



中国科学院大学  
University of Chinese Academy of Sciences

# 硕士学位论文

## 基于深度学习的场景光照估计研究

作者姓名：\_\_\_\_\_

指导教师：\_\_\_\_\_

中国科学院软件研究所

学位类别：工学硕士

学科专业：计算机应用技术

培养单位：中国科学院软件研究所

2019年6月



**Deep Scene Illumination Estimation**

A thesis submitted to the  
University of Chinese Academy of Sciences  
in partial fulfillment of the requirement  
for the degree of  
Master of Engineering  
in Computer Graphics

By

Institute of Software, Chinese Academy of Sciences

**June, 2019**



## 中国科学院大学 学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日 期：

## 中国科学院大学 学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关保存和使用学位论文的规定，即中国科学院大学有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延期后适用本声明。

作者签名：

导师签名：

日 期：

日 期：



## 摘要

光照估计是计算机视觉和计算机图形中的基础问题之一，该问题意图从图片中恢复出场景真实的光照分布。光照估计问题与其它多种视觉和图形学问题息息相关。本文通过对已有光照估计方法的调研和分析，提出了一种基于深度学习的场景光照估计方法。该方法使用智能设备前后置相机拍摄的图片作为输入，预测出使用球形谐波函数近似的场景光照分布。为此本文搭建了一个端到端的深度卷积神经网络，并创新性地提出了一个损失函数——渲染损失（Render Loss），用于监督神经网络的训练，提高光照估计预测效果。本文的方法在可视结果和数值结果上都超过了目前最先进的光照估计算法，而且该方法在真实场景中的效果也非常理想。此外，本文还构建了一个大规模的高动态范围全景数据集，用于训练和测试光照估计网络。最后本文通过在光照估计数据集和光照估计网络上的大量实验，深入探究了数据规模、数据多样性、神经网络结构、损失函数等多个因素对于光照估计效果的具体影响。

**关键词：**光照估计，深度学习，球形谐波函数，高动态范围全景图



## Abstract

Illumination estimation is an essential problem in computer vision and graphics. It is closely related to many vision or graphic problems. In this paper, a learning based method is proposed to recover scene illumination represented as spherical harmonic (SH) functions by pairwise photos from rear and front cameras on mobile devices. An end-to-end deep convolutional neural network (CNN) structure is designed to process images on symmetric views and predict SH coefficients. We introduce a novel Render Loss to improve the rendering quality of the predicted illumination. Experiments show that our model produces visually and quantitatively superior results compared to the state-of-the-arts. A high quality high dynamic range (HDR) panoramic image dataset is developed for training and evaluation. Moreover, a series of experiment is conducted to explore how the data size, data diversity, network structure, loss function and other modules affect the illumination estimation results.

**Keywords:** Illumination Estimation, Deep Learning, Spherical Harmonic Lighting, Hight Dynamic Range Panorama



## 目 录

<b>第1章 绪论 .....</b>	<b>1</b>
1.1 选题的背景及意义 .....	1
1.2 国内外研究现状 .....	2
1.3 论文研究内容.....	12
<b>第2章 光照估计数据集的构建与研究 .....</b>	<b>15</b>
2.1 引言 .....	15
2.2 相关工作 .....	15
2.3 全景图像 .....	16
2.4 构建HDR全景数据集 .....	22
2.5 探究数据集对光照估计的影响.....	23
2.6 总结与讨论 .....	31
<b>第3章 基于深度学习的光照估计方法 .....</b>	<b>33</b>
3.1 引言 .....	33
3.2 相关工作 .....	34
3.3 问题求解范围.....	36
3.4 光照分布的球形谐波表示 .....	37
3.5 光照估计网络结构 .....	38
3.6 损失函数 .....	39
3.7 实验结果与评估 .....	41
3.8 深入研究光照估计网络 .....	44
3.9 讨论 .....	49
3.10 本章总结 .....	51
<b>第4章 总结与展望.....</b>	<b>53</b>
4.1 本文工作总结.....	53
4.2 未来工作展望 .....	55
<b>参考文献 .....</b>	<b>57</b>



## 图形列表

1.1 光照估计的应用 .....	1
1.2 光照表示模型的示例 .....	3
1.3 几种用于光照估计的光探测器 .....	4
1.4 使用特殊设备估计光照 .....	5
1.5 使用额外信息辅助估计光照的方法 .....	7
1.6 用户标记辅助估计光照 .....	10
1.7 光照的自编码器结构 .....	11
2.1 全景图的几种投影方式 .....	18
2.2 使用不同曝光拍摄的LDR图像 .....	18
2.3 LDR图片与HDR图片在调整曝光时的不同 .....	19
2.4 使用HDR全景图与LDR全景图渲染结果的对比 .....	21
2.5 拍摄全景图的示例 .....	21
2.6 光照估计数据集预览 .....	24
2.7 用于评估数据集的卷积神经网络结构 .....	26
2.8 数据规模对光照估计问题影响的趋势图 .....	27
2.9 Laval Indoor数据集中的全景图片 .....	29
2.10 本文数据集与其它工作的可视化结果比较 .....	30
3.1 球形谐波基函数的形状 .....	38
3.2 光照估计网络结果一览 .....	39
3.3 SH系数与渲染结果之间的不一致性 .....	41
3.4 与最先进方法的对比 .....	44
3.5 本文方法在真实场景中的结果 .....	45
3.6 不同的预训练参数引用方式对光照估计的影响 .....	47
3.7 损失函数对光照估计的影响 .....	49
3.8 光照预测失败的例子 .....	50



## 表格列表

2.1 全景相机参数列表 .....	22
2.2 曝光融合参数列表 .....	23
2.3 从HDR全景图中提取普通图像的参数列表 .....	25
2.4 数据规模对光照估计的影响 .....	27
2.5 数据多样性对光照估计的影响 .....	28
2.6 与现有数据集对比 .....	28
3.1 光照估计网络结构 .....	39
3.2 光照估计网络的训练细节。 .....	42
3.3 本文方法与最先进方法的对比 .....	43
3.4 不同特征提取网络的对比 .....	46
3.5 特征融合方式对光照估计的影响 .....	48
3.6 损失函数对光照估计的影响 .....	48



## 缩写列表

### 缩写

AE	Auto Encoder
AR	Augmented Reality
BRDF	Bidirectional Reflectance Distribution Function
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DSSIM	structural Dissimilarity
FOV	Field of View
HDR	Hight Dynamic Range
IBR	Image Based Rendering
LDR	Low Dynamic Range
MAE	Men Absolute Error
MSE	Mean Squared Error
PSNR	Peak Signal to Noise Ration
RELU	Rectified Linear Unit
RMSE	Root Mean Squared Error
SH	Spherical Harmonics
SHL	Spherical Harmonic Lighting
SOTA	State of The Art
SSIM	structural Similarity

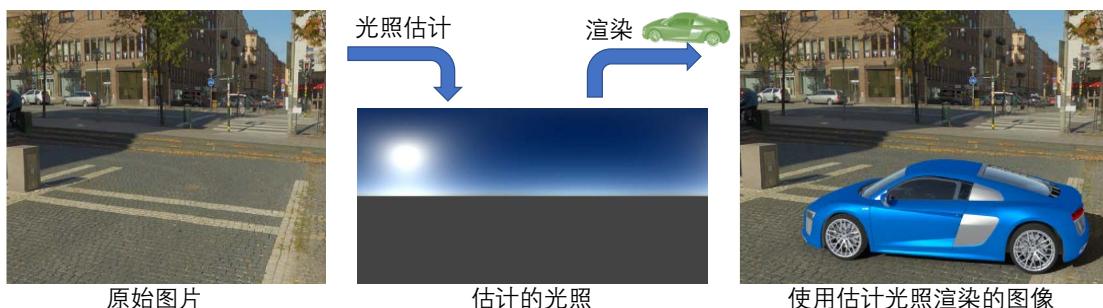


# 第1章 绪论

## 1.1 选题的背景及意义

光照估计（又称光照分布估计）是从已知的彩色图像信息中，预测、估计或恢复出整个场景的光照分布。该问题的输入通常是彩色图片或视频，有时已知的几何或材质信息也会被用来辅助估计光照。场景的光照分布是指场景中各个方向上光照的颜色和强度。较为常见的光照分布表示方法包括高动态范围（High Dynamic Range, HDR）全景图、球形谐波（Spherical Harmonics, SH）光照、基于物理的Sun-Sky光照模型等。其中精度最高、使用比较广泛的是HDR全景图像，而在实时渲染领域使用较多的是SH光照，基于物理的光照模型则多用于室外场景。

光照估计作为计算机图形学和计算机视觉的基础问题之一，有着广泛的实际应用场景。例如：基于图像的渲染（Image Based Rendering, IBR）、增强现实（Augmented Reality, AR）、电影后期制作、真实感虚实交互等。图1.1展示了光照估计的应用之一。光照估计也与这两个学科中的许多其它问题息息相关。例如：双向反射分布函数（BRDF）估计、场景几何重构、本征信息提取、图像增强等等。高质量的光照估计结果通常能够为这些问题的解决带来很大的帮助。



**图 1.1** 光照估计的应用之一。使用单张图片估计场景的光照，并利用估计的光照渲染一个新的物体合成到图像中。可以看出使用估计光照渲染后的3D物体，与原场景在视觉上较为一致。图片引自[1]。

从有限的图像信息估计出整个场景的光照分布是一个复杂的问题。一方面，图像的视野范围有限，例如一张视场角（FOV）为 $60^\circ$ 的照片所拍摄到的区域在其对应的全景图中占比还不足6%。另一方面，一幅图片是光照分布、场景几何结构、物体材质、相机参数等多个单位之间的复杂交互结果（公式 1.1）。

$$Image = ComplexInteraction(Light, Geometry, Material, Camera) \quad (1.1)$$

通过公式1.1可以看出，在其它三个信息未知的情况下，从图像（Image）反推出光照（Light）几乎是一个不适定（ill-posed）问题。不仅如此，在各种条件下拍摄的彩色图像可能存在许多不正确的颜色信息。例如图像中的过曝光/欠曝光区域、相机畸变、不正确的白平衡等，这些都会对光照估计造成一定程度的干扰，增加光照估计的难度。

为了简化问题难度，传统方法常常增加输入信息的数量或缩小光照模型的规模。一部分工作尝试借用额外的输入信息辅助估计场景光照，例如深度信息、几何信息、额外的输入图片、先验知识等等。这类方法通常依赖特殊的探针、特殊的拍摄设备、额外的用户辅助信息等，具有一定的局限性。另一部分工作通过使用低维的光照表示模型来简化光照估计问题，例如使用球形谐波函数（SH）来拟合场景光照、使用小波函数近似场景光照、使用有限的点光源集合近似场景光照、使用基于物理的室外光照模型等等。可以看出，无论是增加输入信息还是使用简化的光照估计模型，传统光照估计方法均具有一定的局限性。

近年来，深度学习在多种计算机视觉问题上大放异彩，用于分割、检测、标识、分类的神经网络层出不穷。一些研究者尝试将深度学习应用在光照估计问题当中。其中Hold-Geoffroy等人[1]和Gardner等人[2]的工作是应用深度学习估计光照的最先进方法。它们在大规模数据集上训练了一个深度卷积神经网络，分别用于室内和室外的场景光照估计。不过，现有的深度学习方法也有一定的局限性。训练一个鲁棒的神经网络往往需要大量的数据，而目前用于训练光照估计神经网络的数据集比较有限，主要包括：大规模的低动态范围全景数据集（如SUN360[3]等）和中小规模特定场景的高动态范围全景数据集（如Laval Indoor等[2]）。这些数据集在规模和质量上很难同时到达训练深度神经网络的要求。

在这样的背景下，本文在光照估计的两个方向开展研究。其一是构建一个具有一定规模和质量的光照估计数据集，用来训练更加高效、鲁棒的光照估计神经网络。其二是在已有数据集和本文构建的数据集基础上，深入探索基于深度学习的光照估计方法，对其中的网络结构，网络参数，损失函数，光照表示等多个模块进行细致的研究和分析。

## 1.2 国内外研究现状

### 1.2.1 光照的表示

光照的表示对光照估计问题至关重要。场景的光照分布有着多种多样的表

示方法。其中高动态范围 (High Dynamic Range, 简称HDR)全景图是一种被广泛使用而且精度较高的光照表示方式。与普通的8位三通道图像不同，HDR图像的颜色值范围可以从0取到非常大。这意味着HDR图像可以更细致地表示每个像素位置的真实亮度值。因此无论是HDR全景图表示的光照，还是使用HDR全景图渲染的物体，都能与真实值更加接近。Reinhard等人在*High Dynamic Range Imaging*[4]中对HDR图像进行了详细的介绍。

研究者们也提出了一些小巧、高效的光照近似模型，代价是牺牲一部分精度。Ramamoorthi和Hanrahan[5]提出使用球形谐波函数来表示场景的光照。这种方法表示的光照，不仅参数规模较小，而且在渲染物体时非常高效。他在文章中指出，该模型仅使用9个系数就可以在漫反射物体的渲染结果上达到平均值不超过1%误差。不过由于球形谐波函数本身的限制，这种方式不能很好的保留光照分布的高频细节。尽管如此，它的简洁和高效使得它成为实时渲染中最常用的技术之一[6, 7]。许多光照估计方法采用了这种模型来近似场景的光照分布。



图 1.2 三种不同的光照表示模型对比

在室外场景中，光线的主要来源是太阳光、光在空气中的散射和折射、和地面对光的反射。因此基于物理的光照表示模型更适合表示室外场景。Perez等人[8]提出了一个参数化的物理模型来表示天空中的发光度 (Luminance) 分布。之后的许多工作 [9–12]在考虑了空气散射和折射、大气浑浊度的影响后，将该模型扩展为多种不同类型的颜色模型。Hosek和Wilkie[13, 14]基于这些工作，考虑并加入了更多因素，整理出一个可调节太阳位置、大气浑浊度、地面反照率的物理参数模型。这类模型的参数量较小，在近似室外光照时精度也很可观，因此成为室外场景光照估计问题中常用的光照表示方法。图1.2展示了不同的几种光照表示示例。

光照分布也有很多其它的表示方式，例如Ng等人[15]提出使用小波函数来

近似光照分布。LeGendre等人[16]提出了一个实用的框架，能够使用LED灯台精确地重构出各个方向的场景光照。近期Weber[17]结合深度学习，使用自编码器（auto-encoder）光照分布进行建模，为光照估计问题提供了新的思路。

### 1.2.2 使用光探测物的光照估计方法

光照估计是一个复杂的问题，研究者们常常采用多种方法的组合来解决这个问题。使用光探测物（light probe）来估计光照就是这些方法之一。常见的光照探测物包括：镜面球体，漫反射球体，镜面反射/漫反射混合球体，人眼，人脸，人手等等。Debevec[18]建立了一个基于光照模型的虚拟物体插入系统。其中所涉及的光照就是使用镜面反射球体作为光探测物来估计的。该工作首次提出高动态范围全景图可以通过拍摄不同曝光下的镜面球体来获取。在之后的工作中，Reinhard等人[4]和Debevec等人[19]指出，使用一个漫反射球体，或者漫反射与镜面反射表面混合的球体，也可以达到同样的效果。值得一提的是，Debevec在其工作[18]中，将远距离场景、近距离场景、以及待插入的虚拟物体分离开，并假设插入的虚拟物体并不会造成远距离场景的变化。之后的图像合成、增强现实等工作都遵从了这个假设。

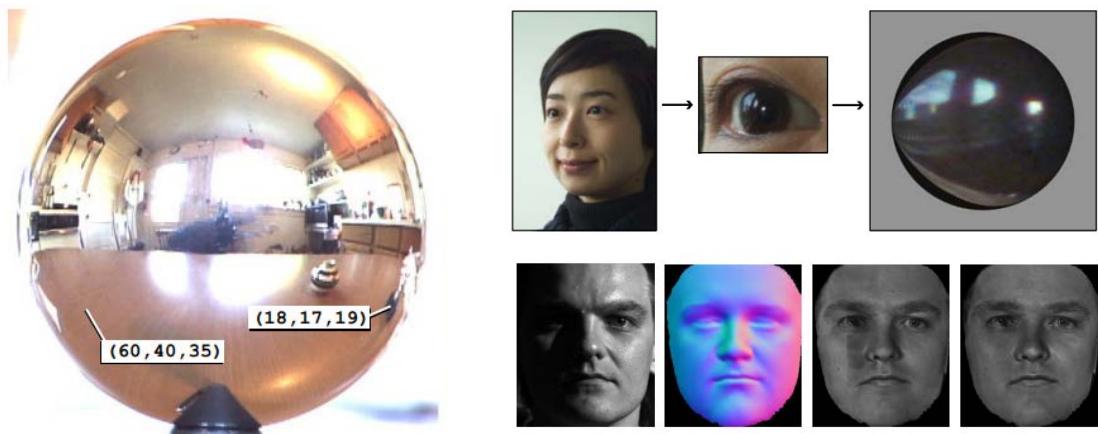


图 1.3 用于光照估计的光探测物——球[18]、人眼[20]、人脸[21]

上述的光探测物大多是已知物体的几何和表面材质、且预先放置在场景中的规则物体。还有一些特殊的“物体”也可以作为光探测物。例如人的眼球，面部等一些经常出现照片中的元素。Tsumra等人[22]在假设人眼是规则球体的前提下，利用眼球上光线的反射，估计场景的光照。Nishino和Nayar[20]则更细致地分析了眼球的大致结构，并利用包含眼球的照片估计场景的光照分布。不过该方法需要分辨率和清晰度较高的相机，而且其文中也指出他们的方法并没有考虑瞳孔颜色、虹膜颜色等因素，而这些因素都会对光照估计的

结果产生较大影响。人脸作为照片中经常出现的物体，也经常被用作光探测物。Wen等人[23]通过一张人脸照片，估计出光照的SH表示进而对人脸实现重照（relight）。Wang等人[21]提出来一种基于马尔科夫随机场的能量最小化框架，意图从正脸照片中恢复出人脸的形状、反照率和光照。在之后的工作中，Shim[24]、Knorr等人[25]、Shahlaei和Blanz[26]进一步探索了如何从人脸照片中估计较为精确的光照。Yao等人[27]使用普通相机和深度相机下的人体手部图像，通过人手的亮度和法线，估计出由球形谐波系数近似的低频光照。

可以看出，这类估计光照的方法都需要一个已知其类型、几何、材质的光探测物存在于场景内，图1.3展示了常见的光探测物。无论是规则的球体，还是人体的各个部位，它们都需要照片中包含指定的物体和元素。而在大多数的光照估计问题很难保证这些标志物或探测物一定存在，这无疑限制了此类方法的应用。

