计算机学报
CHINESE JOURNAL OF COMPUTERS

基于光场焦点堆栈的鲁棒深度估计

吉新新[1)]朴永日[1)]张淼[2)] 贾令尧[1)]李培华[1)]

[1)](大连理工大学信息与通信工程学院大连 116024)

[2)](大连理工大学国际信息与软件学院大连 116024)

摘　要　传统的深度估计方法通常利用普通相机记录的二维图像进行单目或多目深度估计。因这种方式仅利用了光线的强度信息，忽略了它的方向信息，其深度估计的效果并不理想。相比之下，光场记录的信息不仅包含了光线的强度信息，还包含了方向信息。因此，基于深度学习的光场深度估计方法越来越引起该领域研究人员的关注，现已成为深度估计方向的研究热点。然而，目前大部分的研究工作从极平面图像（EPI）或子孔径图像着手进行深度估计，而不能有效利用焦点堆栈含有的丰富深度信息。为此，本文提出了基于光场焦点堆栈的鲁棒深度估计方法。本文设计了一种上下文推理单元(CRU)，它能够有效地挖掘焦点堆栈和 RGB 图像的内部空间相关性。同时，本文提出了注意力引导的跨模态融合模块(CMFA)，对上下文推理单元提取的空间相关性信息进行有效融合。为了验证本方法的准确性，在 DUT-LFDD 和 LFSD 数据集上进行了广泛的验证。实验结果表明，本文方法的准确率相比现有的 EPINet 和 PADMM 分别提高了 1.2%和 2.25%。为进一步证明本方法的有效性，我们在现有公开的手机数据集上进行了反复的测试。可视化测试结果表明，本方法在普通消费级手机获取的图像上亦可取得满意的效果，能够适应现实应用场景。

关键词　光场；焦点堆栈；上下文推理单元；注意力机制；跨模态融合模块

中图法分类号 TP391

Robust Depth Estimation via Light Field Focal Stacks

JI Xinxin[1)]　PIAO Yongri[1)] ZHANG Miao[2)]*　JIA Lingyao[1)] LI Peihua[1)]

[1)](School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China)

[2)](International School of Information and Software Engineering, Dalian University of Technology, Dalian 116024, China)

Abstract　Depth estimation is a key issue in the process of 3D reconstruction, and its purpose is to obtain spatial information of 3D objects. The depth information of the scene can help people better understand the geometric structure of the scene, and at the same time provide effective data support for other visual tasks.It has important applications in the fields of scene restoration, action recognition, 3D reconstruction, and saliency detection.The traditional methods of depth estimation usually estimate the depth from monocular or stereo images captured by ordinary cameras. Since the imaging process only considers the intensity information and ignores the direction information of light rays, these methods usually obtain inaccurate depth maps. Compared with the ordinary two-dimensional images, the light field data records not only the intensity but also the direction information of light at one exposure. Therefore, the depth estimation from light field based on deep learning has attracted the attention of researchers and has become a more and more popular researchdirection. As we all know, the focalstack is an important form of light field data. It consists of a series of images focused on different depth planes,which can focus on targets in any depth range, making it easy for the observer to understand the object distribution in the scene. The way of the objects displaying in the focal stack is more in

本课题得到国家自然科学基金(No.61976035)、大连市科技创新基金(No.2019J12GX034)、中央高校基本科研业务费(No.DUT20JC42)资助.吉新新，硕士研究生，主要研究领域为计算机视觉.E-mail: jxx0709@mail.dlut.edu.cn. **朴永日**，博士，副教授，主要研究领域为计算机视觉.E-mail: yrpiao@dlut.edu.cn.**张淼(通信作者)**，博士，副教授，主要研究领域为计算成像.E-mail: miaozhang@dlut.edu.cn.**贾令尧**，博士研究生，主要研究领域为计算机视觉.E-mail:ayaoabc@mail.dlut.edu.cn.**李培华**，博士，教授，主要研究领域为计算机视觉.E-mail: piehuali@dlut.edu.cn

line with the human visual perception mechanism and the way of collection for the focal stack is easier.However, most of the current researches in this direction only estimate depth information from Epipolar plane images (EPI) or sub-aperture images, and cannot effectively utilize the rich depth information contained in the focal stack. Therefore, a robust light field depth estimation based on the focal stack is proposed in this paper. First of all, we design a new context reasoning unit to effectively excavate the internal spatial correlation in RGB images and the focal stack. Specifically,the context reasoning unit fully extracts the multi-focusness information of the focal stack and the structural information of the RGB image by multi graph convolution and multi dilated convolution.Then, we propose an attention-guided cross-modal fusion module to fuse the spatial correlation information extracted by the contextual reasoning unit. Specifically, the attention-guided cross-modal fusion module uses multi-level attention weights to gradually fuse internal spatial related information from the context reasoning unit, so as to realize the maximum contribution of multi-modal features to the depth prediction. In order to verify the accuracy of the proposed method, we conducted extensive experiments on the DUT-LFDD and LFSD datasets in this paper. The results demonstrate that our method achieves superior performance than existing representative methods on two light field datasets. Specifically, the accuracy of our method is 1.2% higher than the existing representative method EPINet and is 2.25% higher than the PADMM. In addition, we test our method on a mobile phone dataset repeatedly, and the visual results show that our method can also obtain outstanding performance on ordinary consumer-level camera data and be adapted to the real scenarios.

Key words lightfield; focal stack; context reasoning unit; attention mechanism; cross-modal fusion module

1 引言

深度估计是三维重建中的关键问题，其目的是获得三维物体的空间信息。场景深度信息可帮助人们更好理解场景的几何结构，同时为其他视觉任务提供有效的数据支持，在场景复原 **Error! Reference source not found.**、动作识别 **Error! Reference source not found.**、三维重建[3]和显著性检测 **Error! Reference source not found.**等领域有着重要应用。

常见的深度估计方法通常从传统相机捕捉的单幅或多幅图像中提取场景深度信息[5-7]。由于传统相机的成像过程仅仅考虑了光线的强度信息，所以基于单张图像的深度估计存在深度不确定性问题。不同于传统成像，光场相机可通过一次捕获同时记录光线的空间和角度信息。其中，角度信息将更好地反映场景深度。因此，基于光场图像的深度估计成为了研究热点[8-9]。传统光场深度估计方法通常构建各种代价函数，提取更准确的深度信息[10-12]。然而，其计算的时间成本相对较高，对先验知识的强依赖性导致其在泛化到不同场景时面临一些困难。

