基于光场图像的深度估计及快速三维重建研究

马宇航　郑胜男　刘恒亓　陈迪祺
南京工程学院　江苏南京　211167

摘　要：三维重建是计算机视觉中的经典问题之一，其应用领域广泛，一直是相关领域研究的热点。而三维重建的精确性和速度取决于场景深度信息的估计。随着光场成像技术的发展，光场图像的获取越来越便利，光场图像包含四维信息，有利于场景深度信息的精确估计。深度学习在光场图像深度估计中的应用提高了光场图像深度估计的速度和精度，进一步能够实现场景的三维重建。本文研究利用光场图像结合深度学习进行场景深度估计，最终实现近景快速的三维重建。

关键词：光场图像；深度估计；深度学习；快速三维重建

Research On Depth Estimation and Fast 3d Reconstruction Based on Light Field Images

Yuhang Ma, Shengnan, Zheng, HengQi Liu, Diqi Chen
Nanjing Institute of Engineering, Nanjing, Jiangsu 211167

Abstract: Three-dimensional reconstruction is one of the classical problems in computer vision, and its application area is widely used , which has been a hot spot for research in related fields. And the accuracy and speed of 3D reconstruction depends on the estimation of scene depth information. With the development of light field imaging technology, it is more and more convenient to acquire light field images, which contain four-dimensional information and are beneficial to the accurate estimation of scene depth information. The application of deep learning in light field image depth estimation improves the speed and accuracy of light field image depth estimation, and further enables the 3D reconstruction of the scene. In this paper, we study the use of light-field images combined with deep learning for scene depth estimation, and finally realize the near-field fast 3D reconstruction.

Keywords: Light Field Image, Depth Estimation, Deep Learning, Fast 3D Reconstruction

引言：

三维重建技术广泛存在于人们的生产生活中，在工业制造、文物保护、虚拟现实以及智慧城市、自动驾驶等领域都有着重要应用。因此对三维重建的研究经久不衰且分支庞杂。根据数据类型的差异，三维重建技术大致可分类为基于三维数据重建和基于二维数据重建。传统三维数据的获取，主要是利用一些特殊成像设备主动向目标物发射可控光束或电磁波，根据其飞行时间差来获取场景深度信息，如激光扫描、结构光、光栅等，需要专业成像设备、虽然精度较高但是其成本高、应用范围受限。而基于二维数据即图像的三维重建又分为基于单幅图像和多幅图像的三维重建。从目前的研究来看，基于单幅图像三维重建因其缺少深度信息而极具挑战性，基于多幅图像获取深度信息进而实现三维重建是目前主流的研究方向之一。综上所述，三维重建的核心在于场景深度信息的获取。场景深度是指目标物体到相机中心平面的距离。与人眼的视觉系统感知目标深度的机理类似，计算机通过捕捉场景的纹理、遮挡、视差等特征，能够准确计算出目标场景的深度。相较于传统数码相机，光场相机能同时记录光线的位置信息和方向信息，捕获场景完整的光场数据，将传统二维图像扩展到四维。光场相机获取的光场图像为深度估计提供丰富而精确的几何信息支撑，为深度估计精确求解提供了有利条件。微透镜阵列光场成像是一种近年来备受关注的新型单镜头三维成像技术，具有结构简单、一次成像记录光线位置和方向信息、后期数据处理方式多样等特点，可广泛应

用于三维重构与测量、三维测量与识别、虚拟与增强现实等领域。近年来，光场深度估计算法显著提升了场景深度估计的精度，已受到研究者的广泛关注。本文研究利用微透镜光场相机获取的光场图像结合深度学习进行场景深度估计，进一步实现近景快速的三维重建。

1 场景的深度估计方法

深度估计其目的是获得目标与相机之间的距离，输出深度图。基于深度图即可进行三维重建。现有的光场深度估计方法可以分为基于优化的深度估计方法与基于学习的深度估计方法两类：

1）基于优化的光场深度估计方法首先以特定方式估计出场景的初始深度图，然后使用全局优化框架或局部平滑方法来细化深度图。光场图像的本质特征是其具有目标物体多个视角的信息，根据其多视角信息表征方式的不同，可以将现有基于优化的光场深度估计方法分为三种：基于多视角立体匹配的深度估计、基于重聚焦的深度估计、基于EPI（极平面）图像的深度估计。