### 1.2.3 使用特定设备的光照估计方法

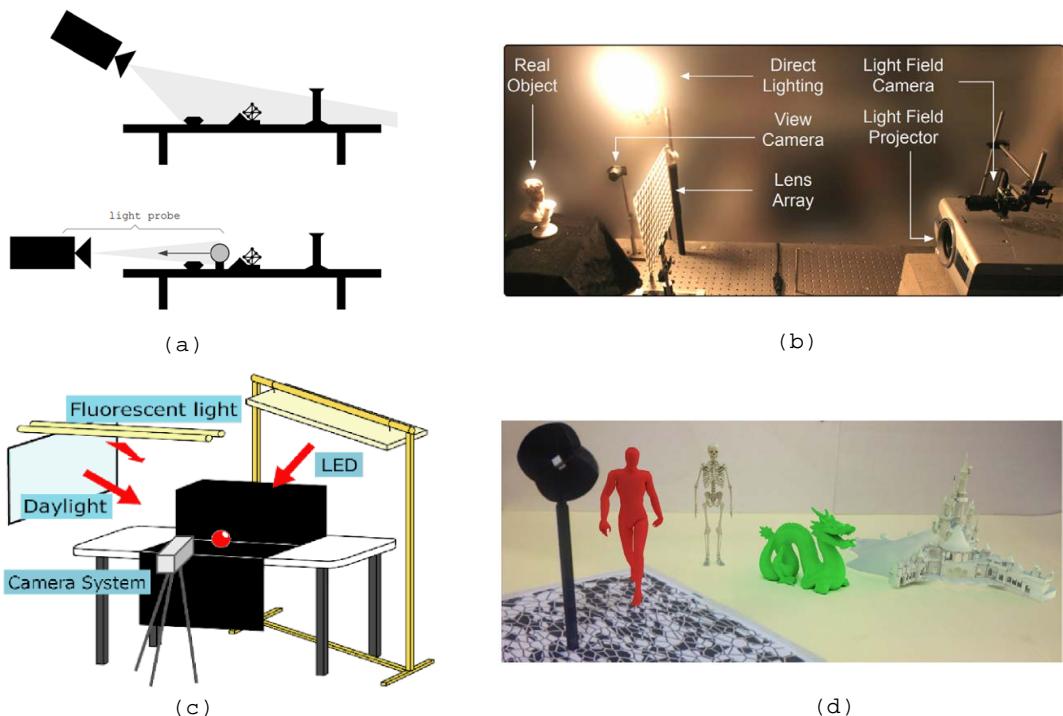


图 1.4 几种使用特殊设备估计场景光照的系统:(a)Debevec[18]通过多次拍摄获取光照的示意图；(b)Cossairt等人[28]的复杂拍摄系统；(c)Imai等人[29]使用的拍摄设备；(d)Cailian等人[30]用于AR光照估计的设备示意图。

降低光照估计问题难度的另一个思路是使用特定的设备、装置或应用辅助估计光照。Pilet等人[31]使用多个不同位置的相机和一个平面标定物构建了一

一个3D估计系统。通过追踪该平面标定物计算其中的高光和阴影，并进一步估计出场景中的几何和光照。Yoo和Lee[32]提出了一个由鱼眼镜头、中性衰减片(Neural Density Filter, ND)和普通相机构成的光照探测系统。他们通过鱼眼镜头获取一个半球面，使用ND片直接感知明亮区域，并通过一系列算法实时地估计场景中的光照。Cossairt等人[28]使用一组透镜、光场相机、光场映射器，构建了一个适合单光源较暗场景下的光照估计系统。随后Imai等人[29]使用多光谱成像设备，探究了可变亮度阈值、色调、偏振滤镜在检测光照条件复杂场景下的不同作用，提出了适用于偏高光反射物体上的光照估计方法。Tocci等人[33]则直接构建了一个光学设备，用以获取影视级别的高动态范围视频。Manakov等人[34]提供了一个插件类型的相机硬件，用以构建与高动态范围图像，多光谱、偏振和光场等相关的应用。Cailian等人[30]使用一个阴影探测器来解决实时增强现实中的光照估计问题。Kán和Peter[35]通过全景图像拼接技术，建立了一个在智能相机中捕提高动态范围全景图的应用。不过，通过这些使用特殊设备或者特定应用获取场景光照的方法并不是严格意义上的光照估计。他们其实是倾向于解决一些需要快速获取高动态范围图像或视频的工程问题。图1.4展示了上述的一些光照估计系统的使用情景，可以看出，这种光照估计方法所需的设备通常比较复杂。虽然在实验室环境中能够有效运行，但是这些设备的类型和数量无疑提高了这类方法的应用门槛。

#### 1.2.4 使用额外信息的光照估计方法

从单张图片恢复或估计出整个场景的光照分布是个复杂的问题。借用额外的信息是光照估计问题中最常用的方法。借助的信息通常有深度图像信息、场景几何形状、物体几何结构、先验知识等等。

深度信息对于光照估计问题有着很大的帮助。Knecht等人[36]借助深度相机和鱼眼镜头重建场景模型并提取了光源位置。Meiland等人[37]利用深度相机实时地创建稠密的高动态范围全景图，并使用K-means算法从环境贴图中提取点光源位置。Barron和Mailik[38]提出了一种从彩色图像和深度图像中提取本征信息的方法。他们使用具有噪声的深度图像来辅助估计场景的几何结构，并在这些信息的基础上进一步提取彩色图像中的本征信息。Zhang等人[39]使用RGBD相机，将拍摄到的室内场景建模为一个包含光源、材质和几何的空的房间模型，并提供了编辑模型场景的方法。

在已知场景几何或物体几何结构的情况下估计场景光照的分布也是主要的研究方向之一。例如，在表面形状类似桌子、平台、平板等结构的场景中，Li等

人[40]的方法可以结合图片中的阴影、明暗、高光信息估计出多个方向的光源信息。Ramamoorthi和Hanrahan[41]使用物体的几何结构和多个视角下的照片作为输入，估计出照片中物体的表面材质并进一步计算所在场景的静态光照信息。不过该方法需要多种假设——曲线表面、各向同性BRDF、没有多重反射等。Sato等人[42]指出了通过物体遮挡关系估计场景光照的可行性和有效性，并通过分析给定几何的物体上的明亮区域和阴影区域，估计出了较为真实的场景光照。在给定朗伯体（Lambertian）反射的物体几何时，Wang和Samaras[43]的方法可以从单张图片估计出多个方向的光照信息。Panagopoulos等人[44]提出了一个新的框架，来从单幅图片和粗糙的3D几何中，恢复出光照环境和估计场景中的投影阴影，该方法描述了一个高阶马尔可夫随机场（MRF）照明模型，将低阶阴影证据与高阶先验知识相结合，用于投影阴影和照明环境的联合估计。

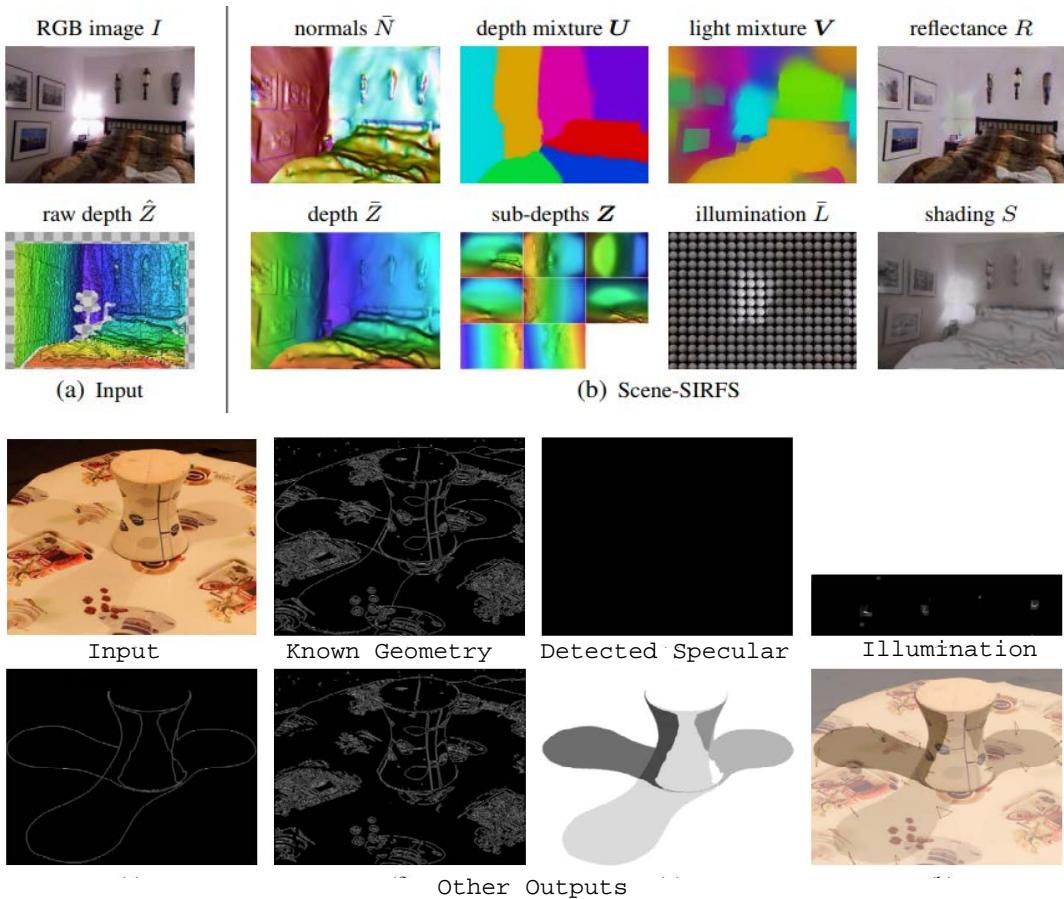


图 1.5 借助深度信息和几何信息估计场景中的光照分布。上图是Barron和Mailk[38]使用Depth信息作为额外的信息估计光照。下图则是Li等人[40]使用额外的几何信息作为额外信息估计光照。

此外还有一些方法在已知物体类型的情况下，利用图片中物体的假设约束

或先验知识来帮助估计光照。例如前文提到的Nishino和Nayar[20]假设人的眼球是规则的球体，并基于此假设从人眼估计场景的光照。类似的方法[45, 46]虽然不需要预先知道精确的场景几何结构，但都要通过图像中物体的类型得到一些先验知识、假设和约束，进而分析、预测、估计、恢复出场景光照。

增加彩色图片的数量也会对光照估计有很大帮助。Sato等人[47]使用两张全方位全角度的照片构建出场景的大致几何，然后根据不同快门速度拍摄的全方位图像序列计算场景的辐射度，并将其映射到构建的几何模型上，这种辐照度分布就可以作为光照来渲染新的3D物体。Nishino等人[48, 49]在已知物体的大致几何的情况下，分别为发光体和普通物体拍摄多张图片来估计场景的光照。Yu等人[50]则通过多视角图片来恢复出固定光照条件下的物体纹理和光照分布。Wu等人[51]建立了一个纯基于图片的形状、表面、光照的估计模型。Shan等人[52]在解决重构场景的问题时，提出了一种从大规模不同光照条件图片集中估计结点反照率和光照参数的方法。受到该方法的启发，Lalonde和Matthews[53]提出了一种从多张室外建筑物图片恢复对应光照分布的方法。

需要注意的是，以上几种增加辅助信息的方式并不是互斥的，研究者们可以选择使用多种额外信息共同辅助估计光照。例如Marschner[54]在已知物体几何的前提下，使用一组照片估计出物体表面的反射情况，进而估计出场景光照信息。Haber等人[55]利用一个物体在多个角度下的图片和已知的几何结构估计其光照分布。使用一种或多种额外的信息来辅助光照估计一直是解决光照估计问题的主要方法，也是能够提升光照估计的效果的主要方式。但是此类方法通常需要使用额外的输入设备（例如深度相机），繁琐的获取步骤（例如多次拍摄），以及一些先验知识。而且使用的额外信息越多，所需的设备就越多，步骤就越繁琐，这难免限制了它们的应用场景。

### 1.2.5 使用简化光照模型的光照估计算法

光照估计是一个已知条件较少，求解结果复杂，涉及因素繁多的问题。除了使用额外的设备增加输入信息规模的方法外，降低该问题难度的另一个思路是简化光照的表示。从有限的信息中估计出复杂的场景光照分布是比较困难的，所以许多方法尝试选取相对简单的光照近似模型表示光照，即通过牺牲一部分精度来降低待估计参数量的规模，从而达到简化光照估计问题的目的。其中最为常见的就是使用球形谐波（spherical harmonic）函数近似的光照模型。该模型最早由Ramamoorthi[5]在2001年提出。通过应用该模型，光照的低频部分可以使用少量的系数（通常为9-16组，约27-48个）来近似。虽然这种表示方式对光照

分布中的高频细节不太友好，但在渲染常见的漫反射物体时却有着极小的误差。因此许多工作[25, 41, 45, 56–58]通过估计少量的SH系数来达到估计光照的目的。与之类似，Barronhe和Malik[59]在光照估计问题中使用小波函数来近似场景光照，并将其与SH表示进行了对比，指出了在表示光照时球形谐波函数相较于小波函数的优点。早期的一些光照估计工作[40, 42–44, 47]将光照分布简化为若干个点光源的集合。进而将光照分布估计问题转化为预测光源数量、位置和大小的问题。预测这种类型的光照比较简单，但遗憾的是真实场景中的光照分布往往比较复杂，能够使用这种类型表示的场景并不是很多。

对于特定的场景（比如室外），基于物理模型的光照表示是一个很好的选择。这种模型往往能够使用很小的参数量近似出精度较高的光照分布。其中最具有代表性的是Perez[8]在1993年提出的天空的发光度分布模型。该模型经过多次补充，修改和完善[9–14]，目前已经成了一个包含太阳位置、大气浑浊度、地面反照率等多个具体物理意义参数的物理模型。因此大部分室外光照估计方法[60–63]都采用了这类表示模型。

值得注意的是，几乎所有的光照估计算法，都对要估计的光照进行一定的简化。选取合理的光照分布简化方法对于光照估计问题至关重要，需要综合考量已知条件，应用需求，核心方法框架等。

### 1.2.6 基于用户交互的光照估计算法

还有一类方法使用交互式方法辅助估计光照，这类方法通常需要用户在给定的图片中标注一些辅助信息。Lopez等人[46]提出了一种需要用户辅助的光照估计方法，该方法需要用户在场景中选择一个凸多面体物体，随后估计出粗糙的光源位置，并使用K-means算法将这些光源进行聚类、合并，最后在计算每个光源的精确位置和光照强度。该方法在面光源较少时能够取得较好的效果。Xing等人[64]通过估计场景的几何结构、推测场景中物体的材质，检测太阳光的方向和强度等多个步骤，从一张室外图片恢复出场景的光照。在这个过程中需要用户手动标注一些辅助线，用来标识出一些简单的平面结构、几何材质、物体和其对应的阴影等，如图1.6展示了该方法中用户的标记和一些渲染结果。Karsch等人[65]的方法与之类似，它需要用户手动标出场景中的一些几何结构和若干个光源位置，随后场景的光照会通过一种基于渲染的优化方法计算出来。

这类方法与使用额外信息的方法比较类似，不过这种方法不需要额外的设备去单独拍摄辅助信息。在一些能够提供用户交互的应用场景下，获取用户的

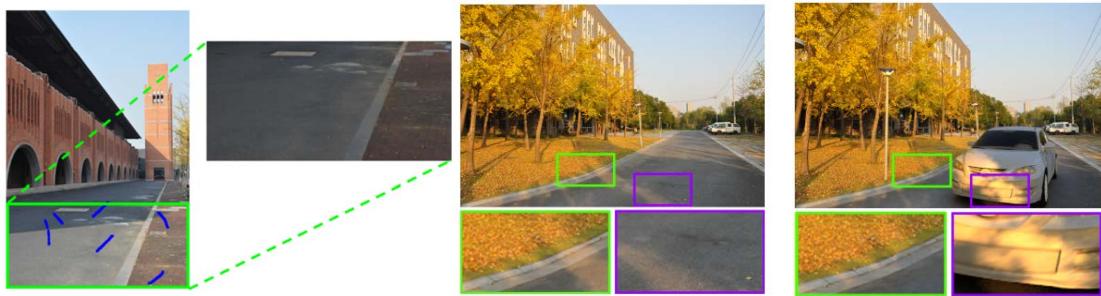


图 1.6 Xing 等人[64]的方法中，用户所标记的平面、阴影等参考线。以及最后的图片合成功效果

辅助信息比其它额外的信息更容易。但显然这种方法也只能应用在能提供用户交互的应用中。

### 1.2.7 基于图像分析的光照估计方法

经过前人多年的工作积累，一些研究者尝试探究仅使用单张图片来进行光照估计，即不借助任何额外的辅助信息、特定设备、几何信息等。Karsch等人[66]在2014年提出了这样的方法。他们提供了一个用户友好的图像编辑系统。用户将任意的3D虚拟物体拖放置场景后，系统会自动地对其进行渲染。用户后期还可以对其进行再编辑。他们的系统是完全自动的，通过一张图片，就可以恢复出综合的三维场景模型（几何、照明、漫反射反照率和相机参数）。为了实现该算法，他们首先从图片中估计出场景的结构信息、深度信息、漫反射反照率，相机参数等信息，然后在SUN360全景数据集[3]中查找外观与输入图像相似的全景图，并使用这些全景图中预先分类的光源作为输入图像的光源，获得了较为可靠的光照估计结果。Chen等人[67]在估计光照时另辟蹊径，对于给定的光照模型，提出了一种基于光线和语义的参数估计方法，该方法利用现有场景理解技术估计出粗糙的场景信息，接着使用语义约束的一组稀疏的小三维曲面进行本征图像分解，最后将得到的粗糙阴影图像视为所选小表面的辐照度。该方法不需要任何已知的三维几何结构、反射率，也不需要额外的采集设备，仅使用一张图片作为输入就可以获得较好的结果。

这类方法致力于不使用额外信息，直接从图片中估计场景光照，是解决光照估计问题时的积极探索。但很显然，这类算法非常复杂，难以做到实时，并且每一步都非常依赖上一步的可靠结果。

### 1.2.8 基于深度学习的光照估计方法

卷积神经网络（CNN）最早由Lecun等人[68]在1998年提出。随着计算机显

卡性能的提高和超大规模数据集（例如ImageNet[69]，ShapeNet[70]）的建立，深度学习在多个领域成为了一种强力的辅助工具。近年来，为了解决多种类型的问题，卷积神经网络结构层出不穷，例如AlexNet [71]，VGG [72]，ResNet [73]。它们在许多视觉问题中超越了传统方法，取得了非常好的成绩，例如物体检测[74]、图像分类[71]、图像分割[75]等等。卷积神经网络也被应用到了传统的图形学问题中，例如渲染降噪[76]、人脸模拟[77]等等，并且已经取得了很好的结果。

近期许多工作尝试使用深度学习解决光照估计问题。和传统方法类似，使用深度学习求解光照估计问题时也会借助一些辅助信息、探测物等。Calian等人[78]使用人脸作为光探测物，搭建了神经网络从照片恢复出室外场景的光照。Yi等人[79]搭建了一个用以提取高光和阴影反照率的神经网络，并使用该网络从人脸照片中恢复出室内外场景的光照信息。Georgoulis等人[80]使用真实反射图作为输入，用两个不同的CNN结构将图片分解为材质变量和光照变量。之后Geogoulis等人[81]又尝试利用深度学习从包含多种已知材质的物体图片中恢复出反射图和光照分布。Mandl等人[82]、Weber等人[17]也是借用了已知的物体几何恢复出场景光照。虽然这些光照方法和传统方法类似，也借助了一些额外的信息。但将传统方法中复杂的算法步骤替换为用深度学习求解后，往往能取得更好的结果，这得益于规模不断扩大、质量不断增加的训练数据。使用深度

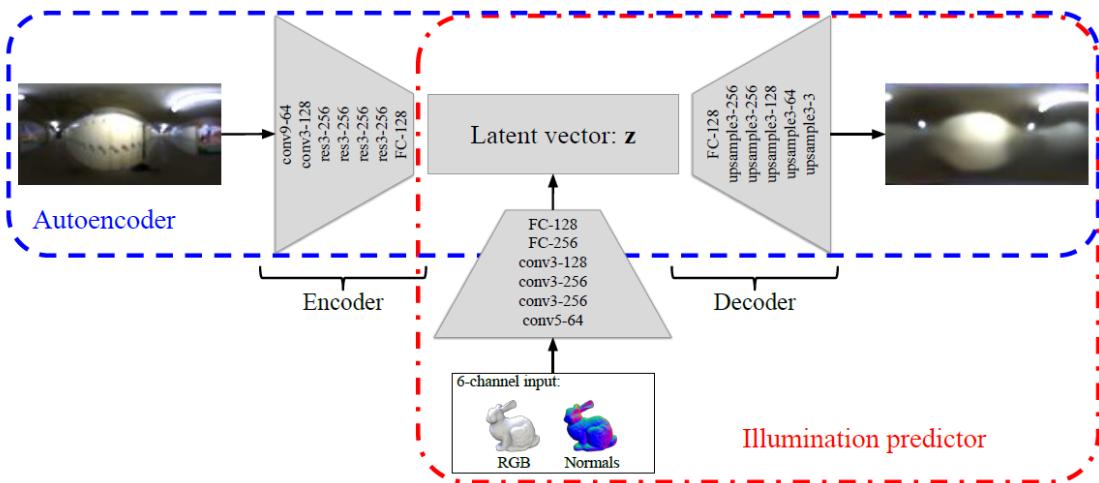


图 1.7 Weber[17]等人结合深度学习工具对光照分布进行编码，该图是其使用的自编码器结构。蓝色边框内为训练自编码器的结构，红色边框内则为训练光照估计的结构。