随着深度学习的蓬勃发展，基于卷积神经网络的光场深度估计 **Error! Reference source not found.**3-**Error! Reference source not found.**应运而生并缓解了这一问题。大多数基于卷积神经网络的光场深度估计方法从极平面图像（Epipolar Plane Image，EPI）[15]或子孔径图像[16-18]中捕获深度线索，而较少关注焦点堆栈。焦点堆栈是由一系列聚焦在不同深度的切片图像组成，可聚焦于任意深度范围的目标，使观察者很容易了解场景中沿纵深方向的目标位置信息，这更加符合人类的视觉感知机制。因此，一些研究者将目光投向了光场焦点堆栈。根据全聚焦信息的引入与否，现有基于深度学习的焦点堆栈深度估计方法可分为两类：(1)独立使用焦点堆栈[9,19]，该方法将深度估计视作回归问题，把若干个沿通道维度级联的焦点切片作为输入，通过堆积大量而简单的二维卷积层以计算深度。但是，由于每个像素的深度值计算依赖相邻像素，而且在局部图像信息不明确的非聚焦像素点处，现有方法很难通过有限的感受野预测每一像素的准确深度值。因此，如何捕获长范围上下文信息是需要解决的一大问题。(2)引入全聚焦信息[20]，该方法采用两个独立的支路分别从焦点堆栈和中心视角图像提取深度语义和结构信息，并将中心视角结构信息作为局部引导，通过乘法操作实现两路特征的后期融合。但是，简单的后期融合不能很好地捕获交叉模态之间的互补性。因此，如何有效融合多模态信息对预测结果具有最大贡献是本工作需要考虑的关键问题。本文提出了一种基于多模态信息的鲁棒光场深度估计方法，此方法从焦点堆栈和 RGB 图像（中心视角）提取并融合多模态信息以获得准确的深度信息。

本方法的关键之处主要体现在以下两个方面：首先，基于空洞卷积[21]和图卷积[22]在其他视觉任务上的优势，本文设计了上下文推理单元 (Context Reasoning Unit, CRU)。CRU 巧妙结合了多重空洞卷积和多重图卷积：空洞卷积主要关注场景中的大物体；图卷积充分推理并建模了场景中的对象共现(Object occurrence)，便于有效关注场景中的小物体。结合两者使 CRU 充分捕获上下文信息，全方位探索不同物体和区域之间的内部空间相关性。其次，为实现多模态特征对深度预测的贡献程度，本工作提出了注意力引导的跨模态融合模块(Attention-guided Cross-modal Fusion Module,CMFA)。CMFA 先通过交叉残差连接从对应层级的焦点堆栈和 RGB 特征中捕获互补信息以增强特征；利用多级注意力权重（自注意力权重和关联注意力权重）逐级融合多模态信息。此过程充分融合了 RGB 图像的全局结构化信息和焦点堆栈的丰富深度信息，有效弥补了散焦模糊造成的细节损失。为验证本方法的有效性，本文在 DUT-LFDD[4]和 LFSD[23]数据集上进行了实验。结果表明，本方法的准确率在两个数据集上超过了现有方法的性能，尤其比目前具有代表性的方法 EPINet[24]和 PADMM[25]分别提高了 1.2%和 2.25%。为进一步证明本文方法的实用性，本文在公开手机数据集[26]上进行了测试，结果表明本方法同样获得了比现存方法优越的效果，为其在生活中的实际应用开辟了道路。

本文章节安排如下：第 2 节介绍本文方法的相关工作；第 3 节对方法详细的介绍；第 4 节在多个数据集上进行大量实验，验证本文方法的有效性与实用性；第 5 节为结论部分，对本文方法进行总结。

2 相关工作

2.1 光场深度估计

现有光场深度估计可以分为传统深度估计方法和基于深度学习的深度估计方法。传统深度估计方法基于优化策略来实现深度图的获取。Wanner 等人 **Error! Reference source not found.** 使用结构化张量计算 EPI 的垂直和水平斜率，将深度图估计问题转化为全局优化方法，并使用快速全变分降噪滤波以获得更为鲁棒的视差；Tao 等人[28-29]利用焦点堆栈中多种线索来获得初始深度，并采用马尔科夫随机场（Markov Random Filed，MRF）置信融合原始深度得到全局优化深度图；Anisimov 等人[30]通过对应搜索将有效的立体匹配算法扩展到多视图来估计密集的深度图。虽然传统方法可以根据场景的深度特征准确求解场景的深度，但时间成本相对较高，对先验知识的依赖性强。考虑到这些先验在描述复杂场景时的局限性，传统光场深度估计方法难以泛化到不同场景。

近年来，卷积神经网络[31]凭借其强大的特征提取能力和泛化能力在光场深度估计任务中亦取得了不错的效果。Johannsen 等人[32]采用了一种特殊设计的稀疏分解，有效提取了 EPI 中的深度关系。Luo 等人[33]将深度估计问题看作分类问题，基于 EPI 的方向-深度关系，分别训练垂直和水平方向的 EPI 切片，利用分类器得到分类概率，并结合优化方法得到全局深度。Heber 等人[34]提出了一种编解码网络对 EPI 切片进行编解码来提取深度信息。但是这些方法仅考虑光场图像一个或两个方向的 EPI 切片，导致预测可靠性较低。为此，Shin 等人[24]提出了一种使用光场几何结构进行深度估计的全卷积网络，将四个方向的 EPI 图像分别编码以预测更准确的深度。但此方法在提高估计结果的同时也增加了计算复杂度。为降低计算量，Faluvegi 等人[35]提出了全卷积 3D 神经网络，以轻量化结构从垂直和水平的 EPI 切片中估计视差图；Li 等人[36]提出轻量化多级集成网络，在不同层级采用不同集成方式以从四个方向的 EPI 切片预测场景深度。这些方法专注于从静态场景进行深度估计而没有考虑时间信息。Kinoshita 等人[37]从每一帧的 EPI 体积中提取空间角度信息，并用卷积长短期记忆网络收集时间信息来估计动态场景的深度。少量研究者利用深度学习从焦点切片来进行深度估计。Anwar 等人[8]利用重叠补丁从单个焦点切片中提取深度。Hazirbas 等人[9]首次实现了利用深度学习的方法从焦点堆栈计算深度。Zhou 等人[20]引入中心视角图像作为局部引导，通过简单乘法操作实现后期的特征融合以改善焦点堆栈深度估计的结果。本文方法将焦点堆栈和 RGB 图像作为多模态输入，通过设计的上下文推理单元和注意力引导的跨模态融合模块先后提取并融合多模态特征以获得更加准确的深度信息。

