深度学习[30]。使用 SH 时，照明分布如式(10)所示：

$$L(\theta)=\sum_{i=1}^{M}S_{\text{SH},i}\times y_i(\theta) \tag{10}$$

式中：θ 是光照方向；$L(\theta)$ 是光照分布函数；M 是 SH 系数的阶数；y_i 是第 i 个 SH 函数的基函数；$S_{\text{SH},i}$ 是第 i 个 SH 函数。

2.6 光照损失函数

本文采用均方误差(MSE)来计算损失函数以优化 ViTLight。计算各阶 SH 系数的平均 MSE 损失，如式(11)所示：

$$L_1=\frac{1}{n}\sum_{a=0}^{n-1}\frac{1}{2a+1}\left(\sum_{m=-a}^{a}(S_{\text{SH},a}^m-\tilde{S}_{\text{SH},a}^m)^2\right) \tag{11}$$

式中：n 是 SH 系数的展开阶数；a 是当前阶数；$S_{\text{SH},a}^m$ 表示当前第 a 阶的第 m 项 SH 系数。

由于本文的最终目标是渲染出与现实光照相一致的虚拟物体，因此在考虑 SH 系数这一中间产物的同时还需要考虑渲染损失，其计算过程如式(12)所示：

$$L_2=\frac{1}{W\times H\times C}\sum_{x=1}^{W}\sum_{y=1}^{H}\sum_{c=1}^{C}(R(\text{sh})-\tilde{R}(\text{sh}))^2 \tag{12}$$

式中：W、H、C 分别为图像的宽、高和通道数；$R(\text{sh})$ 为给定 SH 系数数组时像素的颜色通道值。因此，总损失函数如式(13)所示：

$$L=w_1L_1+w_2L_2 \tag{13}$$

式中：w_1、w_2 分别为 SH 损失和渲染损失的权重，本文取 $w_1=0.8$、$w_2=0.2$。SH 损失用于约束 SH 系数向量的计算准确度，渲染损失用于约束估计光照的一致性。

3 实验与结果分析

3.1 实验环境

为验证算法的有效性，本文针对文献[19]提供

的数据集进行了网络消融实验和方法对比实验。数据集包含 400 张 HDR 全景图，其中室内场景有 79 张，室外场景有 321 张，每张 HDR 图裁剪出 500 对图片，模拟移动设备的前后置相机所拍摄的图片。实验软硬件环境配置如表 1 所示，实验超参数设置如表 2 所示。

表 1 实验环境配置

Table 1 Experimental environment configuration

实验环境	配置
操作系统	Ubuntu20.04.2LTS
GPU	NVIDIA GeForce RTX 4090
CUDA	CUDA 11.3
编程语言	Python 3.7
开发工具	Pycharm 2023.1.2
可视化工具	Tensorboard 2.11.2
深度框架	PyTorch 1.12.0

表 2 实验参数设置

Table 2 Experimental parameter settings

参数	设置
学习率	0.01
优化器	Adam
批处理大小	64
权重衰减	0.000 1
线程数	8
训练轮数	100

3.2 评价指标

本文采用均方根误差(RMSE)作为光照估计任务评价指标。RMSE 计算如式(14)所示：

$$V_{\mathrm{RMSE}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-y_i')^2} \tag{14}$$

式中：y_i 表示真实值；y_i' 表示预测值。RMSE 值越小，表示模型效果越好。然而，在光照估计任务中，仅依靠 RMSE 无法全面地评估预测图像与真实图像之间的相似性。结构相似度(SSIM)不仅可以量化图像在明暗度、对比度和结构性等方面的相似性，还符合人类直觉感知，能够反映图像之间的差异是否在可接受范围内。SSIM 计算公式如式(15)所示：

$$\mathrm{SSIM}(x,y)=\frac{(2u_xu_y+C_1)(2\sigma_{xy}+C_2)}{(u_x^2+u_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \tag{15}$$