学习求解光照估计问题时，也会使用与传统方法中相同的光照表示模型。例如基于物理的Sun-sky模型[1]，球形谐波函数模型[82]等。深度学习在学习输入图

像的特征时非常有效，因此Weber等人[17]结合深度学习，使用自编码器（auto-encoder）对光照分布进行建模（如图1.7），使用卷积神经网络将场景编码为包含少量系数的隐变量，为光照估计问题中光照的表示提供了新的思路。

考虑到CNN强大的学习能力，最近一些方法尝试使用深度学习工具直接从单张图片中恢复整个场景的光照分布。Holdgeoffroy等人[1]搭建了一个深度卷积神经网络从室外图片中恢复出室外场景的参数化光照模型。Gardner等人[2]从室内图片中直接估计HDR全景图像。这两个方法是目前单图片光照估计工作中最先进的方法。由于HDR全景数据有限，这些方法通过一些光源探测算法将低动态范围全景图转化为粗糙的高动态范围全景图用于训练。

在深度学习任务中，训练神经网络需要大规模的高质量数据，但目前来说还没有这样的数据在规模和质量上同时满足要求。用于光照估计问题的数据集主要有两类，一类是规模较大的低动态范围全景数据集（例如SUN 360[3]），这类数据通常规模较大（数万张），但是其动态范围较低无法直接应用于光照估计，在训练光照估计神经网络时，需要使用特定的算法将其转换为粗糙的、低质量的高动态范围全景图；另一类是规模较小的高动态范围全景数据集（例如Laval Indoor[2]，这类数据虽然可以直接应用于光照估计，但是规模较小，难以保证训练的效果。

### 1.3 论文研究内容

将本文的主要工作内容包括：

- 深入调研全景图以及高动态范围全景图的获取步骤、投影方法、存储方式等，使用全景相机和曝光融合算法构建一个用于光照估计的大规模、高质量HDR全景数据集，并通过实验验证该数据集相对于其它数据集在光照估计问题中的优越性。
- 通过详细的实验分析了HDR全景数据集的规模和多样性对于光照估计问题的影响，研究数据的多样性和规模是否能够为基于深度学习的光照估计结果带来有效提升。
- 分析和验证使用视角相对的图片作为光照估计的输入的可行性。这两张图片可以由相机的前后置摄像头拍摄。使用前后摄像头同时拍摄两张照片不仅不会增加获取图片的步骤，还可以极大地降低光照估计的难度。这对于光照估计在智能设备中的应用来说有着很大的实际意义。
- 构建一个基于深度学习的光照估计网络模型，模型使用两张图片作为输

入，估计预测出场景的光照分布，随后通过实验与state-of-the-art方法进行多项对比。

- 对提出的光照估计深度学习模型和损失函数进行十分详尽的实验和分析，对于光照估计网络结构和训练过程中各个模块，设计大量对比实验深入探索使用不同网络结构的特点，以及它们对结果的影响，加深了对基于深度学习光照估计的理解。



## 第2章 光照估计数据集的构建与研究

### 2.1 引言

对于深度学习任务来说，训练数据的规模和质量对网络最终的表现有着直接的影响。小规模、低质量、不平衡的数据都可能会导致网络训练失败。从图片估计光照是一个非常复杂的问题，粗糙的数据往往很难训练出较好的预测网络。目前用于光照估计问题的数据集比较有限，主要包括大规模的低动态范围（LDR）全景图和小规模的高动态范围（HDR）全景图。这两类数据集都无法在规模和质量上同时满足训练一个鲁棒的光照估计网络的条件。因此本文将通过收集、拍摄、筛选等多个严格细致的步骤，构建一个包含多类场景的、多种光照条件、且具有一定规模的光照估计数据集。

本章主要介绍该数据集的构建方式以及在该数据集上的对比试验。首先对低动态范围全景图像和高动态范围全景图像做出介绍；然后阐述高动态范围全景图的拍摄与合成方法；接着在高动态范围全景数据集上进行多项对比实验，给出了数据集质量和规模对于光照估计问题的影响；最后将本文数据集与其它多种光照估计数据集进行对比，实验结果验证了本文数据在训练光照估计网络时相对于其它数据的优越性。

本章主要涉及的工作内容和创新点包括：

- 深入调研了全景图以及高动态范围全景图的获取步骤、投影方法、存储方式等，并使用全景相机和曝光融合算法构建了一个用于光照估计的大规模、高质量HDR全景数据集，而且通过实验验证了该数据集在光照估计问题中的优越性。
- 通过详细的实验分析了HDR全景数据集的规模和多样性对于光照估计问题的影响，证明了丰富的数据多样性和较大的数据规模能够为深度光照估计的表现有效的带来提升。

### 2.2 相关工作

使用深度学习估计场景光照是近几年光照估计问题的主要研究方向。在深度学习任务中，训练神经网络需要大规模的高质量数据，但目前来说还没有数据能在规模和质量上同时满足要求。现有的光照估计问题的数据集有两类，分别是规模较大的低动态范围全景数据集和规模较小的高动态范围全景数据集。

规模较大的低动态范围全景数据集以Xiao等人的SUN 360[3]为代表。该数据集由多种场景组成，包含了数万张低动态范围的全景图片，图片的分辨率为 $9104 \times 4552$ 。虽然这类数据集的规模较大，但是低动态范围全景图无法直接应用于光照估计，需要使用特定的算法将其转换为粗糙的、低质量的高动态范围全景图，这会对光照估计神经网络的训练造成一定的影响。

规模较小的高动态范围全景数据集以Gardner等人构建的Laval Indoor[2]为代表。该数据集包含了2100张分辨率为 $7768 \times 3884$ 的室内高动态范围全景图。拍摄该数据集时，他们将带有Sigma 8mm鱼眼镜头的Canon5D Mark III相机安装在自动的全景机械相机架，并编程拍摄7张不同曝光下的全景图片，这些全景照片以RAW模式拍摄，并使用商业软件PTGui Pro[83]自动拼接，最终融合为HDR全景图片。高动态范围全景图完全能够覆盖真实场景中各个光源的亮度，因此这些数据可以直接应用于光照估计问题中。不过这类数据集的规模通常较小，难以保证光照估计神经网络训练的效果。

## 2.3 全景图像

全景图（panorama）是一种广角图，可以以画作、照片、影片、三维模型的形式存在。全景图这个词最早由爱尔兰画家罗伯特·巴克提出，用以描述他创作的爱丁堡全景画。现代的全景图多指通过相机拍摄并在计算机上加工而成的图片[84]。全景图存储了以相机位置为中心的每个角度的颜色信息，颜色信息与普通图片类似，常用RGB三个通道分别存储。全景图根据其中的颜色数值范围，可分为低动态范围全景图和高动态范围全景图。

### 2.3.1 获取

拍摄全景图像的方式主要有两类。一类是使用专业的全景相机拍摄设备，这类设备大多由数个鱼眼形式的广角相机组成。在拍摄时，设备中的所有相机使用相同的参数同时拍摄，随后内置的固件或软件会对所拍摄到的图像进行投影变换、校正、拼接，形成一张全景图。另外一类是使用普通相机和相机旋转装置，对多个角度拍摄，随后手动使用拼接算法或软件将这些图像连接到一起。由于这种方式拍摄到的图片并不在同一个时刻，所以需要保证场景中不能包含过多的快速运动的物体。此外，大多数的现代智能手机都提供了手动拍摄“全景图”的方式。需要注意的是，由于手机相机的视角范围限制，以人体为轴旋转手机相机拍摄到的“全景图”多称为“宽景图”，在垂直方向上的视场角很难

达到 $180^\circ$ ，这就致使“宽景图”的顶部视角和底部视角区域会有很大缺失。

### 2.3.2 投影方式

全景图像的存储需要考虑投影方式和颜色的数值类型。在全景图像中，以相机为中心的视场可以被视为一个球体的表面，因此在存储和浏览全景图时，需要将全景球表面投影在二维表面中。常见的投影方式有等距投影，圆柱投影，球形投影，立方体贴图投影，立体投影等。

- **圆柱投影** (cylindrical projection) 该投影方式是将全景球置于其外切圆柱内，并由球心向圆柱面做投影，随后将圆柱内表面横向展开后的图像即为球形全景图的圆柱投影。这种投影下的全景图在两极会发生无限的纵向拉伸，因此圆柱投影后的图像无法包含靠近两极的信息，也即这种投影方式无法表示垂直视角为 $180^\circ$ 的全景图。柱面投影是传统摆动镜头全景胶片相机所提供的标准投影方式。相对于全角度的全景图，该投影方式更适合在垂直视角小于 $120^\circ$ 的宽景图，常用于现代智能手机的全景图预览。

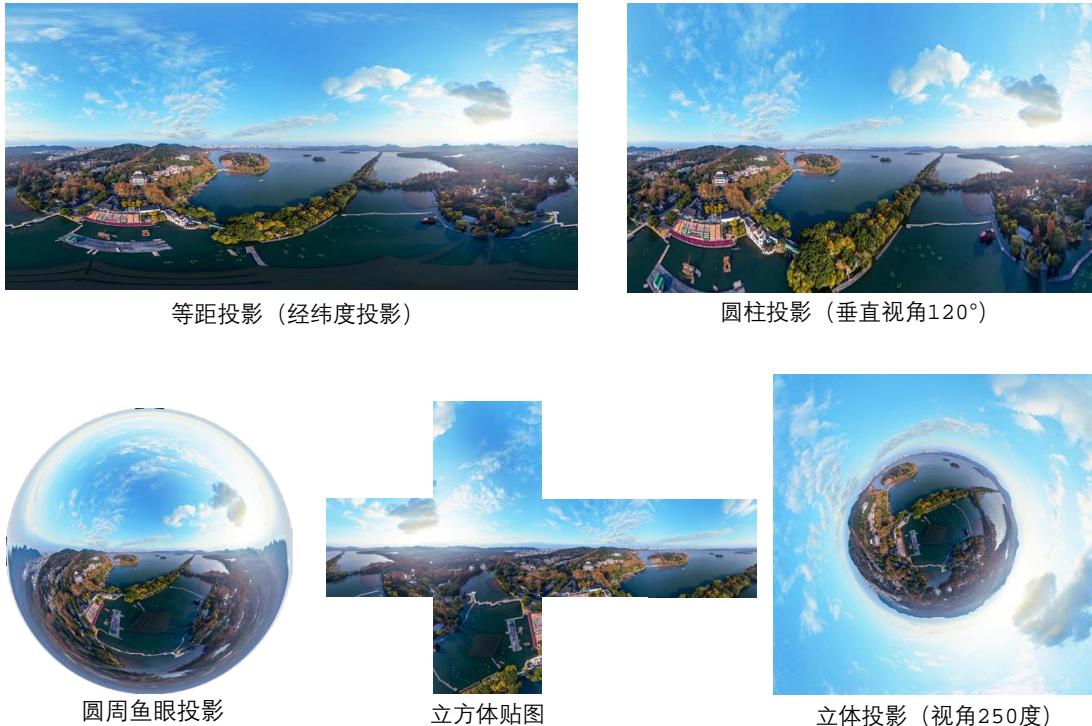
- **等距投影** (equirectangular projection)，也称等距圆柱投影。该投影方式是将球面的经度和纬度坐标线性变换为图像空间的横纵坐标。经过投影处理后的全景图像是一幅宽高比例为2:1的图片。这种投影的特点是越接近两级，图像的变形就越严重。投影后的全景图在预览时一目了然，而且这种投影方式较为简单，是存储和预览全景图最常见的方式之一。

- **圆周鱼眼投影** (circumferential fisheye projection)，也称圆形投影或镜面球投影，是角投影的一种，投影图像看起来像理想圆形鱼眼镜头拍摄的图片。这种投影方式可以覆盖 $360^\circ$ 的视角，但在球面边缘的像素却被极端地扭曲和变形，导致一部分精度损失。这种投影方式常见于全景相机中，全景相机一般由两个朝向相反的 $180^\circ$ 视角鱼眼镜头构成，每个镜头所拍摄到的图片均为一个垂直和水平视角各 $180^\circ$ 的圆周鱼眼投影视图。

- **立方体贴图投影** (cubemap projection) 在计算机图形学中，立方体投影是常用的环境映射方法之一，常用于游戏场景中的天空盒，相当于等距柱状投影的优化版，环境投影到立方体之后可分六个正方形纹理来存储；或者将立方体展开，存储于一个纹理中的六个区域内。在全景图像及视频中，立方体投影将球形视频映射到它的外接立方体上，立方体的上下两个面分别对应两极区域，中间的四个面对应赤道区域。

- **立体投影** (stereographic projection)，这种投影方式常见于一些全景图应用中，由于预览效果类似一颗行星表面，因此这种投影也被称为小行星投影。

除此之外，还有墨卡托投影、正弦投影等、直线投影等应用于不同的领域投影方式。图2.1展示了几种常见的全景图投影方式。



**图 2.1** 全景图的几种投影方式。等距投影是常见的全景图投影方式，包括了所有视角的信息。圆柱投影的特点致使它不可能有 $180^\circ$ 的垂直视角；圆周鱼眼投影可以显示出所有视角的信息，但是从中可以看到，靠近圆形边缘的区域被极度扭曲，这通常会导致精度的损失；立方体贴图是计算机图形学中常用的投影方式；立体投影方式看起来像一颗星球，因此也被成为小行星投影方式，这种方式也无法展示所有视角的信息。

### 2.3.3 动态范围



**图 2.2** 两种不同曝光条件下的LDR图像。在欠曝图像中，黄色框选区域中大部分像素均为黑色，但它们的明暗程度并不同；类似地，在过曝图像中，红色框选区域中像素值均为白色，但实际上太阳所处像素的亮度远远高于周围像素。使用LDR图像难以解决这些问题。

图像的动态范围（dynamic Range）是指一个图像中最亮和最暗部分之间的相对比值[84]。根据动态范围的大小可以将图像分为低动态范围（或称为标准动

态范围)图像和高动态范围图像。传统的8位图像将颜色值存储为[0, 255]范围内的整数, 低动态范围图像中的颜色只能在这256个数中取值。图2.2展示了两种不同曝光条件下的照片, 可以看出每幅图片都有一定的过曝和欠曝区域。例如在过曝图片中, 太阳和其周围区域的颜色均为白色, 但实际上太阳的亮度要远远高过天空的亮度; 同样的, 在欠曝区域中, 虽然大部分区域同为黑色, 但实际上这些区域的明暗程度也是千差万别的。



图 2.3 调整LDR和HDR图像曝光的结果。第一行是从一幅HDR开始, 不断减小场景曝光值; 第二行则是对LDR执行同样的操作。可以看出对于HDR图像, 多次降低曝光后太阳位置的亮度依然很亮, 这说明HDR图像中太阳对应的像素值远远高于周边像素, 与真实场景的光照情况一致。而LDR图像在降低曝光后, 太阳位置的像素值会和周围像素一起降低, 这显然是不符合真实光照条件的。

相比低动态范围的图像, 高动态范围图像可以提供更大更广的动态范围。图2.3展示了调整LDR和HDR图像曝光值时的不同点。高动态范围全景图是具有高动态范围的全景图像。它的动态范围可以高达 $2^{32}$ , 而人类的眼睛所能看到的范围是 $10^5$ 左右[84]。因此高动态范围全景图像可以作为真实的光照信息, 这是低动态范围全景图不能比拟的。

图2.4展示了两组使用低动态范围全景图和高动态范围全景图的渲染结果, 可以看出, 使用低动态范围全景图渲染的结果颜色值对比平缓, 视觉效果不够真实。而使用高动态范围全景图作为光照时, 渲染结果包含了足够的对比度和锐利的阴影。图2.4中的第二行使用包含更亮光源的全景图渲染, HDR和LDR渲染结果之间的差异更加明显。

### 2.3.4 HDR的获取与存储

HDR全景图的获取方式与普通全景图类似，不过由于HDR包含了更大的动态范围，普通相机仅通过单次拍摄是无法满足这个要求的，因此需要对每个视角拍摄不同曝光条件下的图像。HDR全景图的获取分为多曝光拍摄，多视角拍摄，全景图拼接，清洗和筛选，曝光融合五个步骤。

- **多曝光拍摄。** HDR图像通常无法直接由拍摄设备直接获取，因此需要利用不同的曝光值对同一场景多次拍摄。曝光值通常由相机的ISO、快门速度、光圈大小共同决定。

- **多视角拍摄。** 单个相机的视角有限，在合成全景图之前，需要对场景的不同视角拍摄。为了保证拼接后的图片质量，拍摄时相机的镜头不能有太大的平移。因此拍摄全景图时，相机位置多由精密的机械装置自动控制。此外需要注意的是，所有拍摄视角在拍摄时的曝光值序列和其它相机参数（例如白平衡，相机视场范围等）要完全一致，以保证在全景图拼接时的各个视角图片的一致性。

- **全景图拼接。** 使用普通相机对多个视角进行拍摄后，需要将这些图片拼接到一起。这通常需要做图片间的特征匹配等等，目前全景图的拼接有较为成熟的算法和工具，这里不再展开叙述。此外，目前市面上出现了一种专门用于拍摄全景图的全景相机，其中有些全景相机支持通过调节ISO和快门速度拍摄多种曝光条件下的全景图，例如小米全景相机[85]等。使用这种相机可以省去图像拼接的步骤，每次拍摄只需要关注相机的曝光即可，这样可以避免拼接时造成的瑕疵结果。图2.5展示了使用全景相机和自动机械装置拍摄的两组图。

- **清洗和筛选。** 在进行拼接和融合之前，需要过滤掉质量较差的中间结果，包括场景不一致的全景图（通常由移动物体在不同时刻被拍摄导致）、色差过大的图片（通常由错误的白平衡导致）、噪点过多的图片（经常由较暗场景中使用过大的ISO导致）等等。这些都需要人工进行观察、筛选、清理和重拍，以保证最终合成的HDR全景图的质量。

- **曝光HDR融合。** 当获得多张不同曝光的全景图后需要将他们融合为一张HDR图，常见的融合算法有基于信息熵的算法、基于双边滤波的算法、基于亮度梯度大小的算法、基于拉普拉斯金字塔的算法等。曝光融合工具中，PTGUI[83]能够提供很好的融合结果。