2.2 基于图卷积的推理

近年来，图卷积在图形的关系推理中扮演着重要角色。基于图模型，文献[38]提出了基于 CRF（Conditional Random Field）的网络来进行图像分

割；文献[22]将图卷积网络应用于半监督分类任务；Wang 等人[39]使用图卷积来捕获视频识别任务中物体之间的关系；Chen 等人[40]基于图卷积的推理能力探索区域间关系；Fu 等人[41]基于多尺度重构策略提取空间潜在变量的深度拓扑图信息。因此，基于图卷积，本文设计了一个上下文推理单元，此单元有效结合多重膨胀率的空洞卷积和多重图卷积以提取全面的上下文信息。

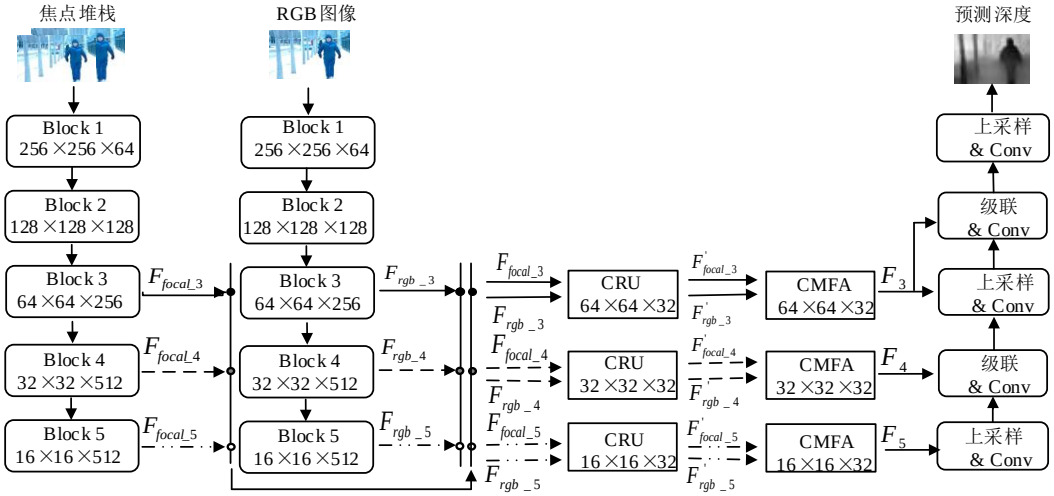


图 1 基于多模态信息的深度估计网络示意图此结构由编码器和解码器两部分组成，编码器部分每一支路包括 5 个卷积块和 3 个 CRU 模块，每一支路的第 3，4，5 层特征将通过一个 CRU 单元；解码器部分包括 CMFA 模块和多个级联、卷积、上采样模块

2.3 注意力机制

人类在接受大量视觉信息时会选择性关注部分信息同时忽略其他可见信息，这种机制被称为注意力机制。在深度学习发展的今天，注意力机制已被广泛地应用于多种视觉任务，如图像分类[42]，单目深度估计[43]和语义分割[44]。Xu 等人[45]提出了一种基于注意力的 CRF，其采用结构化的注意力模型进行不同尺度的特征融合。不同于此工作，本文设计了一个基于多级注意力机制的融合模块，即注意力引导的跨模态融合模块。此模块利用自注意力权重和关联注意力权重逐级融合多模态特征，充分考虑每一特征对预测深度图的贡献。

3 本文方法

本节重点讨论如何有效利用 RGB 图像和焦点堆栈来预测深度。3.1 节简要介绍整体网络结构；3.2 节详细说明了上下文推理单元（CRU）及其子模块的具体实现；3.3 节详细介绍了注意力引导的跨模态融合模块（CMFA）。最后，3.4 节介绍了训练时采用的损失函数。

3.1 整体网络结构

整体网络结构由编码器和解码器两部分组成，旨在从焦点堆栈和 RGB 图像中提取特征并将多模态特征有效融合。如图 1 所示，编码器部分采用对称的双流网络提取特征：焦点堆栈流和 RGB 流。每个支路均由 VGG-16**Error! Reference source not found.**实现的主干网络和多个 CRU 组成。该主干网络丢弃了最后的池化层和全连接层。解码器部分由本文提出的 CMFA 模块和一个渐进式的解码层组成。

具体来说，给定 RGB 图像 I_0 和含 N 个焦点切片 $\{I_1, I_2 \ldots I_N\}$ 的焦点堆栈，本文设置 N=12。首先将其分别输入相应的支路，由主干网络提取原始焦点堆栈特征和 RGB 特征。然后，将来自焦点堆栈流和 RGB 流的特征 $\boldsymbol{F}_{focal\_i} \in R^{N\times W\times H\times C}$ 和 $\boldsymbol{F}_{rgb\_i} \in R^{1\times W\times H\times C}$ 分别输入 CRU 以提取上下文信息，全方位探索焦点堆栈和 RGB 图像的内部空间相关性。其中 N，W，H 和 C 分别表示切片的数量高度，宽度和通道数，$i\,(i=3,4,5)$ 表示主干网络的第 i 个 Block。在解码器中，将 CRU 的输出特征 $\boldsymbol{F}'_{focal\_i} \in R^{N\times W\times H\times C}$ 和 $\boldsymbol{F}'_{rgb\_i} \in R^{1\times W\times H\times C}$ 作为 CMFA 的输入，通过多级注意力机制融合对应的层级特征。最后，自顶向下地逐层对多级特征 $\boldsymbol{F}_i \in R^{1\times W\times H\times C}$ 进行解码，并对预测结果图用真值进行监督。整体实现的伪代码如算法 1 所示。