2）基于学习的光场深度估计方法是利用现有深度学习的框架学习出包含场景深度信息的模型，借助计算机GPU的强大性能，设计各种网络实现深度估计。

由于光场数据的高维性，四维光场数据不能直接应用于现有的深度学习框架中。因此，为了适应网络对输入数据的要求，需进行降维处理，且处理后仍能包含场景空间点的深度关系。相较于光场数据的其他两种表征方式（多视角图像与重聚焦图像），极线图法（Epipolar Plane Image，EPI）切片中的空间几何特性更加直观的反映了场景的深度信息，只需要计算EPI对应斜线的斜率便可得到场景深度信息，并且2D-EPI切片更方便作为卷积神经网络的输入数据。因此现有的基于CNN的光场深度估计框架大多是使用光场EPI块（EPI-Patch）作为网络的输入，根据网络的任务又可分为两类网络实现方式：基于分类任务的光场深度估计和基于回归任务的光场深度估计。基于分类任务的光场深度估计依据数据集的深度范围将深度标签划分为多个类，对每个像素点进行分类。Luo等人[1]设计了两路CNN网络训练垂直和水平方向的EPI切片，结合全局优化方法对输出进行优化，获得最终深度图，没有实现端到端的网络结构。本方法将深度估计问题转化为分类问题，对视差范围小的场景效果较好，对于真实场景中深度连续的情况下，可能会产生输出结果离散、精度降低等问题。Shin et al[2]提出一种四个方向通道的EPI网络，相比较于两个方向的选取，该预处理方式加强了视角保留的信息。同时使用2*2卷积核提取EPI信息，但由于卷积核太小，容易受到噪声影响。Zhou et al.[3]介绍了一种尺度和方向感知的EPI块学习网络，但其深度估计仅基于局部信息。Tsai et al.[4]提出一种基于注意力机制的视图选择网络，该方法通过注意力模块选择对深度图影响较大的子孔径图像，实现了较精确的差异值估计。该类方法实现了端到端的光场深度预测网络且是基于回归任务的网络，但仅基于局部特征估计，深度图易受噪声影响。

2 网络结构

根据现实场景的深度连续性，本文将光场深度获取问题作为回归任务来处理，并在Shin[2]算法的基础上，研究并实现一种端到端的基于EPI和焦点堆栈图像的光场深度估计算法，并基于深度图进行三维重建。与Shin[2]算法中使用的网络不同的是，本次设计在网络中加入注意力机制模块CBAM[5]，加强学习了堆栈图像间的几何关系，并使用较大尺寸的卷积核来提取相对全局的几何特征，有效帮助了网络间的信息流动，提升了网络的整体性能，同时采用了拟合光场数据内部几何关系的数据增强手段以支持网络的训练，网络结构如图1所示。