曝光融合后的全景图就是一幅HDR全景图，这种全景图包含了很高的动态范围，可以作为场景真实的光照表示。HDR全景图像的投影方式和普通全景图完全

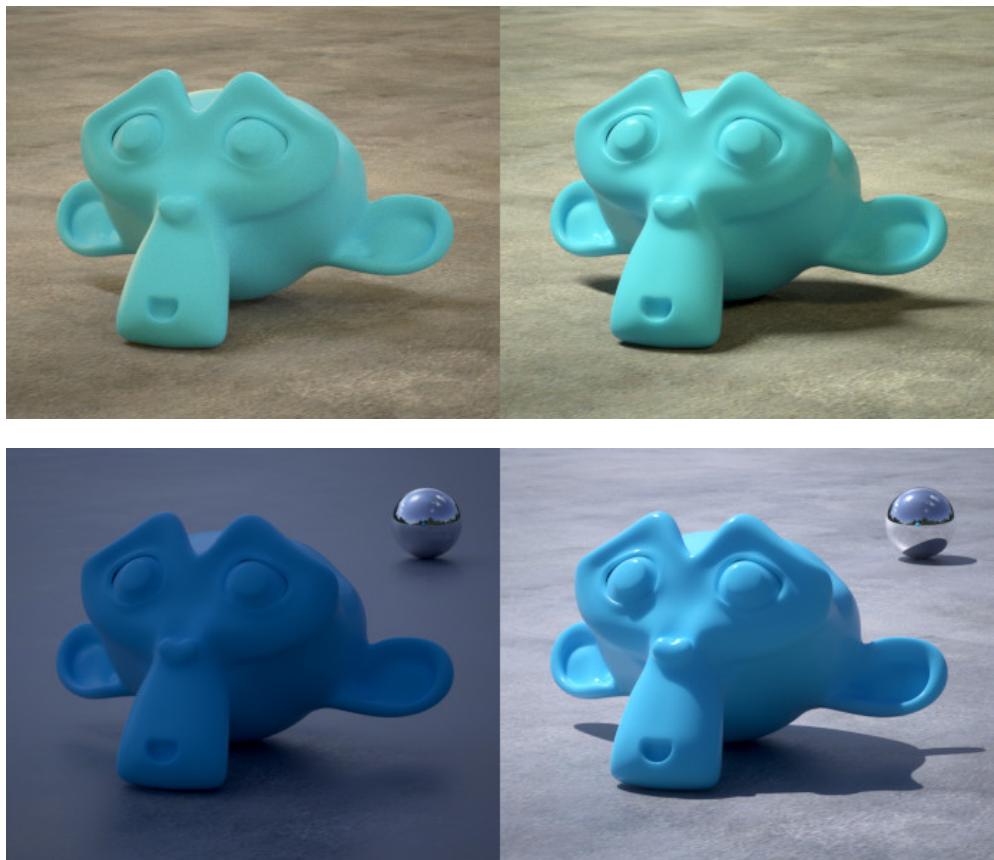


图 2.4 使用**LDR**全景图（左）与使用**HDR**全景图（右）渲染结果的对比，引自[86]。从第一行可以看出，使用低动态范围全景图像渲染的结果，在颜色上对比平缓，看起来很不真实。而使用**HDR**全景图渲染的结果提供了足够的对比度和锐利的阴影，更具有真实感。这种情况在场景中有较强光线时尤为明显（第二行结果）



图 2.5 常见的拍摄全景图的方式。左图是将单反相机固定在能够自动控制旋转的机械架上，对多个方向进行拍摄，最后利用电脑软件将其拼接到一起；右图是使用全景相机拍摄，由于鱼眼镜头和内置硬件的帮助，这种拍摄方式十分简便，只需一次拍摄就可以获得整个全景图。

一致，它们之间的区别只是每个像素位置的数值类型和数值范围不同，常见的HDR图像保存格式有TIFF, HDR, RGBE, EXR等。

## 2.4 构建HDR全景数据集

表 2.1 小米全景相机[85]的详细参数，本文在构造HDR数据集时使用了该相机。

项目	参数
图像传感器	Sony IMX 206
CMOS 尺寸	1/2.3
镜头类型	超大广角折反式镜头模组
单镜头构造	5球面玻璃镜片+2模造非球面玻璃镜片+2片直角玻璃棱镜
视野范围FOV	2 × 190°
光圈与焦距	F2.0, 1.43
ISO范围	自动；支持手动：50、100、200、400、800、1600
曝光补偿	−3 +3, 1/2EV 阶
快门速度：	自动，支持手动：1/6400 ~ 32s
白平衡	自动、户外、阴天、白炽灯、荧光灯
支持存储格式	jpeg
最大分辨率	6912 × 3456

本节主要介绍本文所提HDR全景数据集的构建方法与步骤。构建时需要考虑数据的多样性，在保证数据质量的前提下尽量增加数据规模。

### 2.4.1 场景选择

HDR全景数据集需要考虑到场景的多样性，为此，本文选取了多个场景进行拍摄，包含室内、室外、森林、公园、公寓、小区、建筑群等多种常见场景，晴天、阴天、多云等多种气象条件，清晨、中午、下午、傍晚、夜间等多种拍摄时间，以及春夏秋冬多个拍摄季节。

### 2.4.2 拍摄设备

拍摄HDR全景图的方式有两种，一种是使用普通相机多次拍摄并进行拼接，另一种是直接使用全景相机拍摄。本文采用的方式是后者，即直接使用全景相机拍摄，这样可以避免大量的拼接操作，只需要根据场景调整不同的曝光范

围即可。本文所使用的相机是小米全景相机[85]，该相机可以通过调节快门速度拍摄不同曝光条件下的全景图，该相机的有关参数如表2.1所示。

### 2.4.3 曝光融合工具

在进行曝光融合时，通过多种曝光融合工具的对比，发现PTGUI[83]能够提供很好的融合结果，因此本文在构建该光照估计数据集时主要使用此工具进行曝光融合，融合时的参数如表格2.2所示。

表 2.2 本文进行曝光融合时的硬件设施与使用参数。

项目	参数
硬件平台	处理器Intel 8700，内存16G
操作系统	Windows 10 64位 1809
合成工具	PTGUI [83] 10.0.13
相机响应曲线	拍摄时获得的响应曲线
视觉曝光优化	关闭
插值器	Enblend
插值算法	Lanczos16 (sinc 1024)

通过以上步骤，拍摄了约5000张全景图像，随后经过清理和筛选，共合成了约550组HDR全景图。图2.6展示了部分HDR全景图。由于在拍摄时对拍摄装置的关注，该数据集的HDR全景图片底部没有出现像Laval Indoor数据集[2]中的底部黑色块(图2.9)。该数据集包含了多种场景和光照条件，而且图片质量较高，几乎没有噪点和拼接痕迹。此外，作为工作之一，本文额外从网络上抓取、收集了约500张更高质量的HDR全景数据（均遵循相关的许可文件）。这些数据通常由更专业的单反相机和精密的机械装置拍摄，质量因此相对较高。至此，本文构建的数据集由近千张HDR全景图构成，是目前在光照估计领域，能包含室内外场景的数据集中，规模最大的HDR全景数据集。

## 2.5 探究数据集对光照估计的影响

大规模、多样化的HDR全景数据对光照估计效果的提升需要通过详细严格的实验来验证。本文设计了一系列的对比实验来验证不同的数据规模、数据多样性，以及不同数据集在光照估计网络上的表现。在这些实验中，使用的是光照估计网络由多层卷积层和全连接层组成。网络的输入是单张图片，目标输出

是光照的球形谐波近似系数。对比实验主要分为三个部分，首先是对比不同的数据规模对光照估计网络的影响，其次是对比数据的多样性对光照估计网络的影响，最后是本文构建的数据集和部分已有数据集的对比。



**图 2.6** 数据集中的HDR全景图预览，本文构建的数据集包含了室内、室外、清晨、傍晚、黄昏，公园、广场、建筑群等多个场景。

### 2.5.1 数据准备

使用监督学习方法训练网络时，需要成对的输入和真实目标输出。在进行HDR数据集上的对比实验时，使用的输入是普通的图片，目标输出是光照的球形谐波近似表示。

**图片输入。**输入的图片从HDR全景图中提取。首先将HDR全景图映射到一个球形表面，并将视点置于球心。随后随机选取一个视角方向，经过提取、颜色映射、伽马校正后，获取到对应的视角图片。在Gardenr[2]的工作中，为了避免提取到底部的黑色色块，他们对HDR全景图像的中间部分进行了垂直的增大缩放变形。由于本文数据集中没有类似的黑色色块，因此不需要执行该操作。

**目标输出。**在本节的对比实验中，目标输出是光照的球形谐波表示。使用球形谐波函数，场景光照可以近似地由若干个SH系数表示。这种表示方式能够

保留光照中大部分的低频信息和小部分的高频信息。一般来说，9组或16组三通道的SH系数都可以很好地近似场景的光照。考虑到深度学习强大的学习能力，本节中所有的实验均使用了16组3通道的SH系数作为目标输出信息。

表 2.3 从HDR全景图中提取普通图像的参数列表

项目	配置
视角范围	$\theta$ 在 $[\frac{1}{3}\pi, \frac{2}{3}\pi]$ 上随机均匀取值, $\phi$ 在 $[0, 2\pi]$ 上均匀取值
曝光范围	在 $[-1.5, 1.5]$ 上均匀随机取值
伽马校正系数	固定值, 2.2
视角角度FOV	水平和垂直方向 $70^\circ$
图片宽高尺寸	480×360, 在训练时会resize到224×224

对于1000张HDR全景数据中的每一幅全景图，依照表格2.3所列的参数随机选取128个视角方向。然后在每个视角下提取图片和对应的SH系数，作为一组输入输出数据。图片提取方式和球形谐波函数在章节3中会有更加详细的介绍。最终，共有约12万（ $1000 \times 128$ ）组数据用于光照估计的训练和测试。

## 2.5.2 数据划分

在深度学习任务中，数据集一般会被划分为训练集，测试集和验证集。训练集用于网络的训练，测试集上的表现用于指导调整网络的结构和超参数，验证集的作用是验证网络的最终表现。还有一些划分方式是直接将数据集划分为训练和测试集，测试集兼具验证集的功能。本文在数据集上的所有实验均使用第一种划分方式。本节进行了三种对比实验：不同数据规模上的实验，不同数据多样性上的实验，不同数据集上的实验。下面分别给出三组实验内的详细的数据划分方法。

- **数据规模实验。**该实验意图验证分析不同的数据规模对于光照估计的影响。在该实验中，将HDR全景数据集按照90%/5%/5%的比例划分为训练集、测试集和验证集，即随机选取900/50/50张HDR全景图分别作为训练集/测试集/验证集。在对比数据规模对光照估计的影响时，分别从训练集内随机选取100张到900张HDR全景图（每隔100张），对于每张全景图，提取128组图片和SH系数对用作训练，并均在相同的验证集上测试。

- **数据多样性实验。**该实验意图验证分析数据的多样性对光照估计的影响。与数据规模实验中的划分方案类似，在HDR全景数据集中随机选取900/50/50张

分别作为训练集/测试集/验证集。在训练集中选取160张用于训练光照估计网络，在对比时，选用不同的场景数量（训练集总数一致），构成具有不同多样性的训练集进行对比。

- **不同数据集实验。**意图验证本文所构建数据集在光照估计问题中的优越性。该实验对比了SUN360[3]数据集和Laval Indoor[2]两种用于光照估计的数据集。在每个实验中，数据集中的HDR图像将全部加入到训练集中。为了公平的对比，验证集将不再从本文数据集中选取，而是直接使用真实拍摄的图片和对应的场景光照。这部分真实的光照由100幅图片和10个场景构成。

需要注意的是，为了保证训练集和验证集之间没有重复数据数据，数据的划分是在HDR全景数据集上进行的，之后的图片也是在划分后的HDR数据集上提取，这样就可以保证同一幅HDR全景图不会同时出现在不同的数据集中。

### 2.5.3 网络结构与实现细节

从单张图片预测出所在场景光照对应的SH系数是一种回归问题，因此可以通过卷积层提取图像特征，使用全连接层回归SH系数。图2.7展示了用于对比实验的网络结构，该预测网络是在Resnet-50[73]的基础上修改的，该网络通过将Resnet-50最后的pooling层替换为5个带有批归一化（BN）[87]和Leaky RELU激活函数[88]的全连接层（最后一层不包含激活函数），最终输出48个表示场景光照分布的球形谐波系数(16组×3通道)。预测的48个SH系数可以被视为一维向量，

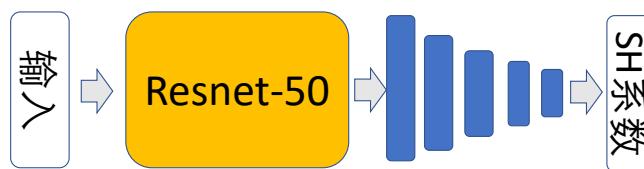


图 2.7 用于评估数据集的卷积神经网络结构。该网络以图片作为输入，输出用来近似光照的SH系数。网络结构以Resnet-50[73]为基础，将最后一层pooling层替换为5个全连接层。除最后一层外，每层之后都有BN和RELU激活函数。

向量之间的误差常用L2距离衡量。因此本节实验中使用真实SH系数和估计SH系数之间的均方误差（MSE）作为损失函数。训练使用RMS优化器[89]，初始学习率为0.0005，衰减系数为0.9，衰减频次为每一个epoch衰减一次。每种实验均在NVIDIA GTX-1080上训练10万步，每步的batch size为16。

### 2.5.4 实验：数据的规模对光照估计的影响

本实验意图分析不同的数据规模对于光照估计的影响。在划分出验证集后，

表 2.4 使用不同数量的**HDR**全景图训练光照估计网络。可以看出，在数据规模较小时，增加训练数据量能够为光照估计带来巨大的提升，但当数据达到一定的规模时，单纯增加训练数据量并不能起到很好的效果。

训练数	100张	200张	300张	400张	500张
RMSE	0.4502	0.3108	0.2161	0.1892	0.1774
训练数	600张	700张	800张	900张	
RMSE	0.1683	0.1633	0.1621	0.1614	

分别使用100张到900张数据用于训练。在测试时，将验证集中的图片输入到网络中，然后使用估计出的光照渲染一些3D物体，之后将这些结果与使用真实光照渲染的结果进行对比，计算他们之间的均方根误差(RMSE)。它们在验证集的表现如表2.4和图2.8所示。可以看出当数据规模较小时，增加数据对于光照估计效果非常明显，但是当数据规模达到一定的数量时，增加数据所带来的光照估计的提升变得有限。

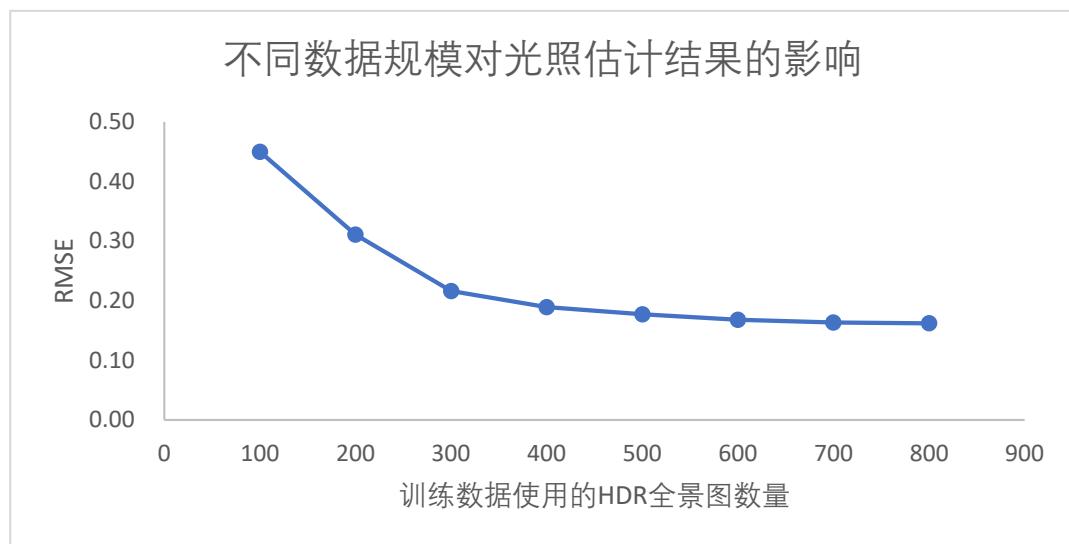


图 2.8 数据规模对光照估计问题影响的趋势图。可以看出，在数据规模较小时，增加训练数据量能够为光照估计带来巨大的提升，但当数据达到一定的规模时，这种提升达到了饱和，单纯增加训练数据量并不能起到很好的效果，这主要是因为这些场景中包含了一定的重复场景，而在数据中添加更多的重复场景对训练几乎没有意义的。

### 2.5.5 实验：数据多样性对光照估计的影响

本实验意图分析数据多样性对于光照估计的影响。在划分出验证集后，分

**表 2.5** 使用具有不同多样性的数据训练光照估计网络。每一项实验中使用了不同的场景数，但使用的全景图总数是相同的**160**张。可以看出，当场景数较少时，光照估计的误差很大。随着场景种类的增多，即使数据规模没有增加，光照估计的效果依然在提升，并且没有出现像增加数据规模一样的饱和现象。

场景数	8	16	32	48	64
RMSE	0.5764	0.5277	0.4507	0.3812	0.3117

别使用覆盖不同类型的HDR数据训练光照估计网络。它们在验证集的表现如表2.5所示。可以看出增加数据多样性对于光照估计的提升非常明显。在本实验中，用以训练光照估计网络所使用的数据量完全相同，均为160张。他们之间的区别仅仅是数据中所包含的场景种类数不同。从表2.5中可以看到，当场景数较少时，光照估计的误差很大。随着场景种类的增多，即使数据规模没有增加，光照估计的效果依然在提升，并且没有出现像增加数据规模一样的饱和现象。这说明丰富数据多样性对于提升光照估计的效果比单纯增加数据规模明显。这为以后数据集的扩充提供了指导和依据。

### 2.5.6 实验：与其它数据集的对比

**表 2.6** 该实验对比了使用不同的数据集训练光照估计网络时，它们在真实数据集上的表现，从数值结果上可以看出，使用本文构建的数据集在训练光照估计网络时，都优于其它网络。

数据集	本文数据	拓展后的SUN360数据	Laval Indoor
RMSE	0.1883	0.2437	0.2014

本实验对比了本文数据集与其它光照估计数据集在光照估计问题中的表现。选取的对比数据集分别是 SUN360[3]和Laval Indoor[2]数据集。这两个数据集是基于深度学习的光照估计方法中最常用的两个数据集，目前多个最先进的（state-of-the-art， SOTA）光照估计方法均是在这两个数据集上进行的训练。其中， SUN360数据集是一个大规模的低动态范围全景图像，为了能将其应用到光照估计问题中， Hold-Geoffroy等人[1]和Gardner等人[2]提出了一种用于低动态范围全景数据集上的光照探测方法，将低动态范围全景图像拓展为高动态范围全景图像。本节实验应用了Hold-Geoffroy的方法将SUN360拓展为HDR数据集。Laval Indoor数据集包含了约2000张室内场景的光照，不过其中有大部分的

类似的场景（例如同一个场景的多次拍摄），全景图的底部也有大面积的黑色色块（图2.9），这些都是会对光照估计问题产生影响的因素。表2.6展示了使用不同数据集训练光照估计网络在真实数据集上的表现，图2.10展示了一些可视结果。本文构建的数据集在训练光照估计网络时，数值表现和可视结果均显著优于其它网络。另外，使用SUN360数据训练的光照估计模型在较暗场景中渲染结果与真实值较为接近，但明亮的场景中预测结果与真实值却相差甚远。这也反映了低动态范围全景图本身的局限性，虽然在通过一些光源探测技术将其拓展为HDR全景图，但这种HDR全景图却很粗糙，限制了其在光照估计问题中的作用。



图 2.9 Laval Indoor[2]数据集中的一幅全景图片。由于该数据集在拍摄时将三脚架的部分暴露在了镜头中，他们在公布数据时选择将这些信息用黑色色块进行了遮挡。虽然经过坐标转换后这些黑色色块的面积没有图中显示的这么大，但依然有可能对光照估计问题造成影响。