算法 1. 基于多模态信息的鲁棒光场深度估计

输入：焦点堆栈{ I_1， $I_2 \ldots I_N$ }、RGB 图像 I_0 和真值 G

输出：预测深度图 $\boldsymbol{D}$

1. 对焦点堆栈和 RGB 图像进行特征提取，得到每一支路的原始特征 $\boldsymbol{F}_{focal\_i}$ 和 $\boldsymbol{F}_{rgb\_i}$ $(i=3,4,5)$

2. FOR i=3 to 5

3. 通过 CRU 分别从两支路的第 i 层原始特征提取上下文信息 $\boldsymbol{F}'_{focal\_i}$ 和 $\boldsymbol{F}'_{rgb\_i}$：

$$\boldsymbol{F}'_{focal\_i} \leftarrow CRU(\boldsymbol{F}_{focal\_i})\ \boldsymbol{F}'_{rgb\_i} \leftarrow CRU(\boldsymbol{F}_{rgb\_i})$$

4. 通过 CMFA 分别集成来自 CRU 第 i 对的成对特征，得到融合特征 F_i：

$$\boldsymbol{F}_i \leftarrow CMFA(\boldsymbol{F}'_{focal\_i}, \boldsymbol{F}'_{rgb\_i})$$

5. END FOR

6. $\boldsymbol{F}_3$, $\boldsymbol{F}_4$, $\boldsymbol{F}_5$ 依次经过上采样&卷积、级联&卷积进行解码，得到预测深度 $\boldsymbol{D}$

7. 利用真值 $\boldsymbol{G}$ 和代价函数 loss 优化深度预测 $\boldsymbol{D}$：

$$\boldsymbol{D} \leftarrow loss(\boldsymbol{G}, \boldsymbol{D})$$

3.2上下文推理单元（CRU）

由于每个像素的深度值与其相邻像素密切相关，当局部图像信息不明确时，有限的感受野限制了每一像素预测深度值的准确性[41]，一些方法采用空洞卷积来增大感受野[6,47]。考虑到每一焦点切片均包含聚焦区域和非聚焦区域，非聚焦区域在表现深度信息的同时也带来了局部信息不明确的问题，因此如何有效捕获上下文信息以全方位探索内部空间相关性对后续操作非常重要。为此，我们提出了一个上下文推理单元（CRU）。CRU 不仅可以通过多重空洞卷积来关注场景中的大物体；还通过多重图卷积充分推理并建模场景中的对象共现，便于关注场景中的细小物体。

如图 2 所示，CRU 模块由三个分支组成。顶部是一个短连接操作，由一个 1×1 的卷积运算实现，可有效学习残差信息 $\boldsymbol{F}_{res}$；中间分支是多重空洞卷积操作；而底部分支是多重图卷积操作。对于输入的原始焦点堆栈特征 $\boldsymbol{F}_{focal\_i}$ 或 RGB 特征 $\boldsymbol{F}_{rgb\_i}$，以 $\boldsymbol{F}_{focal\_i}$ 为例，此单元将中间分支的输出特征 F_{md} 和底部分支的输出特征 $\boldsymbol{F}_{mg}$ 进行级联并卷积得到特征 $\boldsymbol{F}_f = Conv(Cat(\boldsymbol{F}_{mg}, \boldsymbol{F}_{md}))$，其中 $Conv$ 表示卷积操作，Cat 表示级联操作。然后，将该特征与顶部分支的特征 $\boldsymbol{F}_{res}$ 相加以获得最终的细化焦点堆栈特征 $\boldsymbol{F}'_{focal\_i} = \boldsymbol{F}_{res} + \boldsymbol{F}_f$。对于 RGB 图像，则得到细化的 RGB 特征 $\boldsymbol{F}'_{rgb\_i}$。

如图 2 所示，多重空洞卷积由跨通道学习器和空洞空间卷积池化金字塔组成。此模型可以通过 1×1 卷积运算学习复杂的通道交互，并通过膨胀率为 3、5、7 的空洞卷积来捕捉不同的图像上下文信息。生成的特征受大物体支配，可有效地对较大物体之间的空间相关性进行建模。

多重图卷积沿网络特征的通道维度建模图像中的长范围上下文信息。与以前采用图像级编码器

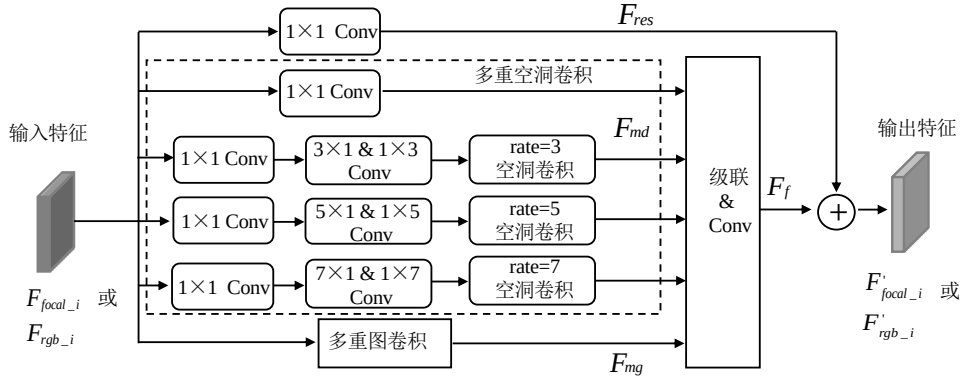


图 2 上下文推理单元（CRU）结构示意图

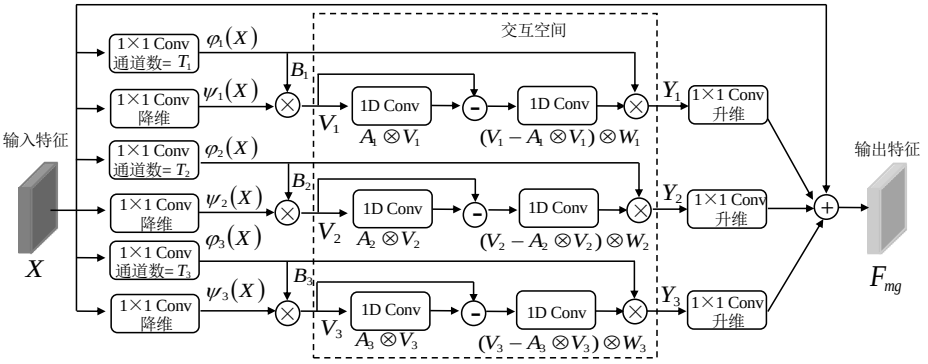


图 3 多重图卷积结构示意图

的方法（例如纯全连接层 **Error! Reference source not found.** 或全局平均池 **Error! Reference source not found.**）不同，此设计可以利用更少的参数和更多的节点对同一特征并行建立多级节点拓扑图。与文献 **Error! Reference source not found.**相比，本文的设计从不同规模进行区域覆盖，从而有效地对区域级线索的信息进行建模和通信。图中节点的数量根据输入特征的空间大小而动态变化。因此，该网络能够改善不同区域之间的空间关系，充分考虑图像中不同对象之间和不同区域之间的差异，以更好地适应场景中的小物体。