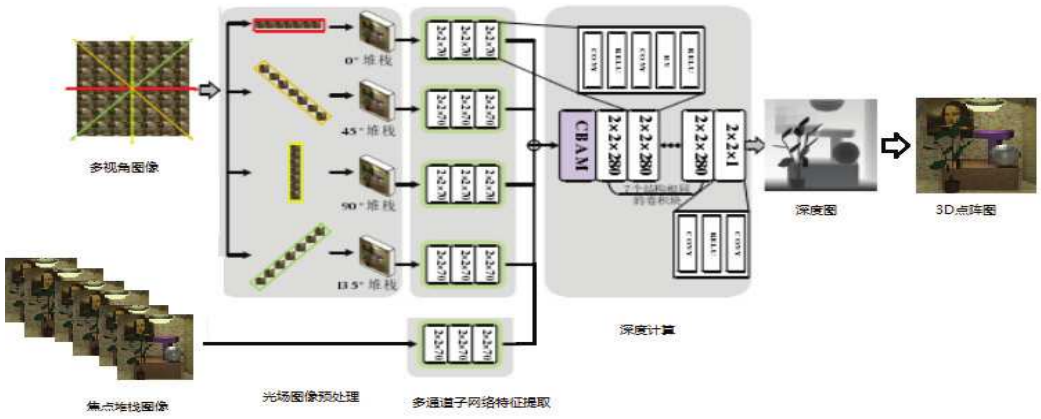


图1 网络结构图

本文使用光场图像的四个方向角（0°、45°、90°、135°）的堆栈图像作为网络的输入，既保证了计算结果的精确性，又降低了计算成本。然后经过一个多通道子网络，分别对0°、45°、90°、135°的EPI块进行网络编码与特征提取。同时增加焦点堆栈图中的信息作为一个通道。由于全卷积网络是能够实现像素级的特征提取，所以该模块由三个全卷积块组成，卷积块结构完全相同，由“Conv-ReLU-Conv-BN-ReLU”组成，用于测量每个局部EPI块中的像素差异。同时为了处理光场图像基线短、视差变化小的问题，卷积块中使用步长为1，尺寸为2×2的小卷积核来进行EPI块中的特征提取。最后是深度计算模块，由三部分组成，第一部分是一个全卷积网络，由7个结构完全相同的卷积块组成。与多通道子网络相同，每个卷积块都是由“Conv-ReLU-Conv-BN-ReLU”构成，用于学习注意力模型传递的特征之间的关系。网络最后一部分由结构为“Conv-ReLU-Conv”的卷积块构成，用于输出亚像素级精度的视差值。最后通过深度估计图进行三维重建。

3 网络训练及评价标准

本次网络训练采用的服务器配置为NVIDIA GeForce GTX 1080 GPU、16GB RAM、Windows64位操作系统，基于TensorFlow架构实现。实验数据集采用的是HCI Old[6]和HCI New[7]光场数据集，两个数据集都由德国海德堡图像处理实验室（简称HCI）实验室推出。HCI Old数据集包含7个合成场景和6个真实场景，每个场景提供81个光场子光圈图像以及真实视差图。HCI New数据集有28个场景都由Blender软件合成，24个场景提供真实视差图，4个不提供。本次网络的训练集从提供真实视差的场景中选择，HCI Old数据集选出10个场景，HCI New数据集中选出20个。其余含有真实视差的场景作为测试集，其余不含有真实视差的场景作为验证集。

网络以为23×23大小EPI图像块作为输入，是由堆栈光场图像随机采样获得。块大小设置为16，学习率为0.1×10^{-4}，每个epoch迭代10000次。为提高速度，训练时卷积不补零。网络模型的损失函数为平均绝对误差（MAE）：

$$\varepsilon(y,y_{gt})=\frac{1}{N}\sum_{i=1}^{N}|(y_i-y_{gt})| \quad （式3.1）$$

式中N表示训练EPI块的个数，y_{gt}为对应像素点的深度标签。

使用均方误差评价算法性能：

$$MSE_{100}=\frac{\sum(d_{GT}-d)^2}{H\times W}\times100 \quad （式3.2）$$

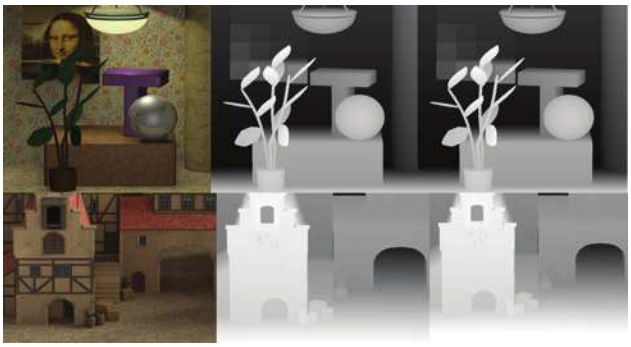
上式中，H和W分别表示图像的高度和宽度，d_{GT}表示真实深度图，d表示深度估计图。