式中：u_x、u_y 分别为图像 x、y 的均值；σ_x^2、σ_y^2 分别为图像 x、y 的方差；σ_{xy} 为图像 x、y 的协方差；C_1、C_2 为常数，通常取 $C_1=(K_1L)^2$、$C_2=(K_2L)^2$，$K_1=0.01$，$K_2=0.03$，$L=255$(像素值的动态范围)。

本文采用的另一个测量环境照明属性的指标是结构相异性(DSSIM)，其计算方法如式(16)所示：

$$\mathrm{DSSIM}(x,y)=\frac{1-\mathrm{SSIM}(x,y)}{2} \tag{16}$$

通常情况下，RMSE、DSSIM 指标值越小，代表模型性能越好，图像失真越小。通过调整学习率和优化器参数，可以有效提升神经网络的收敛性能，进而提高其预测精度和泛化能力。

3.3 网络消融分析

为了验证本文 MSAI 和 LPConv 模块的贡献以及所提网络框架的有效性，采用以下方式进行消融实验：1)编码器仅由标准多头自注意力模块、多层感知机 2 个部分组成，记为 V-A；2)编码器由 MSAI 机制、多层感知机 2 个部分组成，记为 V-B；3)编码器由标准多头自注意力、LPConv 模块、多层感知机 3 个部分组成，记为 V-C；4)编码器由 MSAI 模块、LPConv 模块、多层感知机 3 个部分组成，即本文的光照估计网络 ViTLight。

本文猜测 MSAI 模块的多头个数对网络表现能力有影响，于是针对其进行对比分析。分别选取多头个数 number 为 4、8、12、16 进行实验，结果如图 5 所示，从中可以看出，当 number 为 12 时 RMSE 和 DSSIM 均最低，此时效果最好。

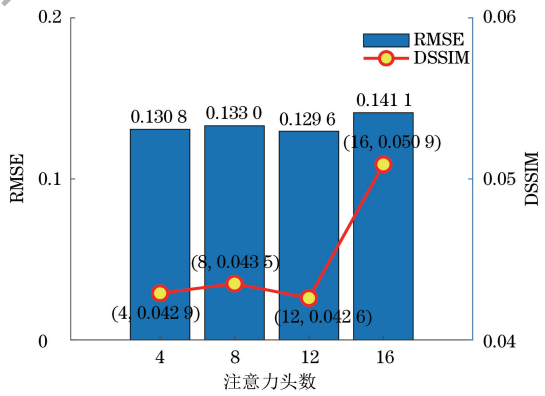


图 5 不同注意力头个数下的指标变化

Fig. 5 Changes of indicators under different numbers of attention heads

本文选取多头自注意力的多头个数为 12，4 种网络结构的消融实验结果如表 3 所示，最优结果加粗标注。从中可以看出：V-B 的 RMSE 和 DSSIM 值相较于 V-A 略低，整体性能有所提升，这表明加入 MSAI 模块后图片分块间的联系增加，使得渲染后的图像更加接近真实图像，从而提高网络的准确性；V-C 的误差相比于 V-A、V-B 大幅下降，这是由于 V-C 中增加了类似于 CNN 的局部特征提取和学习模块，并通过滑动窗口的方式在整个输入数据上移动，从而对整个输入数据进行特征提取；本文的 ViTLight 与前 3 种方法相比误差更小，RMSE 和

DSSIM 相较于标准 ViT(V-A)分别下降 0.124 6、0.034 2,这得益于新增的 2 个模块的作用,使得能够提取更丰富的全局和局部信息。

表 3 网络消融结果分析

Table 3 Analysis of network ablation results

网络	MSAI	LPConv	RMSE	DSSIM	Loss
V-A			0.254 2	0.076 8	**0.330 4**
V-B	√		0.253 8	0.075 4	0.361 4
V-C		√	0.175 7	0.079 2	0.370 2
ViTLight	√	√	**0.129 6**	**0.042 6**	0.357 7