此子模块的具体实现如图 3 所示。具体来说，以原始焦点堆栈特征 $\boldsymbol{F}_{focal\_i}$ 为例。给定输入特征 $\boldsymbol{X}=\boldsymbol{F}_{focal\_i}$ 并将其变形为 $\boldsymbol{X}\in R^{N\times L\times C}$（$L$=$W$×$H$），此模块通过三个并行分支建立三个节点拓扑图，以细化空间关系。在第 $j\,(j=1,2,3)$ 个分支中，其实现过程可以分为以下三个步骤：

（1）空间映射：利用映射函数将输入特征 $\boldsymbol{X}$ 从坐标空间映射到交互空间。为减少计算代价，首先使用 $\psi_j(\boldsymbol{X})$ 对 $\boldsymbol{X}$ 进行降维并形式化投影函数 $\varphi_j(\boldsymbol{X})=\boldsymbol{B}_j$，$\boldsymbol{B}_j\in R^{N\times T_i\times L}$。事实上，$\psi_j(\boldsymbol{X})$ 由具有 $(C_j<C)$ 通道的 1×1 卷积层实现，$\varphi_j(\boldsymbol{X})$ 由具有 $T_j=L/t^{j+1}$ 通道的 1×1 卷积层实现，本文将 t 的取值设为 2。因此，输入特征 $\boldsymbol{X}$ 被投影到交互空间的新特征 $\boldsymbol{V}_j\in R^{N\times T_j\times C_j}$。$\boldsymbol{V}_j$ 整合了焦点堆栈每一切片不同区域的信息。注意，T_j 表示拓补图中的节点数，其根据原始特征的空间大小动态变化。这种设计有助于网络从不同的尺度适应特征。具体而言，每一新特征 $\boldsymbol{V}_j$ 的生成如式（1）所示，其中 ⊗ 表示矩阵乘法操作：

$$\boldsymbol{V}_j=\boldsymbol{B}_j\otimes\psi_j\left(\boldsymbol{X}\right)=\varphi_j\left(\boldsymbol{X}\right)\otimes\psi_j\left(\boldsymbol{X}\right) \quad (1)$$

（2）特征图卷积：通过拓补图进行关系推理。投影后，在交互空间中建立一个具有节点 v_j，边 ε_j 和邻接矩阵 $\boldsymbol{A}_j\in R^{N\times T_j\times T_j}$ 的全连接图 $G_j=(v_j,\varepsilon_j,A_j)$。其中每个节点都存储了新的特征。$\boldsymbol{A}_j$ 的每一点的取值 A_{j_pq} 受区域 p 和区域 q 的影响，若区域 p 与区域 q 相邻，则 $A_{j\_pq}=1$，否则 $A_{j\_pq}=0$。因此，上下文推理问题可以被简化为节点之间的交互性捕获问题。此过程通过沿通道和节点方向的两个 1D 卷积实现。利用邻接矩阵和特定层的可训练边缘权重 $\boldsymbol{W}_j$，可以在节点之间传播信息以获得节点特征矩阵 $\boldsymbol{M}_j\in R^{N\times T_j\times C_j}$。那么，$\boldsymbol{M}_j$ 如式（2）所示：

$$\boldsymbol{M}_j=\left(\boldsymbol{V}_j-\boldsymbol{A}_j\otimes\boldsymbol{V}_j\right)\otimes\boldsymbol{W}_j \quad (2)$$

（3）反向映射：利用反向映射函数将特征从交互空间映射到坐标空间。在交互空间完成推理之后，反向映射函数 $\boldsymbol{B}_j^T$ 将新特征 $\boldsymbol{M}_j$ 映射到原始坐标空间，以获得特征 $\boldsymbol{Y}_j\in R^{N\times L\times C_j}$。实现过程如下所示：

$$\boldsymbol{Y}_j=\boldsymbol{B}_j^T\otimes\boldsymbol{M}_j \quad (3)$$

最后，为更好地与现有的卷积神经网络架构兼容并适应残差信息的引入操作，通过三个 1×1 卷积层将每一分支的输出特征 $\boldsymbol{Y}_1$，$\boldsymbol{Y}_2$，$\boldsymbol{Y}_3$ 分别扩展到原始尺寸。并将三个分支的输出特征与原始特征 $\boldsymbol{X}$ 相加以得到最后的特征 $\boldsymbol{F}_{mg}\in R^{N\times L\times C}$，并将其变形为 $\boldsymbol{F}_{mg}\in R^{N\times H\times W\times C}$。具体实现步骤如下所示：

$$\boldsymbol{F}_{mg}=\boldsymbol{X}+Conv\left(\boldsymbol{Y}_1\right)+Conv\left(\boldsymbol{Y}_2\right)+Conv\left(\boldsymbol{Y}_3\right) \quad (4)$$

其中，式（4）中的 *Conv* 表示 1×1 的升维卷积。

对于 RGB 图像，此单元对来自 RGB 流主干网络的原始 RGB 特征 $\boldsymbol{F}_{rgb\_i}$ 执行相同操作。

3.3注意力引导的跨模态融合模块（CMFA）

焦点切片中的散焦模糊在反映场景中物体深度分布的同时可能会导致细节损失，从而对深度图的准确性造成负面影响。为解决这一挑战，本文重点考虑如何有效融合焦点堆栈特征和 RGB 特征。最直接方法是简单地将焦点堆栈特征和 RGB 特征直接相加、相乘或级联，但这不仅忽略了不同焦点切片特征和 RGB 特征对最终结果的相对贡献，而且严重破坏了焦点切片之间的空间相关性。因此，不同于[20]，本文设计了一个注意力引导的跨模态融合模块，通过多级注意力机制有效地融合焦点堆栈的隐式深度信息和RGB图像的全局结构化信息。