4 实验结果及分析

本文网络分别对合成光场图像和真实场景光场图像进行深度预测并快速三维重建，并对实验结果进行定性和定量分析。

4.1 定性分析

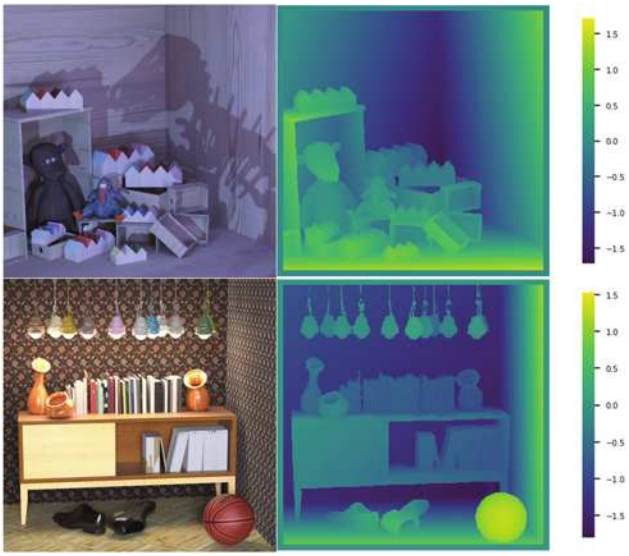
部分HCI Old数据集上本文网络深度估计结果与真实场景深度对比如图2所示。



中心视角图像　　（a）.本文方法　（b）.真实深度值

图2　部分HCI Old数据集本文深度估计结果与真实场景深度对比

本方法主要针对近景三维重建，因此选用结果图为近景图像。从图3可以看出，本文深度估计方法对与近景图像效果较好。因采取多尺度输入和注意力机制，除存在部分噪声影响外，基本轮廓清晰，且深度图亮度对比度接近真实场景深度值。



中心视角图像　　（a）深度估计图

（b）点云图正视图　　（c）点云图侧视图

图3　HCI New数据集上本文网络络深度估计图及其三维重构点云图

HCI New数据集上两个近景图，Dino和Sideboard深度估计结果如图3（a）所示。其中，（b）是三维点云图的正视图，（c）是侧视图。深度估计图中可以看出场景中物体相对位置边缘较清晰。生成的点云图（b）和（c）立体感强，对于纹理的处理也较好。但本文网络对遮挡线索考虑欠缺，且存在部分噪声影响。

4.2 定量分析

在HCI New数据集中深度估计评价函数MSE100数值如表4-1所示。同时使用Lou[1]方法和Shin[2]方法进行对比。表中下划线数值为三种方法中的最优结果。可以看出对于近景图像Dino和Sideboard，本文方法能达到较好的深度估计效果。

表4-1　深度估计MSE_{100}对比

HCI New数据集	Lou[1]方法	Shin[2]方法	本文方法
Backgammon	4.8507	3.6229	3.6578
Dino	0.8743	1.0881	0.7966
Sideboard	1.0861	1.0615	1.0402

表4-2是三种网络的在两个数据集上的每幅图像平均计算时间。下划线部分为最优结果。可以看出基于深度学习且结构的简练Shin[2]的方法时间复杂度最优，远高于传统的Lou[1]方法，虽然本文增加了多尺度聚焦模块和注意力机制模块，但是本文方法与Shin[2]方法所用平均时间十分接近。

表4-2　各种网络深度估计耗时比较（单位：s）

数据集	Lou[]方法	Shin[]方法	本文方法
HCI Old数据集	X	2.55	2.64
HCI New数据集	287	1.63	1.70

5　结语

光场成像具有一次拍摄可实现对三维空间中场景的多维采集，给三维重建等计算机视觉关键问题带来了新的解决方法。尤其是深度学习的应用，极大提高了光场图像的深度估计的处理时间和精确度，进而使得快速三维重建成为可能。本文尝试构建一个深度学习网络对光场图像处理获取深度估计图，并在此基础上实现快速的近景场景的三维重建。深度估计是质量和速度决定了三维重建的精度和速度。但目前仍存在一些问题，如数据集较小，使得网络训练不足；针对主要是朗伯表面，对于镜面反射和折射区域以及遮挡情况欠考虑；下一步工作将对这几个方面进行深入研究。

参考文献：

[1] Luo Y，Zhou W，Fang J，et al. EPI-Patch Based Convolutional Neural Network for Depth Estimation on 4D Light Field[C]// International Conference on Neural Information Processing. Springer, Cham, 2017.

[2] Shin C，Jeon H G，Yoon Y，et al. EPINET: A Fully-Convolutional Neural Network Using Epipolar Geometry for Depth from Light Field Images[J]. IEEE, 2018.

[3] Zhou W，Wei X，Yan Y，et al. A hybrid learning of multimodal cues for light field depth estimation[J]. Digital Signal Processing, 2019, 95(11):102585.

[4] Tsai Y J，Liu Y L，Ming O，et al. Attention-Based View Selection Networks for Light-Field Disparity Estimation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7):12095-12103.

[5] Woo S，Park J，Lee J Y，et al. CBAM: Convolutional Block Attention Module[J]. Springer, Cham, 2018.

[6] Wanner S，Meister S,Goldluecke B.Datasets and benchmarks for densely sampled 4d light field[C]// VMV.2013:225-226.

[7] Honauer K，Johannsen O，Kondermann D，et al. A Dataset and Evaluation Methodology for Depth Estimation on 4D Light Fields[C]// Asian Conference on Computer Vision. Springer, Cham, 2016.