图 6 是框架改进前后的可视化效果对比,分别对应 V-A、V-B 作为特征提取网络进行训练时渲染得到的虚拟物体掩码与 Groud truth 的对比,图 6(a)是标准 ViT 的训练结果,图 6(b)是改进后 ViT 的训练结果。可以看出,图 6(a)得到的光照偏向曝光状态,图 6(b)的虚拟物体上的光照信息比图 6(a)更接近右侧 Groud truth 图片,验证了改进特征提取网络的有效性。

(a)标准ViT的训练结果 (b)改进后ViT的训练结果

图 6 模型改进前后效果

Fig.6 Effect before and after model improvement

为了验证局部感知模块位置对模型性能的影响,将该模块分别放置在多层感知机之前和之后进行实验,图 7 展示了实验结果。

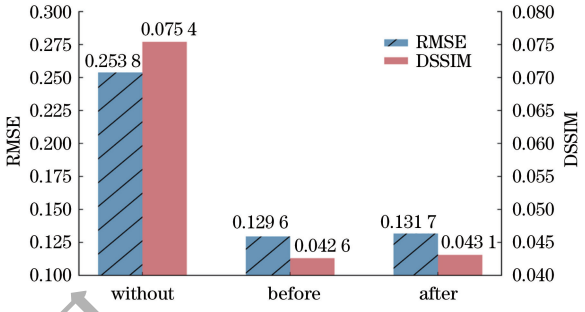


图 7 LPConv 位置对模型性能的影响

Fig.7 The impact of LPConv location on model performance

在图 7 中,横坐标“without”表示没有添加该模块,“before”表示该模块设计在多层感知机之前,“after”表示该模块设计在多层感知机之后。双轴纵坐标分别表示 RMSE 和 DSSIM 值。实验结果表明,将该模块设计在多层感知机之前可以增强模型对图像局部特征和光照信息的感知能力,从而提高模型的性能表现。

图 8 是网络消融的可视化效果对比,第一列和第二列是光照估计框架的成对输入图片,场景 1 和场景 4 选取室内环境,场景 2 和场景 3 选取室外环境,场景 1 光线较暗,场景 3 光线较亮。图 8 表明,V-A 渲染得到的物体与 Ground truth 相差较大,随着网络各模块的加入,V-B、V-C 更接近真实图像。本文的 ViTLight 光照估计结果(即图 8 中的第六列)预测和渲染准确度更高,可以更好地表现真实的光照信息,同时对于不同场景具有更强的鲁棒性。