图 2.10 使用本文数据集训练的光照估计网络与使用其它数据在可视化结果上的对比。使用本文数据集训练的网络，在室内外场景的结果都明显优于其它两种数据集。另外从结果中可以看出，使用**LAVAL**数据集的结果在室内场景的表现中较为理想，在室外场景却得到了很差的结果，这是由于**laval indoor**数据集本身的局限性。使用**SUN360**数据训练的光照估计网络，在较暗场景中的渲染结果与真实值较为接近（第二列），但明亮的场景中预测结果与真实值却相差甚远（第三、四列）。这也反映了低动态范围全景图本身的局限性，虽然在使用一定的光源探测技术将其拓展为**HDR**全景图，但这种**HDR**全景图却很粗糙，限制了其在真实场景光照中的表现。

## 2.6 总结与讨论

本章介绍了全景图像的两种获取与拍摄方式，以及存储和预览全景图需要的多种投影方法。根据全景图中颜色的动态范围，可以将全景图分为低动态范围（LDR）全景图和高动态范围（HDR）全景图。LDR全景图与HDR全景图在投影方式、视角范围等并无二致，但在能表示的亮度动态范围上却相差甚远。HDR全景图能够支持大于人类眼睛的动态范围，这使得HDR全景图能够作为真实场景光照的一种表示。

HDR全景图的获取通常需要多视角拍摄、多曝光拍摄、全景图拼接、筛选清理、曝光融合五大步骤，由于本文使用一类可以调节曝光的全景相机，因此省去了多视角拍摄和拼接的步骤。通过拍摄不同光照条件下的多个场景，本文构建了一个包含约550张HDR全景图的高质量数据集，同时额外地从网络上收集了约500张HDR全景图片，这近千张HDR全景图构成了一个大规模、高质量、具有丰富数据多样性的HDR全景数据集。该数据集不仅可以用于光照估计，也可以用于与光照估计相关的其它多种问题。

大规模、多样化的HDR全景数据对光照估计效果的提升需要通过详细严格的实验来验证。本文设计了一系列的对比实验来分析数据规模、数据多样性，以及数据集类型在训练光照估计网络时的表现。在这些实验中，光照估计网络是在Resnet-50的基础上修改的，网络使用单张图片作为输入，预测出光照的球形谐波近似系数。本文设计了三组实验进行对比，分别是数据规模对光照估计网络的影响实验，数据的多样性对光照估计网络的影响实验，以及与其它相关数据集的对比实验。实验结果表明，数据的多样性对于深度学习光照估计方法预测的结果提升很大。在数据规模较小时，增加数据也能够明显的提高光照估计网络的表现。与其它光照估计数据集的对比实验表明，使用本文所构建的数据集训练的光照估计网络有着更好的效果，验证了该数据集在光照估计问题中的优越性。

虽然本文所构建的全景数据集达到了数千张，使用该数据集训练光照估计网络已经能够取得较好的效果，但是该数据集仍然有一定的改进空间，例如可以拍摄更多场景下的数据、使用曝光范围更大的全景相机等。此外，从HDR全景图的获取过程可以看出，HDR全景图的拍摄需要对同一场景多次拍摄，因此快速变动的场景难以被拍摄为HDR全景图。不过对于光照估计问题来说，该数据集的多样性已经足够保证较好的光照估计结果，相信随着以后各种新技术的提出，动态场景的HDR全景图也能够被轻松获取。



## 第3章 基于深度学习的光照估计方法

### 3.1 引言

从有限的图像信息估计出整个场景的光照分布是一个复杂的问题。首先，图像的视野范围比较有限，例如一张视场角（FOV）为 $60^{\circ}$ 的照片所拍摄到的区域，在其对应的全景图中占比不足6%。此外，一幅图片是光照分布、场景几何结构、物体材质、相机参数等多个单位之间的复杂交互结果（公式 3.1）。

$$\text{Image} = \text{ComplexInteraction}(\text{Light}, \text{Geometry}, \text{Material}, \text{Camera}) \quad (3.1)$$

通过公式3.1可以看出，在其它三个信息未知的情况下，从图像（Image）反推出光照（Light）是一个严重的不适定（ill-posed）问题。不仅如此，在不同的条件下拍摄的彩色图像可能存在颜色或几何偏差。例如相机畸变、不正确的白平衡、过曝光、欠曝光等。这些都会对光照估计造成一定程度的干扰，增加光照估计的难度。

为了降低问题的难度，研究者们尝试对该问题进行约束或简化。在利用传统方法估计光照时，简化的方式通常是增加输入的信息或者减少要估计的光照模型规模。一部分工作借助额外的输入信息辅助估计场景光照，例如深度信息[36–39]、几何信息[40–42]、多张图片[47, 48, 50]、先验知识[20, 45, 46]、用户标记[46, 65]等等。另一部分工作通过使用低维的光照表示模型来简化光照估计问题，例如使用球形谐波函数（SH）来拟合场景光照[25, 41, 45, 56–58, 90]、使用小波函数近似场景光照[59]、使用有限的点光源的集合近似光照[40, 42–44, 47]、使用基于物理的室外光照模型[60–63]等。

近年来，一些研究者尝试将深度学习方法应用在光照估计问题当中。Hold-Geoffroy等人[1]搭建了一个深度卷积神经网络，意图从室外图片中恢复出场景的参数化光照模型。Gardner等人[2]直接使用单张室内图片估计HDR全景图像。这两个方法是目前单图片光照估计工作中最先进的方法。不过现有的深度学习方法也有一定的局限性。训练一个鲁棒的神经网络往往需要大量的数据，而目前用于光照估计问题的数据集比较有限，主要包括大规模的低动态范围全景数据集（SUN360[3]等）和中小规模特定场景的高动态范围全景数据集（Laval Indoor等[2]）。这些数据集在规模和质量上很难同时到达训练深度神经网络的要求。

在这样的背景下，本文在光照估计的两个方向上开展研究。其一是构建一个具有一定规模和质量光照估计的数据集。这样的数据集不仅能被用来训练更加鲁棒的光照估计网络，也可以被应用到其它多种相关的深度学习问题当中，上一章节已经对这一部分工作进行了详细的介绍。其二是在已有数据集和本文构建的数据集基础上，深入探索基于深度学习的光照估计方法，对其中的网络结构，网络参数，损失函数，光照表示等多个模块进行细致的对比和研究。这是本章节工作的主要目标。

本章主要所涉及的工作内容和创新贡献主要有以下几点：

- 首次提出使用前后两张相对视角的图片作为光照估计的输入。两幅图片多由相机的前后置摄像头拍摄，这不仅不会增加获取图片的步骤，还可以极大地降低光照估计的难度。现代移动设备几乎都至少包含前后两个摄像头，因此本文所提方法对于智能手机应用来说有着很大的实际意义。
- 构建了一个基于深度学习的光照估计网络模型，该网络模型使用两张图片作为输入，估计预测出场景光照对应的球形谐波系数，该模型在光照估计问题中非常有效，目前已经取得了超过state-of-the-art的结果。
- 提出了一个新的损失函数 - Render Loss，该损失函数巧妙地利用了SH的特性，将部分渲染过程置于神经网络中，这样可以在训练神经网络时，添加渲染结果上的监督信息。Render Loss的算法简洁但非常有效，极大地提高了光照估计的表现。
- 对提出的光照估计深度学习模型和损失函数进行十分详尽的实验和分析，对于光照估计网络结构和训练过程中的各个模块，通过几十组对比实验深入探索了这些模块对光照估计结果的影响，促进和加深了对基于深度学习光照估计问题的理解。

## 3.2 相关工作

### 3.2.1 深度学习

卷积神经网络（CNN）最早由Lecun等人[68]在1998年提出。随着计算机显卡性能的提高和超大规模数据集（例如ImageNet[69]，ShapeNet[70]）的建立，深度学习在多个领域成为了一个强力的工具。近年来，为了解决各种类型的问题，卷积神经网络结构层出不穷，例如AlexNet [71], VGG [72], ResNet [73]。它们在许多视觉问题中超越了传统方法中取得的非常好的成绩，例如物体检测[74]、图像分类[71]、图像分割[75]等等。近期，卷积神经网络也被应用到了传统的图

形学问题中，例如渲染降噪[76]、人脸模拟[77]等等，都取得了很好的结果。

### 3.2.2 传统光照估计方法

传统方法估计光照时，常常通过增加输入信息或者简化光照模型来降低求解难度。Sato等人[47]、Nishino[48]等人、Yu[50]等人使用多张图片作为输入。Knecht[36]等人、Meilland[37]等人、Zhang等人[39]、Barron和Mailk[38]等使用额外的深度信息估计光照。还有一些工作使用已知的几何信息辅助估计光照[40–42]。另外一部分工作通过使用低维的光照表示模型来简化光照估计问题，例如使用球形谐波函数（SH）来拟合场景光照[25, 41, 45, 56–58, 90]。通过使用SH，场景光照的低频部分可以使用少量的系数（通常为9-16组，约27-48个）来近似，这极大地减少了光照估计的难度，与之类似，Barronhe和Malik[59]在光照估计问题中使用小波函数来近似场景光照。还有一些光照估计工作[40, 42–44, 47]将光照分布简化为若干个点光源的集合，进而将光照分布估计问题转化为预测光源数量、位置和大小的问题。对于室外场景中的光照估计问题，使用基于物理的室外光照模型[60–63]可以使用更少的参数表示更精确的室外场景光照信息。

### 3.2.3 基于深度学习的光照估计方法

近期许多工作尝试使用深度学习解决光照估计问题。和传统方法类似，使用深度学习求解光照估计问题时也会借助一些辅助信息、探测物等。Calian等人[78]使用人脸作为光探测物，搭建了神经网络从照片恢复出室外场景的光照。Yi等人[79]搭建了一个高光提取神经网络和阴影反照率估计网络，用于从人脸照片中恢复出室内外场景的光照信息。Georgoulis等人[80]使用真实的反射图作为输入，并搭建了两个不同的CNN结构将图片分解为材质变量和光照变量。之后Geogoulis等人[81]又尝试利用深度学习从包含多种已知材质的物体图片中恢复出反射图和光照分布。Mandl等人[82]、Weber等人[17]也是借用了已知的物体几何恢复出场景光照。虽然这些光照方法和传统方法类似，也借助了一些额外的信息。但通过将传统方法中复杂的算法步骤替换为用深度学习求解，往往能取得更好的结果，这通常得益于大规模的训练数据。

使用深度学习求解光照估计问题时，也会使用与传统方法中相同的光照表示模型。例如基于物理的Sun-sky模型[1]，球形谐波函数模型[82]等。深度学习在学习特征学习上非常有效，因此Weber等人[17]结合深度学习，使用自编码器（auto-encoder）对光照分布进行建模，使用卷积神经网络将场景编码为由少量系数组成的隐变量，为光照估计问题中光照的表示提供了新的思路。

考虑到CNN强大的学习能力，最近一些方法尝试使用深度学习工具直接从单张图片中恢复整个场景的光照分布。Holdgeoffroy等人[1]搭建了一个深度卷积神经网络，从室外图片中恢复出场景的参数化光照模型。Gardner等人[2]从室内图片中直接估计HDR全景图像。这两个方法是目前单图片光照估计工作中最先进的方法。由于HDR全景数据有限，这些方法设计了一类光源探测系统，将LDR全景图转化为粗糙的HDR全景图，用于在大规模LDR数据集上进行训练。

### 3.3 问题求解范围

从视角有限的单张图片预测完整的场景光照分布是一个复杂的问题，目前已有的算法大多通过增加输入信息或简化光照模型的方式降低问题的难度。在增加输入信息方面，许多工作使用多张图片、或额外的深度信息、或已知的几何信息等。这种方式虽然可以简化光照估计问题的难度，提升光照估计效果，但增加输入信息意味着要使用额外的设备（例如如借助深度相机）或更多的获取步骤（例如多次拍摄）。因此一个合理的光照估计方法需要在输入信息的获取步骤与光照估计的效果之间找到一个平衡点，即划定输入信息的规模，使得使用这个规模的信息不仅不需要过多地增加额外的步骤和设备，而且可以更大的提升光照估计的表现。

现代智能设备拍摄的成对照片可以作为这样的一个平衡点。这些设备通常都具有前后两个摄像头，经过在各类手机的验证，绝大部分设备的两个摄像头可以同时运行（常用的手机应用很少需要同时用到前后摄像头，因此并没有应用同时开启前后摄像头，但经过在多种机型验证，这确实是简单做到的）。因此可以在同一时刻使用现代智能设备的前后置相机拍摄两幅图片。这和拍摄一幅图片的步骤完全相同，用户既不需要多余的拍摄步骤，更不需要使用额外的设备。同时，小节3.7中的实验也表明，使用前后两个视角的图片作为输入时，能够极大地提高光照估计的效果。

光照估计的输入决定了光照估计问题的求解范围。使用单张图片作为输入时，光照估计问题的求解范围是图片，使用额外的深度信息辅助估计光照时，问题的求解范围则为RGBD图像。使用成对的图片作为输入时，问题的求解范围依然是图片，而且同时拍摄两张图片并不会增加额外的步骤。据此本文首次提出使用现代设备前后相机拍摄的两幅图片作为光照估计方法的输入，这对于光照估计问题在智能设备中的应用来说有着很大的实际意义。

### 3.4 光照分布的球形谐波表示

使用简化的光照分布近似模型是降低光照估计问题难度的另一个思路。目前的光照估计方法中使用较多的有球形谐波（SH）函数近似，点光源集合近似，小波函数近似，以及基于物理模型的室外光照近似。类似于这些工作，本文也选择使用SH来近似场景的光照。光照的SH近似最早由Sloan等人[91]提出，这种近似方式有很多优点：一方面SH系数小巧而且高效，场景的光照可以使用少量的系数来表达，通过预算算一部分信息，SH渲染过程能够很容易地达到实时。另一方面SH在近似场景光照时可以保留绝大部分的低频信息和小部分的高频信息，使用真实光照和SH系数近似的光照在渲染漫反射物体时差别非常小。通用、高效是本文和目前大部分光照估计工作选取球形谐波函数来近似光照的主要原因。

使用SH近似场景光照需要用到SH基函数。勒让德多项式（Legendre polynomial）是SH函数的核心，这是一个定义球表面上的，类似于傅里叶变换的数学系统。SH基函数通常是在虚数上定义的，不过在近似光照时，只考虑近似球体上的实函数。SH基函数通常由符号 $y$ 表示。

$$y_l^m(\theta, \phi) = \begin{cases} \sqrt{2}K_l^m \cos(m\phi) P_l^m(\cos\phi) & m < 0 \\ \sqrt{2}K_l^m \sin(|m|\phi) P_l^{|m|}(\cos\phi) & m > 0 \\ \sqrt{2}K_l^0 P_l^0(\cos\phi) & m = 0 \end{cases} \quad (3.2)$$

其中 $l$ 表示基函数的阶， $-l \leq m \leq l$ ； $P$ 为勒让德多项式， $K$ 为归一化系数：

$$K_l^m = \sqrt{\frac{(2l+1)}{4\pi} \frac{(l-|m|)!}{(l+|m|)!}} \quad (3.3)$$

图3.1是引自[6]的图片，展示了SH基函数的大致形状。

场景的光照分布可以映射到一个球面上，即 $L(s)$ ， $s$ 表示从球心到球面上的一个方向。通过SH基函数可以将球面上的光照分布映射为SH系数：

$$SH_l^m = \int_s L(s) y_l^m(s) ds \quad (3.4)$$

根据 $l$ 和 $m$ 的关系，可以看出， $l$ 阶SH需要 $2l$ 个系数。将光照映射到 $n$ 阶时，共需要 $n^2$ 个SH系数。因此对于 $n$ 阶的SH系数，可以将SH展开为一维向量 $SH_i, 0 \leq i < n^2$

球面上的 $n$ 阶SH近似光照 $\tilde{L}(s)$ 可以通过SH系数计算：

$$\tilde{L}(s) = \sum_{l=0}^{n-1} \sum_{m=-l}^l SH_l^m y_l^m(s) = \sum_{i=0}^{n^2} SH_i y_i(s) \quad (3.5)$$

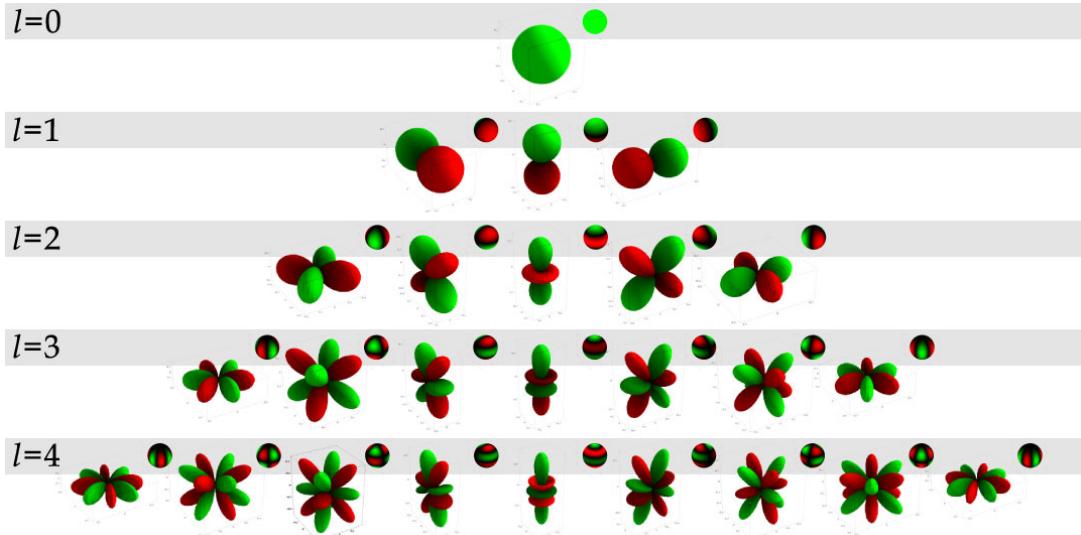


图 3.1 球形谐波基函数的形状 [6]

HDR全景图像映射为SH系数时，需要先将HDR全景图像投影到球面上，随后再将球面上的光照分布映射为一组SH系数。考虑到深度神经网络强大的学习能力，本文使用了4阶SH来近似场景光照，此外由于光照信息包含了三个通道，因此共需要48个SH系数( $4^2 \times 3$ )。

### 3.5 光照估计网络结构

在本文的光照估计方法中，输入为两幅由前后置相机拍摄的相对图片，目标输出为48个SH系数。因此需要搭建一个从两幅图片预测出48个参数的深度卷积神经网络（CNN）。该网络需要从图片中提取特征图，并使用全连接层回归出用以表示光照的SH系数。本文仔细地设计了光照估计的卷积神经网络结构，如图3.2所示。

该网络结构主要包含两个部分：第一部分提取两幅输入图像的特征，第二部分融合这些特征并预测SH系数。

本文构建的HDR数据集虽然有近千张，但是对于训练神经网络来说还是略显不足，因此需要考虑使用预训练的网络和参数提取图像特征。在调研时发现，场景的光照分布与场景的类型密切相关（例如，室外，室内，白天，夜晚等），因此在特征提取部分采用了Zhou预训练的场景分类网络[92]，这个网络在大规模数据集上进行了训练，有可能提取出较好的图像特征。在预测回归阶段，提取的两组图像特征在通道维度被拼接在一起，随后经过卷积层和FC层，最终回归出SH系数，表3.1展示了详细的网络结构，除了最后一层外，其它层之后都