如图 4 所示，此模块的实现可以分为两步：（1）捕获补充信息以增强特征；（2）集成增强后的多模态特征。以 CRU 输出的焦点堆栈特征 $\boldsymbol{F}_{focal\_i}^{'}$ 和 RGB 特征 $\boldsymbol{F}_{rgb\_i}^{'}$ $(i=3,4,5)$ 为例，考虑到焦点堆栈特征和 RGB 特征之间的差异，首先对他们进行增强以突出多模态信息的互补性。在第一步中，首先引入简单 3D 卷积和 2D 卷积实现跨模态残差连接，以从对应层级的焦点堆栈特征 $\boldsymbol{F}_{focal\_i}^{'}$ 和 RGB 特征 $\boldsymbol{F}_{rgb\_i}^{'}$ 中捕

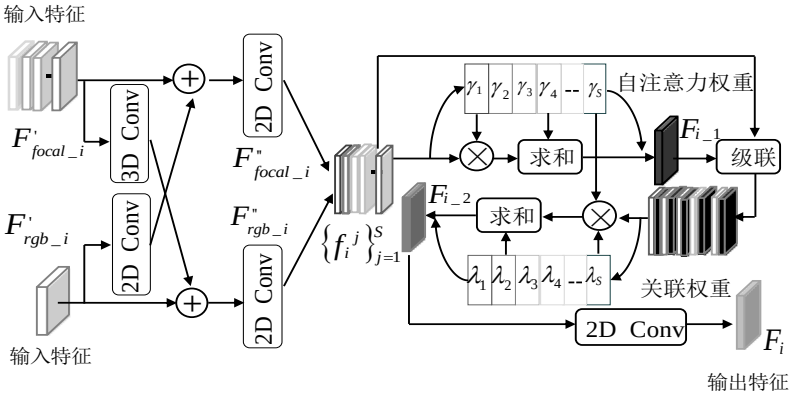


图 4　注意力引导的跨模态融合模块（CMFA）结构示意图

获互补信息并将互补信息分别加到对应的特征中。然后采用一个 1×1 的 2D 卷积深入学习从而获得增强后的特征 $\boldsymbol{F}_{focal\_i}^{''}\in R^{N\times W\times H\times C}$ 和 $\boldsymbol{F}_{rgb\_i}^{''}\in R^{1\times W\times H\times C}$。具体实现如下：

$$\boldsymbol{F}_{focal\_i}^{''}=\boldsymbol{F}_{focal\_i}^{'}+2DConv(\boldsymbol{F}_{rgb\_i}^{'}) \quad (5)$$

$$\boldsymbol{F}_{rgb\_i}^{''}=\boldsymbol{F}_{rgb\_i}^{'}+3DConv(\boldsymbol{F}_{focal\_i}^{'}) \quad (6)$$

注意，使用交叉模态残差连接从对应层级的焦点堆栈特征和 RGB 特征中提取互补信息可以近似等效为残差函数。这种重新定义消除了多模态融合的歧义。式（5）和式（6）中的 $2DConv$ 和 $3DConv$ 分别表示 2D 卷积和 3D 卷积。

在第二步中，将增强的焦点堆栈特征和 RGB 特征通过多级注意力机制分配权重进行融合，以充分考虑每一切片对预测结果的相对贡献。首先，将增强的特征 $\boldsymbol{F}_{focal\_i}^{''}$ 和 $\boldsymbol{F}_{rgb\_i}^{''}$ 沿切片维度进行级联，并将级联后的特征表示为 S 个切片特征的集合 $\{\boldsymbol{f}_i^j\}_{j=1}^{S}$ $(S=13)$。首先，为了专注于每个焦点切片的深度信息和 RGB 的内容信息，为每个切片特征 $\boldsymbol{f}_i^j$ 分配粗略的自注意力权重 γ_j。通过自注意力权重，我们将所有切片特征初步融合以得到全局特征 $\boldsymbol{F}_{i\_1}\in R^{1\times W\times H\times C}$。由于 $\boldsymbol{F}_{i\_1}$ 包含所有焦点堆栈完整的深度信息和 RGB 图像的结构化信息，因此将每个切片特征与全局特征 $\boldsymbol{F}_{i\_1}$ 再次进行关联学习，以引入更可靠的关联注意力权重优化融合结果。如上所述，利用关联注意力权重对所有切片特征进行融合以得到细化特征表示 $\boldsymbol{F}_{i\_2}\in R^{1\times W\times H\times C}$。最后，通过一个简单的卷积层，我们可以获得最终的融合结果 $\boldsymbol{F}_i\in R^{1\times W\times H\times C}$。该过程可以定义为：

$$\gamma_j=\sigma(fc(dropout(avgpool(\boldsymbol{f}_i^j)))) \quad (7)$$

$$\boldsymbol{F}_{i\_1}=\sum_{j=1}^{S}\gamma_j\boldsymbol{f}_i^j\Big/\sum_{j=1}^{S}\gamma_j \quad (8)$$

$$\lambda_j = \sigma(fc(dropout(avgpool(Cat(\boldsymbol{f}_i^j, \boldsymbol{F}_{i\_1}))))) \quad (9)$$

$$\boldsymbol{F}_{i\_2} = \sum_{j=1}^{S} \gamma_j \lambda_j Cat(\boldsymbol{f}_i^j, \boldsymbol{F}_{i\_1}) \Big/ \sum_{j=1}^{S} \gamma_j \lambda_j \quad (10)$$

$$\boldsymbol{F}_i = Conv(\boldsymbol{F}_{i\_2}) \quad (11)$$

这里 σ 表示 *sigmoid* 函数，*avgpool* 表示平均池化，*dropout* 表示随机失活，*fc* 表示全连接操作，*Cat* 表示级联操作，γ_j 表示第 j 个切片的自注意力权重，λ_j 维第 j 个切片的关联注意力权重。此模块有效利用了焦点堆栈和 RGB 图像之间的互补性。

3.4损失函数

为提高预测的准确率，本文采用文献[49]中的代价函数对网络进行优化，此损失函数包括三项：深度损失、梯度损失和表面法线损失。每一损失函数的计算方式如下所示：

$$l_{depth} = \frac{1}{n}\sum_{i=1}^{n} \ln(\left\|\boldsymbol{D}_i - \boldsymbol{G}_i\right\|_1 + \alpha) \quad (12)$$