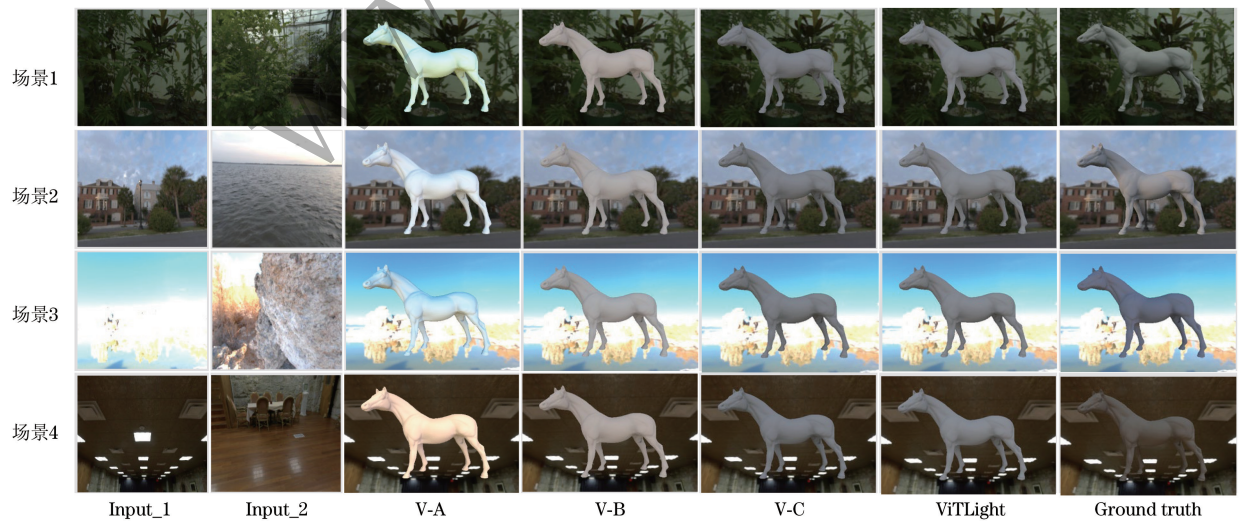


图 8 网络消融可视化分析

Fig.8 Visualization analysis of network ablation

3.4 对比实验分析

本文首先将 ViTLight 与 6 种经典特征提取网络进行对比,包括不同尺度模型大小的 ViT[27] 和 MobileNetV3[31]、ShuffleNet[32]、GhostNet[33],以寻找最适合光照估计的主干网络。选择轻量化网络的主要目的是提升计算效率,降低训练成本。不同大小 ViT 的比较旨在保证准确率同时提高效率,以满足场景实时和资源受限的应用需求。各网络模型的

光照估计效果如图 9 所示。图 9 表明，采用 ShuffleNet 和 GhostNet 模型只能得到少量光照信息，虚拟物体表面渲染结果较暗，ViT 和 MobileNetV3 尽管能得到较多光照信息，但效果仍不够准确，而 ViTLight 不仅能较好地恢复场景光照的准确率，而且可以实现视觉体验的一致性。

为了验证算法的稳定性，选择在室内场景、室外场景和室内外混合场景下分别进行实验，不同特征提取网络的对比结果如表 4 所示。轻量级网络的实验损失结果普遍高于 ViT，且与其他 2 种不同尺度的 ViT 相比，ViT-large 在处理室内外混合场景光照时的损失较低，但由于其参数量庞大，因此未得到选择使用。ViT-small和ViT-base的效果相近，但在提取室外场景光照信息方面，ViT-base 更为出色，其 RMSE 和 DSSIM 值最低，因此，将 ViT-base 作为主干特征提取网络是合适且有效的选择。

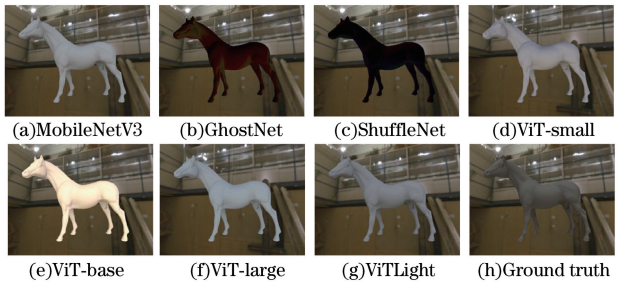


图 9 不同网络模型的光照估计效果

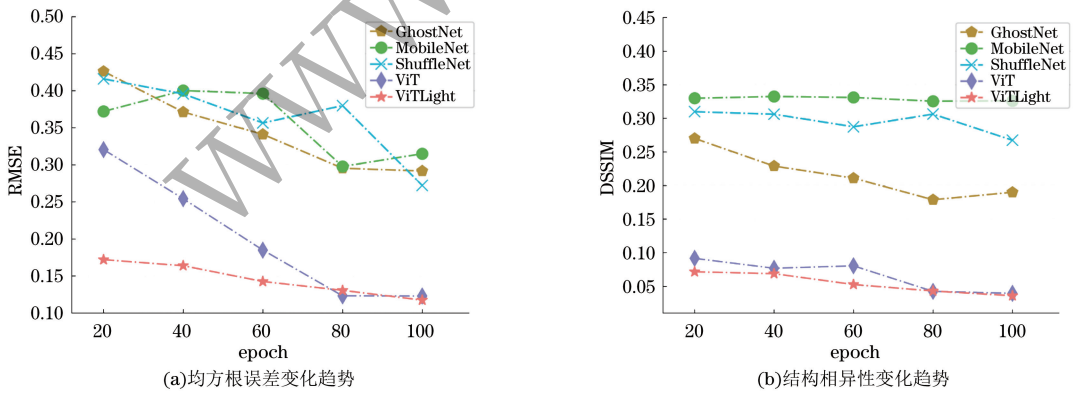
Fig. 9 The effectiveness of illumination estimation using different network models