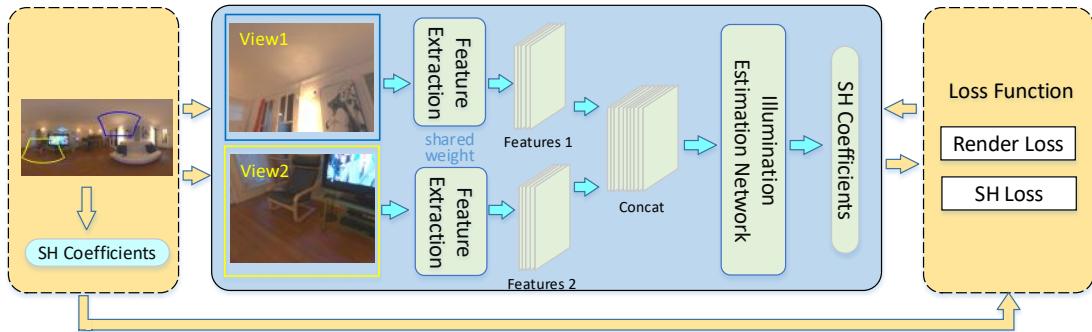


图 3.2 光照估计网络结果一览。在训练阶段（蓝色区域+黄色区域），使用HDR全景图分别提取普通图片和SH系数分别作为网络输入和目标输出信息，并使用**Render Loss**<sup>3.6</sup>作为监督信息。在测试阶段（蓝色区域），成对的图片输入到网络中直接预测光照对应的SH系数。网络结构由两部分组成，一部分是共享权重的特征提取层，另一部分是由卷积层和全连接层构成的光照估计网络。

有batch normalization和leaky RELU激活函数层。

表 3.1 光照估计网络结构。该网络分为两个部分，首先特征提取层从输入图片中提取特征图，之后特征图会在通道层拼接（concatenate）在一起。紧接通过5个卷积层和6个全连接层输出48个球形谐波系数，除最后一层外，每层后都会有batch normalization 和 Leaky RELU激活函数。

	View 1	View 2
Input	$224 \times 224 \times 3$	$224 \times 224 \times 3$
Feature Extraction Network	$5 \times 5 \times 256$	$5 \times 5 \times 256$
Concat	$5 \times 5 \times 512$	
Convs $3 \times 3$	$5 \times 5 \times 64$	
Dense Layers	2048->1024->512->256->128	
Output	$16 \times 3$	

### 3.6 损失函数

在训练神经网络时，损失函数的意义重大。合理的损失函数能够加快网络的收敛，提高网络的表现。本文通过分析SH和渲染结果之间的差异，提出了一个新的用于优化光照估计神经网络的损失函数。

### 3.6.1 球谐参数损失函数

为了预测的SH系数与真实SH系数之间的差异，一个比较直观的做法是使用MSE作为损失函数。不过在SH系数中，不同阶的参数个数是不同的，每一个SH系数在近似光照时所占的权重并不相同。为了平衡这种不均衡的权重，需要先在每一阶的SH系数上计算MSE，然后再不同阶之间做均值。这种均衡化之后的SH距离可以作为训练网络的损失函数。SH Loss的定义如公式3.6

$$\mathcal{L}_{SH} = \frac{1}{n} \sum_{l=0}^{n-1} \left( \frac{1}{2l+1} \sum_{m=-l}^l (SH_l^m - SH_l^m)^2 \right) \quad (3.6)$$

其中， $n$ 为近似光照所使用的SH的阶数，该公式为每个颜色通道上的SH Loss，在整个颜色空间中对每个通道的Loss做均值即可。

### 3.6.2 渲染结果损失函数

虽然MSE是评估两个向量之间距离的常用指标，但是对于光照估计问题来说，仅用MSE来作为损失函数优化SH系数是不够的。实际上，较低的SH误差并不能保证对应光照的渲染结果足够好。SH系数上的微小差异也可能导致渲染结果上非常大的误差。图3.3展示了一个比较SH系数和渲染结果差异的实验，该实验首先选取一张HDR全景图，然后每隔5°地水平旋转该HDR全景图，每次旋转记为一帧；对于每一帧，计算当前帧HDR全景图对应的SH系数，同时使用HDR渲染一个三维物体；通过分别计算当前帧的SH系数和渲染结果与前一帧的差异，绘制了SH系数差异和渲染结果差异之间的关系图。图中可以看出，经过5°的旋转，虽然SH之间的差异非常小，但是渲染结果的变化却十分巨大。因此需要在训练时对渲染结果进行监督。

因此本文提出一种Render Loss，意图在训练光照估计神经网络时，增加渲染结果上的监督信息。具体的做法是先将一些3D物体渲染为一张SH Map（渲染方式参考[6]）。使用SH Map与近似光照的SH系数可以计算出该光照下物体的渲染图像：

$$R(SH, x, y, c) = \sum_{i=1}^{n^2} SHMap(x, y, i) * SH_{c,i} \quad (3.7)$$

其中 $n$ 为近似光照所使用的SH的阶数， $c$ 表示渲染图的某个通道， $(x, y)$ 为渲染结果和SH map的图片坐标。

上述公式表明，在预算算SH map之后，SH系数与渲染结果之间是线性的乘

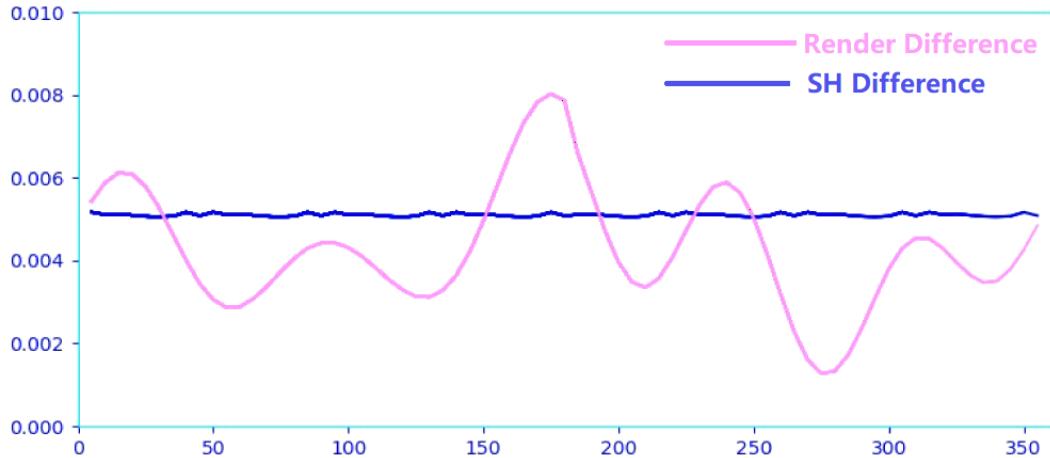


图 3.3 每隔5°地旋转一幅HDR全景图，计算对应的SH系数，随后分别使用旋转后的HDR全景图和SH渲染一个物体，并计算他们之间的差异，它们之间的差异分布如图。

加关系，可以将其添加在神经网络中并计算梯度。因此可以定义Render Loss:

$$\mathcal{L}_{render} = \frac{1}{W \times H \times C} \sum_{x=1}^W \sum_{y=1}^H \sum_{c=1}^C (R(SH, x, y, c) - R(\hat{SH}, x, y, c))^2 \quad (3.8)$$

其中  $R(SH, x, y, c)$  SH的渲染结果在坐标(x, y)，通道c的颜色值。W, H 分别是渲染图像的宽高，C = 3表示RGB三个通道数。最终的损失函数定义为SH Loss和Render Loss的加权和：

$$\mathcal{L} = w_1 * \mathcal{L}_{SH} + w_2 * \mathcal{L}_{render} \quad (3.9)$$

其中  $w_1$  和  $w_2$  是用来平衡  $\mathcal{L}_{sh}$  和  $\mathcal{L}_{render}$  权重的超参数。

### 3.7 实验结果与评估

为了验证本文的深度光照估计方法的有效性，本文设计了光照估计实验，对比了提出的光照估计方法与当前最先进的（SOTA）方法，实验结果表明本文方法在室内室外场景下均优于其他方法。同时，该方法在真实场景下进行了测试，也获得了比较理想的结果。

#### 3.7.1 实现细节

- **数据准备** 通过拍摄和收集，本文构建了包含近千张HDR全景图的数据集。在训练光照估计网络时，需要使用这些HDR全景图生成大批量的训练和测试数据。首先HDR全景图按照90%，5%，5%的比例划分为训练集，测试集和验证集，

训练集用来训练光照估计网络，测试上的结果用来指导网络的超参数调整。验证集用以评估光照估计网络的表现。

接下来生成输入图片。首先将划分后的每张HDR全景图映射到单位球体的表面，并将视点置于球心，然后选取一个视线方向 $(\theta, \phi)$ 。根据球面坐标系的定义可知其相对方向为 $(\pi - \theta, -\phi)$ ，这两个方向上的视图可以视为前后摄像头拍摄的两幅照片。由于大部分智能设备的前后置相机并不会很好的对齐，选取相对方向时，在垂直和水平方向分别添加了一个标准差为 $5^\circ$ 的高斯扰动。同时，为了通过数据增强提高光照估计神经网络的泛化能力，在提取图片时会使用随机的曝光值 $e$ ，即HDR全景图乘以 $2^e$ ，其中 $e$ 服从 $[-1.5, 1.5]$ 之间的均匀分布。

在选取好方向 $(\theta, \phi)$ 之后，该方向上的SH系数也可以被提取出来，该实验使用4阶的SH来近似场景光照，因此这里SH的系数共有48个（ $4^2$ 通道）。

对于每张HDR全景图，会均匀地随机128个方向来提取图片和SH系数，在过滤掉过度曝光和欠曝光的图片后，用于训练光照估计网络的图片/SV系数数据对大概为12万组。此外，数据的划分是在HDR全景数据集上进行的，所以同一幅HDR图像不会同时出现在训练集、测试集或验证集中，规避了训练集和测试集中包含相同图片的可能。

- **训练细节** 训练光照估计网络的硬件平台、操作系统、优化器、学习率等参数列于表3.2中。另外在训练时，用以平衡 $\mathcal{L}_{sh}$ 和 $\mathcal{L}_{render}$ 权重的 $w_1$  和  $w_2$ 被设为0.8和0.2，这是通过探究实验（小节3.8）获取的最佳配置。

表 3.2 光照估计网络的训练细节。

项目	配置
硬件平台	处理器Intel i9-9900K，内存16G，显卡NVIDIA GTX 1080
操作系统	Ubuntu 18.04 64位
优化器	RMS Prop 优化器
批大小	16，又称batch size
初始学习率	0.0005
学习率衰减	每训练2万步，学习率乘以0.9
训练时间	100个epoch，约1200万对数据，750万步

- **评估方式** 评估光照估计的常用方式是计算真实渲染结果（通常是图像）与估计光照渲染结果之间的差异。衡量图片之间距离的常用度量指标有均方差(mean squared error, MSE)、均方根 (root mean squared error, RMSE)、平均

表 3.3

与最先进的方法进行定量比对, Hold-Geoffroy等人[1]和Gardner等人[2]的工作分别适用于室外和室内图像。本文的结果在数值上超过了这两个方法。

Image	Metric	Ours	[HGSH]	[GSY]
Indoor	RMSE	<b>0.1437</b>	0.1676	0.2182
	DSSIM	<b>0.0729</b>	0.0759	0.1065
Outdoor	RMSE	<b>0.1185</b>	0.1609	0.1984
	DSSIM	<b>0.0670</b>	0.0780	0.0985
Total	RMSE	<b>0.1239</b>	0.1622	0.2027
	DSSIM	<b>0.0686</b>	0.0776	0.1003

绝对误差差 (mean absolute error, MAE)、结构相似性 (structural similarity, SSIM)、结构差异性 (structural dissimilarity, DSSIM)、峰值信噪比 (peak signal to noise ratio, PSNR) 等等。与现有的光照估计方法类似, 本实验中选用的指标为RMSE和DSSIM。对于验证集中的每一对数据, 真实的渲染结果使用原始的HDR全景图进行渲染, 将其作为ground truth与使用预测SH渲染的结果对比, 计算它们之间的RMSE和DSSIM, 最后使用验证集上所有的RMSE和DSSIM均值作为评价光照估计方法的最终指标。

### 3.7.2 与最先进方法的对比

本文与目前的两个最先进的 (state-of-the-art, SOTA) 光照估计工作进行了对比: Hold-Geoffroy等人[1]的室外光照估计工作, Gardner等人[2]的室内光照估计工作。前者通过卷积神经网络, 从室外图片预测出sun-sky模型[14]的几个物理参数, 进而达到估计光照的目的; 后者通过在大规模低动态范围全景图上预训练、在小规模室内高动态范围全景图上微调 (fine-tune), 构建了一个从室内图片预测场景光照的深度学习模型; 这两个工作都提供了从图片到HDR全景图的在线应用, 上传一张图片后就可以获取估计的光照分布。在验证集上, 本文方法与这两个工作进行了对比, 数值结果如表3.3所示。可以看出, 无论是室外场景还是室内场景, 本文方法在RMSE和DSSIM两个指标上均明显优于其它工作。图3.4是本方法和两个SOTA方法的可视化对比, 结果显示是用本文方法估计的光照渲染的3D物体更接近真实的渲染结果。

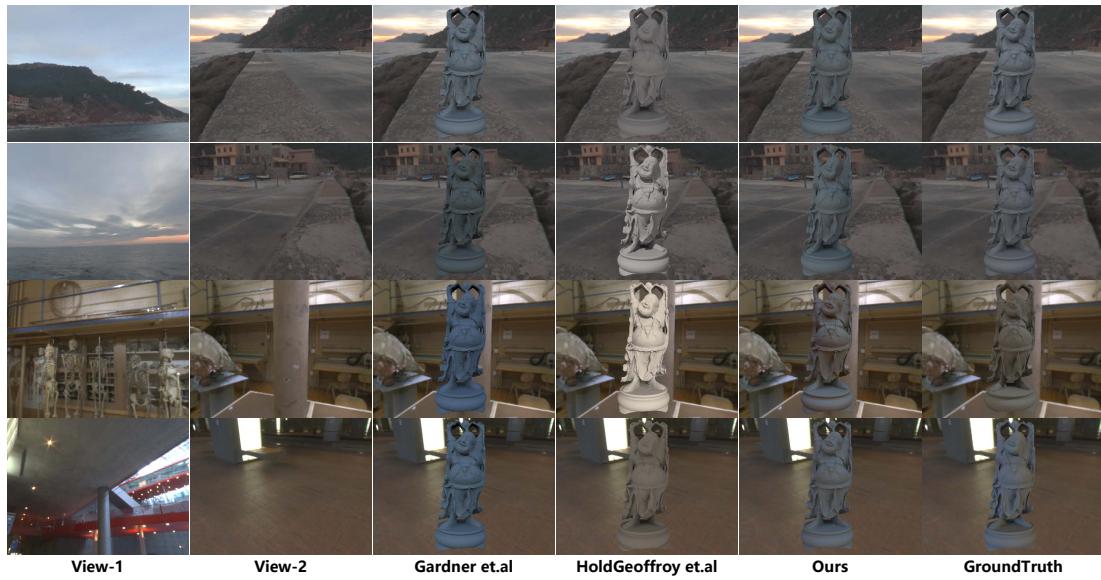


图 3.4 本文工作与Gardner等人[2]、Hold-Geoffroy等人[1]两个最先进方法的对比。可以看出，本文的方法在室内和室外场景中结果均优于这两个方法。

### 3.7.3 在真实数据上的表现

本文的光照估计方法使用视角相对的两张图片作为输入，这对于具有前后相机的现代设备来说有着实际的意义。因此本文测试了该光照估计方法在真实场景中的表现。用于测试的设备是一部普通的智能手机和一台全景相机。首先使用智能手机的前置和后置相机同时拍摄两张图片，然后使用全景相机在同样的位置获取HDR全景图。拍摄的两张图片输入到光照估计网络，随后使用预测的SH系数渲染一个3D物体插入到图像中，最后将此结果与使用HDR全景图渲染的结果进行对比。图3.5展示了本文方法在真实场景中的表现，可以观察到本文方法在多种场景下均能取得较好的预测结果。

## 3.8 深入研究光照估计网络

本文方法中的网络结构包含了多个模块：特征提取模块，特征融合模块，光照估计模块等等，训练过程中也有多个可调的参数，例如损失函数中的权重系数 $w_1$ 和 $w_2$ 。这些模块和参数的选取与使用方式都在一定程度上影响了光照估计效果。本节通过几十组详细的实验，探究这些模块、参数对光照估计结果的影响，从可视结果与数值结果上就行定性和定量的分析。这些实验和分析对以后的深度学习、光照估计工作来说有着一定的指导意义。



图 3.5 本文方法在真实图像上的估计结果。(a) 场景的HDR全景图; (b) 前置摄像头的图像; (c) 后置摄像头的图像; (d) 预测结果; (e) 来自HDR全景图渲染的参考图像。由于真实场景中输入图片和HDR全景图并不是使用一个设备拍摄的,为了得到输入图片对应的真实渲染结果,需要手动对齐输入图像与HDR全景图,这导致了它们之间有一些小的偏移,不过这对于结果评估来说没有影响。结果显示本文方法在真实场景中也能取得很好的预测结果。

### 3.8.1 探究特征提取模块

在光照估计网络中，特征提取模块使用的是Zhou等人[92]的场景分类网络结构。不过Zhou等人[92]在进行场景分类时，使用了多种类型的网络模型，例如AlexNet[71]，GoogleNet[93]，VGG-16[72]等。即使使用同样的数据和训练方式，这些网络在场景分类时也有着不同的表现。使用不同的网络作为光照估计网络中的特征提取模块，也会极大地影响光照估计的效果。因此需要通过实验对比使用不同的网络结构时结果的差异。

**表 3.4 使用不同特征提取网络的结果比较，评价指标采用RMSE和DSSIM。**对于每种引入的网络结构和参数，使用三种方式来评估网络性能：(a) 从头开始训练整个网络 (**train from scratch**)；(b) 从预训练模型中微调(**fine-tune**)。(c) 固定特征提取网络的预训练参数，仅训练照明估计网络 (**freeze**)。结果表明，在不同网络中处理特征提取网络的最佳方式是不同的。这取决于采用的骨干模型以及训练数据的大小。基于此表格，我们使用固定参数的**AlexNet**作为特征提取网络。

	GoogLeNet		VGG-16		AlexNet	
	RMSE	DSSIM	RMSE	DSSIM	RMSE	DSSIM
(a)	0.1304	0.0656	0.1269	<b>0.0642</b>	0.1638	0.0799
(b)	<b>0.1292</b>	<b>0.0656</b>	<b>0.1303</b>	0.0661	0.1318	0.0717
(c)	0.1329	0.0696	0.1336	0.0718	<b>0.1239</b>	<b>0.0686</b>

对于同一个特征提取网络，预训练的网络参数的使用方式也有多种：

- **Freeze**，固定住预训练好的网络参数。这种方式将直接使用预训练的参数进行特征提取，在训练整个网络时，这部分参数保持固定，不参与变量的更新。
- **Fine-tune**，以预训练的网络参数作为初始参数。在训练的过程中与网络中的其它部分一起计算梯度，更新参数。
- **From scratch**，不实用预训练的参数。这种方式只使用引入的网络结构，网络中变量使用随机的初始值，并参与网络中的变量更新。

对于不同的问题、模型、训练集，选取合理的预训练参数引入方式十分重要。本节设计了九组实验，分别使用不同的网络结构和参数引入方式组合。表格3.4展示了这些方式在验证集上的数值结果，选取的两个评价指标是渲染结果上的RMSE和DSSIM。可以看出，对于不同的网络结构，最佳的预训练网络参数的引用方式并不相同。对于GoogleNet和VGG-16这两种结构来说，最好的方式