其中，l_{depth} 为深度误差的对数。其中，$\boldsymbol{D}_i$ 为像素 i 处的预测深度值，$\boldsymbol{G}_i$ 像素 i 处的真值深度值，n 表示真值深度图中像素值不为 0 的有效像素点个数。ln 为取对数操作，α 为超参数。

$$l_{grad} = \frac{1}{n}(F(\nabla_x(\left\|\boldsymbol{D}_i - \boldsymbol{G}_i\right\|_1)) + F(\nabla_y(\left\|\boldsymbol{D}_i - \boldsymbol{G}_i\right\|_1))) \quad (13)$$

其中，$\nabla_x(*)$为沿 x 方向第 i 个像素处$\left\|\boldsymbol{D}_i - \boldsymbol{G}_i\right\|_1$的空间导数。深度梯度损失函数 l_{grad} 可以有效处理由卷积神经网络训练引起的边缘失真问题。

$$l_{normal} = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{\left\langle \boldsymbol{n}_i^D, \boldsymbol{n}_i^G \right\rangle}{\sqrt{\left\langle \boldsymbol{n}_i^D, \boldsymbol{n}_i^D \right\rangle}\sqrt{\left\langle \boldsymbol{n}_i^G, \boldsymbol{n}_i^G \right\rangle}}\right) \quad (14)$$

其中，l_{normal} 为表面法线损失函数，其值取决于预测深度值和真值深度值之间的表面法线精度。$\langle *,* \rangle$ 表示向量的内积，$n_i^G = [-\nabla_x(G_i), -\nabla_y(G_y), 1]^T$ 表示深度真值的表面法线。

故，本文采取的总代价函数如下所示：

$$L = l_{depth} + \lambda l_{grad} + \mu l_{normal} \quad (15)$$

其中，$\lambda, \mu \in \mathrm{R}$ 为权重系数。

4 实验

4.1 数据集介绍

本节使用了 2 个光场数据集 DUT-LFDD[4]、LFSD[23] 和 1 个手机数据集[26]对基于多模态信息的深度估计方法进行性能评估。所有消融实验是在 DUT-LFDD 和 LFSD 数据集上进行的。

4.1.1 DUT-LFDD 数据集

此数据集共包含 1462 个真实世界的光场样本，这些样本均由 Lytro 相机捕获。每个场景包含多种图片类型：如 RGB 图像（中心视角图像），焦点堆栈图像和相应的地面真值深度图。在实验过程中，本文从此数据集中选择 967 个包含 12 个焦点切片的焦点堆栈，并且以其中的 630 个样本进行训练，其余 337 个样本进行测试。先前研究表明，数据增强可以有效避免过拟合问题。因此，本文从以下几个方面对数据进行了增强：（1）以 50%的概率随机水平翻转输入图像；（2）在一定的角度范围[-5，5]内随机旋转所有图片；（3）在范围[06，1.4]内通过均匀采样实现亮度，对比度和饱和度的随机变化以对图片进行色彩变换。

4.1.2 LFSD 数据集

该数据集由 Li 等人提出。它包含了 Lytro 相机捕获的 100 个场景，其中 60 个为室内场景，40 个为室外场景。每个场景都包含一幅 RGB 图像、若干个焦点切片组成的焦点堆栈和一个深度图。为适应网络参数，本文选取了焦点切片数为 12 的焦点堆栈进行对比实验。

4.1.3 手机数据集

手机数据集是由 Samsung Galaxy 手机通过自动聚焦捕获的。每个场景都包含一系列聚焦在不同深度的焦点切片。此数据集共包含 13 个场景，如植物，水果，窗户等。每个图像的大小为 640×340。

4.2 实验设置

4.2.1 实现细节

本次实验采用的主机配置如下：CPU 为 Intel Core i7-8700，3.20GHz，16G 内存，GPU NVIDIAGTX2080Ti。本文算法由 Pytorch 深度学习工具包实现。RGB 流和焦点堆栈流的主干网络均采用在 ImageNet 数据集上训练的参数进行初始化，其他模块进行随机初始化。本文在 DUT-LFDD 的训练集上进行网络训练。本文 RGB 流输入的图像大

小为$1\times256\times256\times3$，焦点堆栈流输入的图像大小为$12\times256\times256\times3$。网络优化采用自适应矩估计算法，训练过程中将学习率初始化为 le-4, 迭代 30 个周期后将其调整为 3×le-4，并且继续迭代 20 个周期。考虑到输入数据的尺寸与数量，将训练过程中的 batchsize 设置为 1。通过相关任务的经验和反复尝试，本文最终设置梯度损失l_{grad}的权重$\lambda=1$，表面法线损失l_{norma}的权重$\mu=1$，深度损失l_{depth}中的$\alpha=1$。

4.2.2 评估指标

为全面评估各种方法，本文采用了深度估计中常用的评价指标，即：均方根误差（Root Mean Square Error，RMSE），平均相对误差绝对值（Average Absolute Relative Error, Abs Rel），相对均方根误差（Square Relative Error，SquRel），精度为门限$\delta_i=1.25^i$ $(i=1,2,3)$的准确率。具体表示如下所示：

- RMSE: $\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{D}_i-\boldsymbol{G}_i\right\|^2$ （16）
- Abs Rel: $\frac{1}{n}\sum_{i=1}^{n}\frac{\left\|\boldsymbol{D}_i-\boldsymbol{G}_i\right\|}{\boldsymbol{G}_i}$ （17）
- SquRel: $\frac{1}{n}\sum_{i=1}^{n}\frac{\left\|\boldsymbol{D}_i-\boldsymbol{G}_i\right\|^2}{\boldsymbol{G}_i}$ （18）
- 准确率：$\delta_i=\frac{card(\{D_i:\max\{\frac{\boldsymbol{D}_i}{\boldsymbol{G}_i},\frac{\boldsymbol{G}_i}{\boldsymbol{D}_i}\}<1.25^i\})}{card(\{\boldsymbol{D}_i\})}$ （19）

其中，$\boldsymbol{D}_i$为像素 i 的预测深度值，$\boldsymbol{G}_i$为像素 i 的真值深度值，n 表示真值深度图中的有效像素点个数。