表 4 不同特征提取网络的评价结果

Table 4 Evaluation results of different feature extraction networks

网络	室内场景		室外场景		室内外混合场景	
	RMSE	DSSIM	RMSE	DSSIM	RMSE	DSSIM
MobileNetV3	0.275 6	0.274 5	0.374 6	0.329 6	0.134 1	0.044 2
GhostNet	0.227 3	0.163 7	0.283 3	0.186 5	0.359 8	0.153 3
ShuffleNetV1	0.272 2	0.267 3	0.356 5	0.287 3	0.365 4	0.151 9
ViT-small	0.209 5	0.068 7	0.201 5	0.063 4	0.240 3	0.084 9
ViT-base	0.184 9	0.080 4	**0.122 9**	**0.039 5**	0.254 2	0.076 8
ViT-large	0.227 3	0.066 8	0.216 9	0.062 5	0.223 4	0.067 7
ViTLight	**0.139 1**	**0.063 6**	0.125 8	0.046 0	**0.129 6**	**0.042 6**

图 10 显示了不同特征提取网络的评价指标变化趋势，结果表明，本文模型的评价指标始终优于其他模型，并且整体上呈现较好的稳定性。



(a)均方根误差变化趋势　(b)结构相异性变化趋势

图 10 不同网络的评价指标变化

Fig. 10 Changes in evaluation indicators of different networks

不同尺寸的分块图像对 ViTLight 框架光照估计的影响如表 5 所示。随着分块尺寸的减小，单张图像的分块数量增加，导致训练时间上升。当图像分块大小 $P=16$ 时，场景光照损失最小，能够最大程度地保证虚实光照一致性效果。

其次，为了评估 ViTLight 框架在光照估计方面的性能，本文将其与光照估计领域中的文献[12]方法、文献[15]方法、文献[19]方法、文献[25]方法

表 5 不同图像分块大小对光照估计的影响

Table 5 The influence of different image block sizes on illumination estimation

指标	图像分块大小			
	32	16	14	8
RMSE	0.221 0	**0.129 6**	0.208 7	0.234 6
DSSIM	0.064 3	**0.042 6**	0.067 7	0.070 0
Training Time/h	30	**27**	42	51

这 4 种方法进行对比，结果如表 6 所示。从结果可以看出，本文的光照估计模型在面对室内外混合场景时能够达到较优的综合性能，其 RMSE 和 DSSIM 值分别低至 0.129 6 和 0.042 6，原因是本文提出的框架能够获取更加丰富的场景信息和照明细节。

表 6 不同光照估计方法的评价结果

Table 6 Evaluation results of different illumination estimation methods

方法	室内场景		室外场景		室内外混合场景	
	RMSE	DSSIM	RMSE	DSSIM	RMSE	DSSIM
文献[12]方法	0.197 1	0.088 4	0.254 2	0.076 8	0.240 2	0.066 4
文献[15]方法	0.184 9	0.080 4	0.126 9	0.068 6	0.174 6	0.072 4
文献[19]方法	0.131 1	0.042 9	**0.119 4**	**0.040 6**	0.134 1	0.044 2
文献[25]方法	0.154 4	0.049 7	0.143 7	0.072 9	0.130 8	0.042 9
ViTLight	**0.130 4**	**0.042 8**	0.125 8	0.046 0	**0.129 6**	**0.042 6**