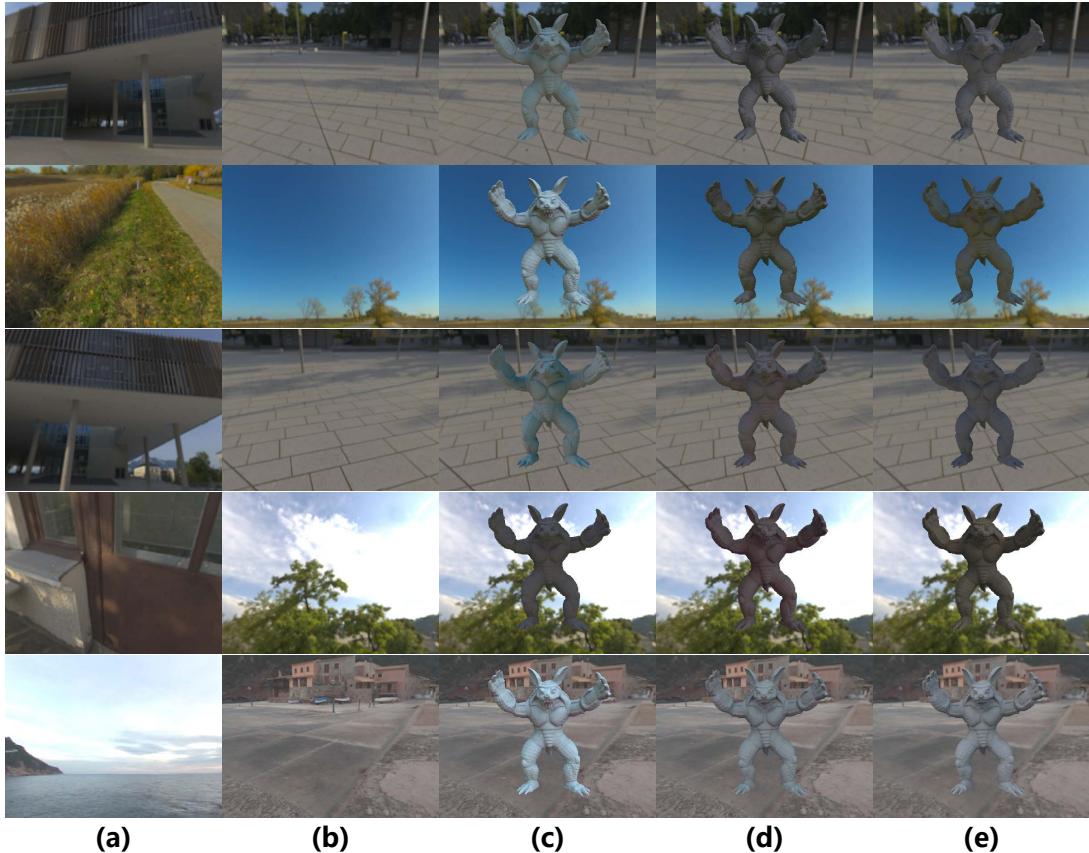


图 3.6 不同的预训练参数引用方式对光照估计的影响。(a) (b) 输入图片; (c) Train from scratch的结果; (d) Freeze参数的结果; (e) 真实渲染结果。

是fine-tune预训练参数。而对于Alexnet来说最好的使用方式是Freeze，即固定住预训练参数，其中的变量不参与参数更新。图3.6展示了在不同的参数引入方式下，光照估计的结果对比，可以看出固定预训练网络参数的方式更接近真实的渲染结果。

不过值得一提的是，目前的结果只是在当前数据集上的表现。由于数据集规模限制，如果不使用预训练参数，网络可能难以从有限的数据中学到足够好的特征选取方式，因此固定预训练参数的方式表现最好。在数据集扩充以后，最佳的参数引入方式极有可能发生变化，当然，这需要在以后研究中使用更大规模的数据进行更详细的验证。

在目前的数据集上，表现最好的网络结构与参数引用方式组合是使用固定参数的Alexnet，因此后续的探究中均使用这种结构和方式进行对比实验。

### 3.8.2 探究特征融合方式

本文的光照估计网络的输入是两张图片，使用预训练的网络从这两幅图片

表3.5 使用两种不同的方式融合特征提取层提取的两组特征图像，并对比它们之间的差异，结果显示在该方法中，特征拼接总是优于特征相减。

Backbone	RMSE		DSSIM	
	Concat	Sub.	Concat	Sub.
GoogLeNet	<b>0.1329</b>	0.1390	<b>0.0696</b>	0.0729
VGG-16	<b>0.1336</b>	0.1399	<b>0.0718</b>	0.0719
AlexNet	<b>0.1239</b>	0.1262	<b>0.0686</b>	0.0690

中提取出两组特征图像后，需要将其进行合并才能输入到后续的网络中，因此需要选择合理的特征融合方式。常见的特征图融合方式都是在通道层进行的，例如通道层拼接、通道层相加、通道层相减、通道层相乘等等。深度学习任务中使用做多的特征融合方式是通道层拼接（concatenate）。

在该问题中，特征图在通道层相加是不合理的。由于两幅图片使用的是完全相同的特征提取网络，所以当融合方式是相加时，交换两幅图片产生的结果不会有任何变化，但实际上这两个方向的光照是截然不同的，这显然不合常理。而拼接和相减却没有这个问题，因此本节实验对比了这两种方式对光照估计结果的影响，实验结果如表3.5所示，可以看出通道层拼接在各种网络结构和评价指标上均优于通道层相减。

### 3.8.3 探究损失函数

表3.6 使用具有不同权重配比的损失函数训练光照估计网络，并在测试集上分析它们之间的优劣。其中 $w_1$ 和 $w_2$ 对应于公式3.9中定义的权重系数。我们可以观察到使用混合损失函数可以提高光照估计的性能。

$(w_1, w_2)$	(0.0, 1.0)		(0.2, 0.8)		(0.5, 0.5)		(0.8, 0.2)		(1.0, 0.0)	
	RMSE	DSSIM	RMSE	DSSIM	RMSE	DSSIM	RMSE	DSSIM	RMSE	DSSIM
GoogLeNet	0.1422	0.0742	0.1334	0.0712	0.1445	0.0785	<b>0.1329</b>	<b>0.0696</b>	0.1376	0.0732
VGG-16	0.1447	0.0749	0.1561	0.0768	0.1587	0.0765	<b>0.1336</b>	0.0718	0.1656	<b>0.0674</b>
AlexNet	0.1247	0.0678	0.1268	<b>0.0675</b>	0.1479	0.0770	<b>0.1239</b>	0.0686	0.1267	0.0682

为了探究提出的render loss对光照估计结果的影响，并找到最优的损失函数权重系数 $w_1$ ,  $w_2$ (公式3.9)，本节设计和实现了十几组实验，详细地对比了使用不同的 $w_1$ ,  $w_2$ 的损失函数对光照估计结果的影响。表3.6列出了详细的实验结果，结果表明使用 $w_1 = 0.8$ ,  $w_2 = 0.2$ 的损失函数能够达到最好的结果。图3.7的可视结果也表明，相较于只使用一类损失函数，综合使用两个损失函数可以得出更

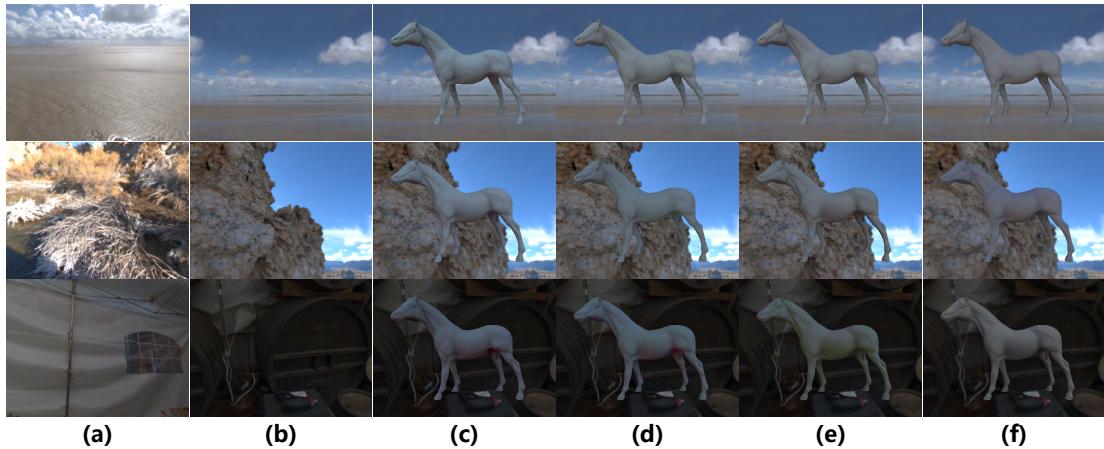


图 3.7 损失函数对光照估计的影响。 (a)(b) 输入图像(c) 只使用SH Loss(公式3.6); (d)只使用RenderLoss (公式3.8); (e) 使用加权loss(公式3.9); (f) 真实结果

加接近真实值的效果。

需要注意的是，在训练时计算render loss所用的三维物体，与在测试时渲染的三维物体是不相同的。因此render loss能够提升效果的原因并不是来源于3D物体的先验知识，而是render loss本身对于渲染过程和场景光照的关注。这也是只使用render loss的效果比较差的原因之一。

### 3.8.4 探究光照估计性能

本文所提出光照估计算法非常适合现代智能手机应用，因此模型的性能也是需要考量的一个因素。本文测试了光照估计的性能表现，在普通的消费级桌面显卡NVIDIA GTX 1080上，从一对图片估计出SH系数的耗时为0.0391秒，在CPU（INTEL i76800k）上则需要0.3039s。虽然该耗时无法在大量低端设备中达到实时，但是通过优化和裁剪仍然可以达到不错的交互速度。此外，目前的网络结构使用的是AlexNet，虽然该模型在当前数据集上最好的选择，但随着数据集规模的不断增加，一些轻量级的网络可能会更加合适。同时随着现代智能设备处理器性能的不断提高，该方法有望在移动设备中达到实时交互的效果。

## 3.9 讨论

本章提出了一个从图片恢复场景光照的深度学习模型。该模型使用现代智能设备的前后相机拍摄的图片作为输入，预测出场景光照分布对应的球形谐波系数。该网络由特征提取模块，特征融合模块，光照估计模块组成，其中特征提取模块引用了Zhou[92]的网络结构和参数。结合创新提出的损失函数render loss，该网络模型不仅能够获得超过最先进的方法的结果，其在真实场景的表现也很理

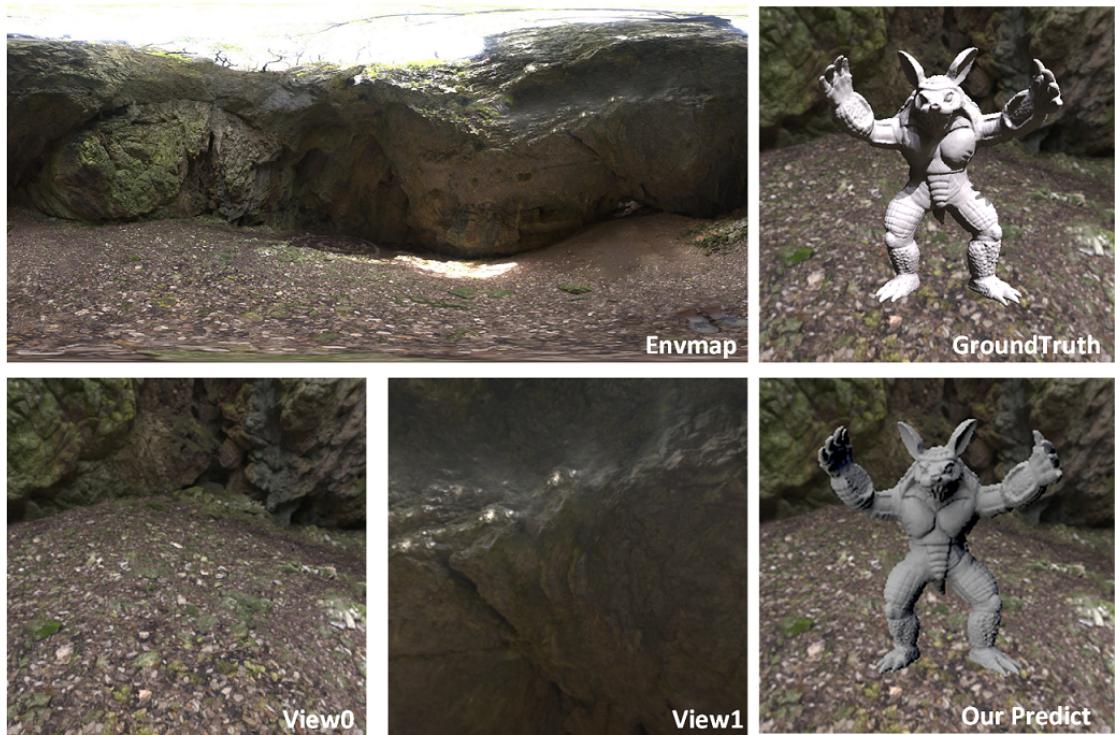


图 3.8 一种本文方法无效的情况。从图中可知全景图的顶部有一片十分强的光照区域，但是两幅输入图片都不包含推断这个光照的线索。此外这束强光属于该图中的高频信息，难以使用**SH**近似。由于这两个原因，在该图片上的预测结果与真实值相差甚远。

想。不过该方法仍然有一些局限性。一方面是使用**SH**表示光照的局限性，使用**SH**表示的光照往往会忽略掉一些高频信息，这对于具有镜面反射表面的物体来说很不友好。另一方面是仅使用图片预测光照时，对输入信息的依赖比较严重，如果输入图片既没有拍摄到场景的光照也没有拍摄到能推理出光照的阴影，则光照估计的效果很可能会不太理想。图3.8展示了一种因为这两种局限性导致的失败情况。

使用本文的方法在利用前后置摄像头估计光照时会有一个特定的问题，就是人脸经常会出现前置摄像头中。比较好的解决办法是用户平移一下手机或者移动一下头部，来通过前置相机获取到更丰富的场景。不过，出现在前置摄像头中的人脸并不能完全视为光照估计问题的一个阻碍。在相关工作中提到，一些研究者使用人脸作为标志物，辅助估计场景的光照。这些工作为基于前后摄像头估计光照的方法提供了一个新的思路——可以将基于人脸的光照估计方法嵌入到本文的方法中，从前置相机的人脸和后置相机的普通图片中共同恢复场景光照，这也是进一步的工作之一。

### 3.10 本章总结

从有限的图像信息估计出整个场景的光照分布非常复杂。从图像（Image）反推出光照（Light）是一个严重的不适定（ill-posed）问题。相对于整个场景，有限视野的图像不仅包含的信息有限，而且在不同的条件下拍摄的彩色图像可能存在很多误差。这些都会对光照估计造成一定程度的干扰，增加光照估计的难度。

为了降低问题的难度，研究者们尝试对该问题进行约束或简化。传统方法通常增加输入的信息或者减少要估计的光照模型规模。例如增加深度信息、几何信息，或者使用球形谐波模型、Sun-Sky物理模型表示场景的光照等。近年来，深度学习也被应用在光照估计问题当中，不过，目前已有的深度学习方法也有一定的局限性。训练一个鲁棒的神经网络往往需要大量的数据，而目前用于光照估计问题的数据集比较有限，如SUN360[3]和Laval Indoor[2]等很难在规模和质量上同时到达训练深度神经网络的要求。

在这样的背景下，本文在基于深度学习的光照估计中的两个方向开展研究。其一是构建一个具有一定规模和质量光照估计的数据集。其二是对深度网络模型进行研究，这也是本章的主要内容。

本文提出使用前两张相对视角的图片作为光照估计的输入。这两张图片多由相机的前后置摄像头拍摄，使用前后摄像头同时拍摄两张照片不仅不会增加获取图片的步骤，还可以极大地降低光照估计的难度。现代移动设备几乎都包含前后至少两个摄像头，因此本文所提方法对于智能手机应用来说有着很大的实际意义。

接着介绍了使用球形谐波技术表示光照的方法，以及光照估计问题中使用这种方法的优点，并基于此提出一种新的损失函数——render loss。它通过将渲染过程的可导部分置于卷积神经网络，增加对渲染结果的监督，达到了优化光照估计效果的目的。

随后展示了本文所构建的基于深度学习的光照估计网络模型，该网络模型使用两张图片作为输入，估计预测出场景中光照对应的球形谐波系数。目前该方法不仅已经取得了超过state-of-the-art的结果，它在真实场景的中的表现也非常理想。

最后，本文对深度学习模型的各个模块、损失函数、运行性能、局限性等进行了十分详尽的实验和分析。通过几十组对比实验深入探索了使用不同网络结构的特点和对光照估计结果的影响。这些实验和分析对以后的深度学习、光

照估计工作来说有着一定的指导意义。

## 第4章 总结与展望

### 4.1 本文工作总结

光照估计（又称光照分布估计）是从已知的彩色图像信息中，预测、估计、恢复出整个场景的光照分布。场景的光照分布是指场景中各个方向的光照的颜色和强度。该问题的输入通常是若干张彩色图片，或者是一段视频，有时已知的几何或材质信息也被用来辅助估计光照。光照估计作为计算机图形学和计算机视觉的基础问题之一，有着广泛的实际应用场景。例如：基于图像的渲染（Image Based Rendering, IBR）、增强现实（Augmented Reality, AR）、电影后期制作、真实感虚实交互等。图1.1展示了光照估计的应用之一。光照估计也与这两个学科中的许多其它问题息息相关。例如：双向反射分布函数（BRDF）估计、场景几何重构、本征信息提取、图像增强，等等。高质量的光照估计结果通常能够为这些问题的解决带来很大的帮助。

从有限的图像信息估计出整个场景的光照分布是一个复杂的问题。首先，图像的视野范围比较有限，例如一张视场角（FOV）为 $60^{\circ}$ 的照片所拍摄到的区域，在其对应的全景图中占比不足6%。此外，一幅图片是光照分布、场景几何结构、物体材质、摄相机参数等多个单位之间的复杂交互结果，从图像反推出光照是一个严重的不适定（ill-posed）问题。不仅如此，在不同的条件下拍摄的彩色图像可能存在很多误差。例如图像中的过曝光/欠曝光区域、相机畸变、不正确的白平衡等。这些都会对光照估计造成一定程度的干扰，增加光照估计的难度。

为了简化问题难度，传统方法常常增加输入信息的数量或缩小光照模型的规模。其中一部分工作使用更多的输入信息辅助估计场景光照。例如深度信息、几何信息、多张图片、先验知识、用户标记等等。这类方法或依赖特殊的探针、或依赖特殊的拍摄设备、或依赖额外的辅助信息与模型假设，均具有一定的局限性。另一部分工作通过使用低维的光照表示模型来简化光照估计问题，例如使用球形谐波函数（SH）来拟合场景光照、使用小波函数近似场景光照、使用若干个点光源的集合近似场景光照、使用基于物理的室外光照模型等等。可以看出，无论是增加输入信息还是使用简化的光照估计模型，传统光照估计方法都具有很大的局限性。

近年来，深度学习在多种计算机视觉问题上大放异彩，用于分割、检测、

标识、分类的神经网络层出不穷。一些研究者尝试将深度学习应用在光照估计问题当中。其中Hold-Geoffroy等人[1]和Gardner[2]等人的工作是应用深度学习估计光照的最先进方法。它们在大规模数据集上训练了一个深度卷积神经网络，分别用于室内和室外的场景光照估计，能达到较好的效果。不过，目前已有的深度学习方法也有一定的局限性。训练一个鲁棒的神经网络往往需要大量的数据，而目前用于光照估计问题的数据集比较有限，主要包括：大规模的低动态范围全景数据集（SUN360[3]等）和中小规模特定场景的高动态范围全景数据集（Laval Indoor等[2]）。这些数据集在规模和质量上很难同时到达训练深度神经网络的要求。