4 结束语

本文提出了一种基于改进 ViT 的光照一致性估计框架 ViTLight。ViTLight 框架通过引入多头自注意力交互机制和局部感知卷积模块，有效解决了传统光照估计框架在相机视角受限和复杂光照环境中局部特征提取能力不足的问题。实验结果验证了所提方法的有效性和实用性。对于室内外混合数据集，本文方法在均方根误差与结构相异性评价指标上优于文献[12,15]所提方法。下一步将考虑如何降低模型的训练和推理成本，并提升用户在增强现实场景下的真实体验质量。

参考文献

[1] 刘万奎，刘越. 用于增强现实的光照估计研究综述[J]. 计算机辅助设计与图形学学报，2016，28(2)：197-207.
LIU W K，LIU Y. A review of illumination estimation for augmented reality[J]. Journal of Computer-Aided Design & Computer Graphics，2016，28(2)：197-207. (in Chinese)

[2] MOHAMMADKHORASANI A，MALEK K，MOJIDRA R，et al. Augmented reality-computer vision combination for automatic fatigue crack detection and localization [J]. Computers in Industry，2023，149：103936.

[3] CAO J，LAM K Y，LEE L H，et al. Mobile augmented reality：user interfaces，frameworks，and intelligence[J]. ACM Computing Surveys，2023，55(9)：1-36.

[4] 滕嘉玮，赵岩，张艾嘉，等. 基于二维仿射变换的几何一致性虚实融合[J]. 光学 精密工程，2022，30(11)：1374-1382.
TENG J W，ZHAO Y，ZHANG A J，et al. Virtual-real fusion with geometric consistency based on two-dimensional affine transformation[J]. Optics and Precision Engineering，2022，30(11)：1374-1382. (in Chinese)

[5] LEGENDRE C，MA W C，FYFFE G，et al. DeepLight：learning illumination for unconstrained mobile mixed reality[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D.C.，USA：IEEE Press，2019：5918-5928.

[6] BHATT D，PATEL C，TALSANIA H，et al. CNN variants for computer vision：history，architecture，application，challenges and future scope[J]. Electronics，2021，10(20)：2470.

[7] GARON M，SUNKAVALLI K，HADAP S，et al. Fast spatially-varying indoor lighting estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D.C.，USA：IEEE Press，2019：6908-6917.

[8] ZHU Y J，ZHANG Y D，LI S，et al. Spatially-varying outdoor lighting estimation from intrinsics[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D.C.，USA：IEEE Press，2021：12834-12842.

[9] DONG S，WANG P，ABBAS K. A survey on deep learning and its applications[J]. Computer Science Review，2021，40：100379.

[10] SONG S R，FUNKHOUSER T. Neural illumination：lighting prediction for indoor environments[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D.C.，USA：IEEE Press，2019：6918-6926.

[11] WANG G C，YANG Y N，LOY C C，et al. StyleLight：HDR panorama generation for lighting estimation and editing[EB/OL]. [2023-10-05]. https：//arxiv.org/abs/2207.14811.

[12] GARDNER M A，SUNKAVALLI K，YUMER E，et al. Learning to predict indoor illumination from a single image[J]. ACM Transactions on Graphics，2017，36(6)：1-14.

[13] 曹天池，李秀实，李丹，等. 基于图像分解的光照估计算法[J]. 计算机工程与科学，2021，43(8)：1422-1428.
CAO T C，LI X S，LI D，et al. Illumination estimation based on image decomposition [J]. Computer Engineering & Science，2021，43(8)：1422-1428. (in Chinese)

[14] SOMANATH G，KURZ D. HDR environment map estimation for real-time augmented reality[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D.C.，USA：IEEE Press，2021：11298-11306.

[15] HOLD-GEOFFROY Y，SUNKAVALLI K，HADAP S，et al. Deep outdoor illumination estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington D.C.，USA：IEEE Press，2017：7312-7321.

[16] ZHAO J，CHALMERS A，RHEE T. Adaptive light estimation using dynamic filtering for diverse lighting conditions [J]. IEEE Transactions on Visualization and Computer Graphics，2021，27(11)：4097-4106.

[17] CHENG H，XU C，WANG J，et al. Fast and accurate illumination estimation using LDR panoramic images for realistic rendering[J]. IEEE Transactions on Visualization and Computer Graphics，2023，29(12)：5235-5249.

[18] 吴广运，周治平. 基于阴影检测的增强现实光照一致性实现[J]. 激光与光电子学进展，2022，59(2)：350-355.

WU G Y, ZHOU Z P. Realizing illumination consistency in augmented reality based on shadow detection[J]. Laser & Optoelectronics Progress, 2022, 59(2): 350-355. (in Chinese)
[19] CHENG D C, SHI J, CHEN Y Y, et al. Learning scene illumination by pairwise photos from rear and front mobile cameras[J]. Computer Graphics Forum, 2018, 37(7): 213-221.
[20] SUN Y K, LI D, LIU S, et al. Learning illumination from a limited field-of-view image[C]// Proceedings of the IEEE International Conference on Multimedia & Expo Workshops. Washington D.C., USA: IEEE Press, 2020: 1-6.
[21] LI Z Q, YU L, OKUNEV M, et al. Spatiotemporally consistent HDR indoor lighting estimation[J]. ACM Transactions on Graphics, 2023, 42(3): 1-15.
[22] LIU C L, WANG L Y, LI Z, et al. Real-time lighting estimation for augmented reality via differentiable screen-space rendering[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 29(4): 2132-2145.
[23] ZHANG A J, ZHAO Y, WANG S G. An improved augmented-reality framework for differential rendering beyond the lambertian-world assumption[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 27(12): 4374-4386.
[24] LI M T, GUO J, CUI X F, et al. Deep spherical Gaussian illumination estimation for indoor scene[C]//Proceedings of the ACM Multimedia Asia. New York, USA: ACM Press, 2019: 1-6.
[25] ZHAN F N, ZHANG C G, YU Y C, et al. EMLight: lighting estimation via spherical distribution approximation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(4): 3287-3295.
[26] 赵宏, 陈志文, 郭岚, 等. 基于 ViT 与语义引导的视频内容描述生成[J]. 计算机工程, 2023, 49(5): 247-254.
ZHAO H, CHEN Z W, GUO L, et al. Video content caption generation based on ViT and semantic guidance[J]. Computer Engineering, 2023, 49(5): 247-254. (in Chinese)
[27] HAN K, WANG Y H, CHEN H T, et al. A survey on Vision Transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 87-110.
[28] CHAUDHARI S, MITHAL V, POLATKAN G, et al. An attentive survey of attention models[EB/OL]. [2023-10-05]. http://arxiv.org/abs/1904.02874v3.
[29] ZHANG Q, YANG Y B. ResT: an efficient transformer for visual recognition[J]. Advances in Neural Information Processing Systems, 2021, 34: 15475-15485.
[30] GREEN R. Spherical harmonic lighting: the gritty details[EB/OL]. [2023-10-05]. https://www.cse.chalmers.se/~uffe/xjobb/Readings/GlobalIllumination/Spherical%20Harmonic%20Lighting%20-%20the%20gritty%20details.pdf.
[31] HOWARD A, SANDLER M, CHEN B, et al. Searching for MobileNetV3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Washington D.C., USA: IEEE Press, 2019: 1314-1324.
[32] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D.C., USA: IEEE Press, 2018: 6848-6856.
[33] HAN K, WANG Y H, TIAN Q, et al. GhostNet: more features from cheap operations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D.C., USA: IEEE Press, 2020: 1580-1589.

编辑 吴云芳