在这样的背景下，本文在基于深度光照估计问题的两个主要方向上进行了细致的研究。其一是严格仔细地构建了一个具有一定规模和质量的光照估计数据集。数据集由高质量的高动态范围全景图构成，这样的数据集不仅能被用来训练更加鲁棒的光照估计网络，也可以被应用到其它多种相关的深度学习问题当中。其二是在已有数据集和本文构建的数据集基础上，深入探索基于深度学习的光照估计方法，对其中的网络结构，网络参数，损失函数，光照表示等多个模块进行了细致的对比和研究。作为总结，将本文的主要工作内容和创新贡献罗列如下：

- 深入调研了全景图以及高动态范围全景图的获取步骤、投影方法、存储方式等，使用全景相机和曝光融合算法构建了一个用于光照估计、大规模、高质量的HDR全景数据集，并通过实验证明了该数据集相对于其它数据集在光照估计问题中的优越性。
- 通过详细的实验分析了HDR全景数据集的规模和多样性对于光照估计问题的影响，证明了丰富的数据多样性和较大的数据规模能够为基于深度学习的光照估计结果带来有效提升。
- 首次提出使用视角相对的图片作为光照估计的输入，这两张图片多由相机的前后置摄像头拍摄。使用前后摄像头同时拍摄两张照片不仅不会增加获取图片的步骤，还可以极大地降低光照估计的难度。这种方法对于光照估计在智能设备中的应用来说有着很大的实际意义。
- 构建了一个基于深度学习的光照估计网络模型，该网络模型使用两张图片作为输入，估计预测出场景中光照对应的球形谐波系数。已有的实验结果表明，该模型在光照估计问题中非常有效，目前已经取得了超过state-of-the-art的结果。

- 提出了一个新的损失函数 - Render Loss，该损失函数巧妙地利用了SH的特性，将部分渲染过程置于神经网络中，进而在训练中对渲染结果进行监督，指导网络的优化与调整。在网络中加入这个损失函数的方法算法简洁但非常有效，极大地提高了光照估计的表现。
- 对提出的光照估计深度学习模型和损失函数进行十分详尽的实验和分析，对于光照估计网络结构和训练过程中各个模块，通过几十组对比实验深入探索了使用不同网络结构的特点和对结果的影响，促进了对基于深度学习光照估计的理解。

## 4.2 未来工作展望

尽管本文构建的数据集和光照估计网络取得了不错的效果，但在进行光照估计相关的实验时，依然发现了可以进一步拓展、提高、优化的部分。这些内容或因与本文工作相关性不大、或受限于目前的硬件和技术限制无法开展实验、或由于实验周期较长难以在有限时间内完成，目前没有被包含在本文的工作中。这些内容在满足研究条件后，都可能成为一些新的、独立的研究课题。

在光照估计数据集方面，本文构建的光照估计数据集包含了近千张HDR全景图像。这个数量还可以增加。通过第2章的实验可以发现，在训练用于光照估计的深度网络时，增加数据的多样性比单纯的增加数据更加有用，这可以作为后期扩充数据时的主要依据和指导。

在输入信息方面，本文的方法是使用现代智能设备前后相机拍摄的两幅图片作为输入。在实际使用中，人脸可能会经常出现在前置摄像头中。本文所提出的方法中没有针对人脸的部分，因此为了避免人脸所占区域过多，需要用户让手机或者头部做一下水平偏移。不过，出现在前置摄像头中的人脸并不能完全视为光照估计问题的一个阻碍。一些光照估计方法专门使用人脸作为标志物，辅助估计场景的光照。这些工作为基于前后摄像头估计光照的方法提供了一个新的思路——可以将基于人脸的光照估计方法嵌入到本文的方法中，从前置相机的人脸和后置相机的普通图片中共同恢复场景光照，相信这是一个值得研究和探索的方向。

在光照表示方面，球形谐波函数本身具有一定的局限性，在表示光照时难以处理非常高频的光照细节，在渲染时对镜面反射表面不太友好。一些工作使用自编码器（Auto-Encoder）对光照进行建模，这种方式仍然会丢失一部分精度信息，并且非常依赖用于构建自编码器的训练数据。如何在尽可能不损失精度

的情况下，表示兼具高频和低频的光照信息是一个值得探索的领域。

在深度学习模型方面，模型的性能也是一个需要考虑的因素。目前的模型中，主要的网络结构是AlexNet，虽然这是在目前光照估计数据集上表现较好的网络，但Alexnet本身参数量较大，运行时间也不是很短。轻量化和高效性是网络模型能否用于移动设备中的两个重要影响因素，因此在训练数据得到扩充之后，可能会有更加轻量和高效的网络模型能够应用到光照估计问题中来。

此外，视频也可以作为光照估计问题的输入，视频往往包含了更丰富的信息。传统方法使用视频估计光照时通常能获得较好的效果，但目前还没有使用深度从视频估计光照的方法。相信这也是一个值得探索的研究方向。

## 参考文献

- [1] 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. 2017: 7312-7321.
- [2] GARDNER M A, SUNKAVALLI K, YUMER E, et al. Learning to predict indoor illumination from a single image[J]. ACM Transactions on Graphics (TOG), 2017, 36(6):176.
- [3] XIAO J, EHINGER K A, OLIVA A, et al. Recognizing scene viewpoint using panoramic place representation[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 2695-2702.
- [4] REINHARD E, WARD G, PATTANAIK S, et al. High dynamic range imaging: Acquisition, display, and image-based lighting[M]. Elsevier, 2005.
- [5] RAMAMOORTHI R, HANRAHAN P. An efficient representation for irradiance environment maps[C]//Proceedings of the 28th annual conference on Computer graphics and interactive techniques. ACM, 2001: 497-500.
- [6] GREEN R. Spherical harmonic lighting: The gritty details[C]//Archives of the Game Developers Conference: volume 56. 2003: 4.
- [7] SLOAN P P. Stupid spherical harmonics (sh) tricks[C]//Game developers conference: volume 9. 2008: 42.
- [8] PEREZ R, SEALS R, MICHALSKY J. All-weather model for sky luminance distribution—preliminary configuration and validation[J]. Solar energy, 1993, 50(3):235-245.
- [9] NISHITA T, DOBASHI Y, NAKAMAE E. Display of clouds taking into account multiple anisotropic scattering and sky light[C]//Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. ACM, 1996: 379-386.
- [10] SIRAI T N T, NAKAMAE K T E. Display of the earth taking into account atmospheric scattering[C]//Siggraph: volume 93. Citeseer, 1993: 175.
- [11] PREETHAM S. A practical analytic model for daylight[J]. 1999.
- [12] RAAB M, SEIBERT D, KELLER A. Unbiased global illumination with participating media [M]//Monte Carlo and Quasi-Monte Carlo Methods 2006. Springer, 2008: 591-605.
- [13] HOSEK L, WILKIE A. An analytic model for full spectral sky-dome radiance[J]. ACM Transactions on Graphics (TOG), 2012, 31(4):95.
- [14] HOŠEKHOŠEK L, WILKIE A. Adding a solar-radiance function to the hošek-wilkie skylight model[J]. IEEE computer graphics and applications, 2013, 33(3):44-52.
- [15] NG R, RAMAMOORTHI R, HANRAHAN P. All-frequency shadows using non-linear wavelet lighting approximation[C]//ACM Transactions on Graphics (TOG): volume 22. ACM, 2003: 376-381.

- [16] LEGENDRE C, YU X, LIU D, et al. Practical multispectral lighting reproduction[J]. ACM Transactions on Graphics (TOG), 2016, 35(4):32.
- [17] WEBER H, PRÉVOST D, LALONDE J F. Learning to estimate indoor lighting from 3d objects[C]//2018 International Conference on 3D Vision (3DV). IEEE, 2018: 199-207.
- [18] DEBEVEC P. Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography[C]//Proceedings of the 25th annual conference on Computer graphics and interactive techniques. ACM, 1998: 189-198.
- [19] DEBEVEC P, GRAHAM P, BUSCH J, et al. A single-shot light probe[C]//ACM SIGGRAPH 2012 Talks. ACM, 2012: 10.
- [20] NISHINO K, NAYAR S K. Eyes for relighting[J]. ACM Transactions on Graphics (TOG), 2004, 23(3):704-711.
- [21] WANG Y, LIU Z, HUA G, et al. Face re-lighting from a single image under harsh lighting conditions[C]//2007 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2007: 1-8.
- [22] TSUMURA N, DANG M N, MAKINO T, et al. Estimating the directions to light sources using images of eye for reconstructing 3d human face[C]//Color and Imaging Conference: volume 2003. Society for Imaging Science and Technology, 2003: 77-81.
- [23] WEN Z, LIU Z, HUANG T S. Face relighting with radiance environment maps[C]//2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.: volume 2. IEEE, 2003: II-158.
- [24] SHIM H. Faces as light probes for relighting[J]. Optical Engineering, 2012, 51(7):077002.
- [25] KNORR S B, KURZ D. Real-time illumination estimation from faces for coherent rendering [C]//2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, 2014: 113-122.
- [26] SHAHLAEI D, BLANZ V. Realistic inverse lighting from a single 2d image of a face, taken under unknown and complex lighting[C]//2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG): volume 1. IEEE, 2015: 1-8.
- [27] YAO Y, KAWAMURA H, KOJIMA A. The hand as a shading probe[C]//ACM SIGGRAPH 2013 Posters. ACM, 2013: 108.
- [28] COSSAIRT O, NAYAR S, RAMAMOORTHI R. Light field transfer: global illumination between real and synthetic objects[C]//ACM Transactions on Graphics (TOG): volume 27. ACM, 2008: 57.
- [29] IMAI Y, KATO Y, KADOI H, et al. Estimation of multiple illuminants based on specular highlight detection[C]//International Workshop on Computational Color Imaging. Springer, 2011: 85-98.
- [30] CALIAN D A, MITCHELL K, NOWROUZEZAHRAI D, et al. The shading probe: Fast

- appearance acquisition for mobile ar[C]//SIGGRAPH Asia 2013 Technical Briefs. ACM, 2013: 20.
- [31] PILET J, GEIGER A, LAGGER P, et al. An all-in-one solution to geometric and photometric calibration[C]//2006 IEEE/ACM International Symposium on Mixed and Augmented Reality. IEEE, 2006: 69-78.
- [32] YOO J D, LEE K H. Real time light source estimation using a fish-eye lens with nd filters[C]// 2008 International Symposium on Ubiquitous Virtual Reality. IEEE, 2008: 41-42.
- [33] TOCCI M D, KISER C, TOCCI N, et al. A versatile hdr video production system[C]//ACM Transactions on Graphics (TOG): volume 30. ACM, 2011: 41.
- [34] MANAKOV A, RESTREPO J, KLEHM O, et al. A reconfigurable camera add-on for high dynamic range, multispectral, polarization, and light-field imaging[J]. ACM Transactions on Graphics, 2013, 32(4):47-1.
- [35] KÁN P. Interactive hdr environment map capturing on mobile devices.[C]//Eurographics (Short Papers). 2015: 29-32.
- [36] KNECHT M, TRAXLER C, MATTAUSCH O, et al. Reciprocal shading for mixed reality[J]. Computers & Graphics, 2012, 36(7):846-856.
- [37] MEILLAND M, BARAT C, COMPORT A. 3d high dynamic range dense visual slam and its application to real-time object re-lighting[C]//2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, 2013: 143-152.
- [38] BARRON J T, MALIK J. Intrinsic scene properties from a single rgb-d image[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 17-24.
- [39] ZHANG E, COHEN M F, CURLESS B. Emptying, refurbishing, and relighting indoor spaces [J]. ACM Transactions on Graphics (TOG), 2016, 35(6):174.
- [40] LI Y, LU H, SHUM H Y, et al. Multiple-cue illumination estimation in textured scenes[C]// Proceedings Ninth IEEE International Conference on Computer Vision. IEEE, 2003: 1366-1373.
- [41] RAMAMOORTHI R, HANRAHAN P. A signal-processing framework for inverse rendering [C]//Proceedings of the 28th annual conference on Computer graphics and interactive techniques. ACM, 2001: 117-128.
- [42] SATO I, SATO Y, IKEUCHI K. Illumination from shadows[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(3):290-300.
- [43] WANG Y, SAMARAS D. Estimation of multiple illuminants from a single image of arbitrary known geometry[C]//European conference on computer vision. Springer, 2002: 272-288.
- [44] Panagopoulos A, Wang C, Samaras D, et al. Illumination estimation and cast shadow detection through a higher-order graphical model[C/OL]//CVPR 2011. 2011: 673-680. DOI: [10.1109/CVPR.2011.5995585](https://doi.org/10.1109/CVPR.2011.5995585).
- [45] BARRON J T, MALIK J. Shape, illumination, and reflectance from shading[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(8):1670-1687.

- [46] LOPEZ-MORENO J, HADAP S, REINHARD E, et al. Compositing images through light source detection[J]. *Computers & Graphics*, 2010, 34(6):698-707.
- [47] SATO I, SATO Y, IKEUCHI K. Acquiring a radiance distribution to superimpose virtual objects onto a real scene[J]. *IEEE transactions on visualization and computer graphics*, 1999, 5(1):1-12.
- [48] NISHINO K, ZHANG Z, IKEUCHI K. Determining reflectance parameters and illumination distribution from a sparse set of images for view-dependent image synthesis[C]//Proceedings Eighth IEEE international conference on computer vision. ICCV 2001: volume 1. IEEE, 2001: 599-606.
- [49] NISHINO K, IKEUCHI K, ZHANG Z. Re-rendering from a sparse set of images[J]. Department of Computer Science, Drexel University, 2005.
- [50] YU T, WANG H, AHUJA N, et al. Sparse lumigraph relighting by illumination and reflectance estimation from multi-view images[C]//ACM SIGGRAPH 2006 Sketches. ACM, 2006: 175.
- [51] WU C, WILBURN B, MATSUSHITA Y, et al. High-quality shape from multi-view stereo and shading under general illumination[C]//CVPR 2011. IEEE, 2011: 969-976.
- [52] SHAN Q, ADAMS R, CURLESS B, et al. The visual turing test for scene reconstruction[C]// 2013 International Conference on 3D Vision-3DV 2013. IEEE, 2013: 25-32.
- [53] LALONDE J F, MATTHEWS I. Lighting estimation in outdoor image collections[C]//2014 2nd International Conference on 3D Vision: volume 1. IEEE, 2014: 131-138.
- [54] MARSCHNER S R, GREENBERG D P. Inverse lighting for photography[C]//Color and Imaging Conference: volume 1997. Society for Imaging Science and Technology, 1997: 262-265.
- [55] HABER T, FUCHS C, BEKAER P, et al. Relighting objects from image collections[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 627-634.
- [56] KEMELMACHER-SHLIZERMAN I, BASRI R. 3d face reconstruction from a single image using a single reference face shape[J]. *IEEE transactions on pattern analysis and machine intelligence*, 2011, 33(2):394-405.
- [57] GARRIDO P, VALGAERTS L, WU C, et al. Reconstructing detailed dynamic face geometry from monocular video.[J]. *ACM Trans. Graph.*, 2013, 32(6):158-1.
- [58] LI C, ZHOU K, LIN S. Intrinsic face image decomposition with human face priors[C]// European Conference on Computer Vision. Springer, 2014: 218-233.
- [59] OKABE T, SATO I, SATO Y. Spherical harmonics vs. haar wavelets: Basis for recovering illumination from cast shadows[C]//Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.: volume 1. IEEE, 2004: I-I.
- [60] LALONDE J F, NARASIMHAN S G, EFROS A A. What does the sky tell us about the camera?[C]//European conference on computer vision. Springer, 2008: 354-367.
- [61] LALONDE J F, NARASIMHAN S G, EFROS A A. What do the sun and the sky tell us about the camera?[J]. *International Journal of Computer Vision*, 2010, 88(1):24-51.

- [62] LALONDE J F, EFROS A A, NARASIMHAN S G. Estimating the natural illumination conditions from a single outdoor image[J]. International Journal of Computer Vision, 2012, 98(2):123-145.
- [63] SUNKAVALLI K, ROMEIRO F, MATUSIK W, et al. What do color changes reveal about an outdoor scene?[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008: 1-8.
- [64] XING G, ZHOU X, PENG Q, et al. Lighting simulation of augmented outdoor scene based on a legacy photograph[C]//Computer Graphics Forum: volume 32. Wiley Online Library, 2013: 101-110.
- [65] KARSCH K, HEDAU V, FORSYTH D, et al. Rendering synthetic objects into legacy photographs[C]//ACM Transactions on Graphics (TOG): volume 30. ACM, 2011: 157.
- [66] KARSCH K, SUNKAVALLI K, HADAP S, et al. Automatic scene inference for 3d object compositing[J]. ACM Transactions on Graphics (TOG), 2014, 33(3):32.
- [67] CHEN X, JIN X, WANG K. Lighting virtual objects in a single image via coarse scene understanding[J]. Science China Information Sciences, 2014, 57(9):1-14.
- [68] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.
- [69] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database [C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.
- [70] CHANG A X, FUNKHOUSER T, GUIBAS L, et al. Shapenet: An information-rich 3d model repository[J]. arXiv preprint arXiv:1512.03012, 2015.
- [71] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
- [72] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
- [73] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
- [74] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
- [75] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, 2015: 234-241.
- [76] CHAITANYA C R A, KAPLANYAN A S, SCHIED C, et al. Interactive reconstruction of monte carlo image sequences using a recurrent denoising autoencoder[J]. ACM Transactions on Graphics (TOG), 2017, 36(4):98.

- [77] KARRAS T, AILA T, LAINE S, et al. Audio-driven facial animation by joint end-to-end learning of pose and emotion[J]. ACM Transactions on Graphics (TOG), 2017, 36(4):94.
- [78] CALIAN D A, LALONDE J F, GOTARDO P, et al. From faces to outdoor light probes[C]// Computer Graphics Forum: volume 37. Wiley Online Library, 2018: 51-61.
- [79] YI R, ZHU C, TAN P, et al. Faces as lighting probes via unsupervised deep highlight extraction [C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 317-333.
- [80] GEORGOULIS S, REMATAS K, RITSCHEL T, et al. Delight-net: Decomposing reflectance maps into specular materials and natural illumination[J]. arXiv preprint arXiv:1603.08240, 2016.
- [81] GEORGOULIS S, REMATAS K, RITSCHEL T, et al. Natural illumination from multiple materials using deep learning[J]. arXiv preprint arXiv:1611.09325, 2016.
- [82] MANDL D, YI K M, MOHR P, et al. Learning lightprobes for mixed reality illumination [C]//2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, 2017: 82-89.
- [83] PTGUI. new hourse internet services b.v, holland[EB/OL]. 2000. <https://www.ptgui.com/>.
- [84] WIKIPEDIA. panorama[EB/OL]. 2019. <https://www.wikipedia.org/>.
- [85] XIAOMI. Xiaomi Panoramic Camera[EB/OL]. 2016. <https://www.mi.com/mj-panorama-camera/>.
- [86] GREGZAAL.COM. How to create your own hdr environment maps[EB/OL]. 2016. <http://blog.gregzaal.com/2016/03/16/make-your-own-hdri/>.
- [87] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.
- [88] MAAS A L, HANNUN A Y, NG A Y. Rectifier nonlinearities improve neural network acoustic models[C]//Proc. icml: volume 30. 2013: 3.
- [89] TIELEMAN T, HINTON G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude[J]. COURSERA: Neural networks for machine learning, 2012, 4(2): 26-31.
- [90] REMATAS K, RITSCHEL T, FRITZ M, et al. Deep reflectance maps[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4508-4516.
- [91] SLOAN P P, KAUTZ J, SNYDER J. Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments[C]//ACM Transactions on Graphics (TOG): volume 21. ACM, 2002: 527-536.
- [92] ZHOU B, LAPEDRIZA A, KHOSLA A, et al. Places: A 10 million image database for scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.
- [93] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.