4.3 消融实验

4.3.1 基线网络模型

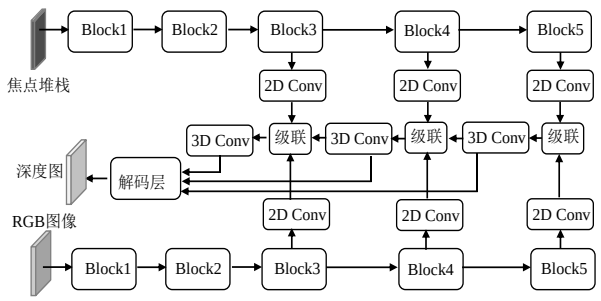


图 5 基线网络模型示意图

如图 5 所示，本文的基线网络由焦点堆栈流和 RGB 流两部分构成。每一支路均采取 VGG-16 作为主干网络以提取特征。相比较本文方法，基线网络采用 2D Conv 和 3D Conv 分别替代提出的 CRU 和 CMFA 模块。为保证公平性，每个 2D Conv 或 3D Conv 均由 6 个普通 2D 卷积层或 3D 卷积层实现。此设计充分保证基线网络具备较高的表达能力。此网络引入普通的级联操作逐层融合对应层级的焦点堆栈特征和 RGB 特征。最后通过解码层预测深度信息。

4.3.2 多模态输入的有效性分析

为证明多模态输入的有效性，本节分别从三个方面进行了消融对比实验：仅输入 RGB 图像、仅输入焦点堆栈、联合输入焦点堆栈和 RGB 图像。具体而言，焦点堆栈流和 RGB 流分别利用基线网络模型中对应的支路实现，并将其分别表示为“焦点堆栈流”和“RGB 流”。如表 1 和表 2 所示，相比较单模态输入的“RGB 流”和“焦点堆栈流”，基线网络的误差显著下降。图 6 为本文方法在 DUT-LFDD 数据集上的消融实验结果。图 6 结果表明，焦点堆栈和 RGB 图像的结合实现了更准确的预测结果，深度图中的信息更加完整。

4.3.3 上下文推理单元的有效性

CRU 可以从焦点堆栈特征和 RGB 特征中提取上下文信息以探索内部空间相关性。CRU 模块既可以有效地推理出焦点切片中聚焦区域和非聚焦区域的空间关系，又可以捕捉 RGB 图像中不同区域的结构化关联。为了验证 CRU 的有效性，我们使用 CRU 代替基线网络中的 2D Conv 并将其表示为“+CRU”。表 1 和表 2 展示了本文方法在 DUT-LFDD 和 LFSD 数据集上的消融实验结果。如表 1 和表 2 所示，在所有评估指标上，“+CRU”显著胜过基线网络。为了直观地表明 CRU 单元的有效性，本文将使用各模块预测的深度图在图 6 中进行了展示。如图 6 所示，相比较基线网络，CRU 模块更好的关注场景中的细小物体，使得细小物体处的深度变化更加明显。

4.3.4 注意力引导的跨模态融合模块的有效性分析

本文提出的 CMFA 模块可以有效地融合焦点堆栈和 RGB 图像的信息，从而补偿由散焦模糊造成的细节损失。为证明此融合方法的有效性，本小节采用 CMFA 代替基线模型中对应层级特征间的简单级联和暴力 3D 卷积操作，并记作“+CMFA”。如表 1 所示，相比基线网络，“+CMFA”在 RMSE 上性能提高了 9.8%；如表 2 所示，相比基线网络，

“+CMFA”在 RMSE 上性能改善了 9.5%。这主要归功于多级注意力机制充分考虑了不同模态信息对深度预测的贡献。图 6 结果表明，“+CMFA”有效提高了深度图的质量，保留了更多结构信息。

RGB 图像 深度真值 RGB 流焦点堆栈流 基线网络 +CRU +CMFA 本文方法

图 6 在 DUT-LFDD 数据集上的消融实验结果比较

4.3.5 CRU 和 CMFA 联合使用的有效性分析

为证明上下文推理单元和注意力引导的跨模态融合模块的联合使用可以有效提取并融合多模态信息，本文在最终方案中将其集成。同时，本小节进一步证明 CRU 每一子模块的有效性。为了便于表示，将 CRU 中的多重空洞卷积和多重图卷积操作分别记作“md”和“mg”。相比 4.3.3 节的“+CRU”和 4.3.4 节的“+CMFA”，联合使用“md”和“CMFA”即“md+CMFA”，与联合使用“mg”和“CMFA”模块即“mg+CMFA”均显著提高了性能。如表 1 和表 2 所示，相比基线网络，本文方法使得均方根误差 RMSE 在 DUT-LFDD 数据集上降低了约 0.7，在 LFSD 数据集上降低了约 0.9。这说明每一子模块均可以有效地实现各自的功能，且不会由于组合使用而相互干扰。从图 6 可知，本文方法预测的深度图具有更加明显且准确的深度变化，物体的边缘更加清晰，内部信息更加完整。

4.3.6 损失函数各项成分的有效性分析

为了证明本文采用的损失函数中各项成分的有效性，本小节采用不同的损失函数对提出的网络进行了训练。图 7 展示了采用不同损失函数训练后的模型预测的深度图。如图 7 所示，仅仅采用 l_{depth} 训练的网络预测的深度图更加模糊。而且，本文的损失函数提供了显着的改进，尤其是对于场景中物体的精细细节和边界的估计。表 3 展示了采用各项损失函数的数值结果。表 3 的各项指标表明，分别使用深度损失 l_{depth}，深度损失+梯度损失 $l_{d\ e}+_p\lambda l_{g\ hr}$ 和总损失函数 $l_{d\ e}+_p\lambda l_{g\ hr}+_a\mu l_{h\ o}$ 的模型逐步获得了场景更准确的细节信息。

表 1 在 DUT-LFDD 数据集上的消融实验结果

方法	误差	准确率


RGB 图像


深度真值


焦点堆栈流


基线网络

		Abs Rel	SquRel	δ_1	δ_2	δ_3
焦点堆栈流	.3856	.1977	.1140	.6520	.9164	.9867
RGB 图像	深度真值	.1830	.0978	.6927	.9298	.9890
基线网络	.3720	.1727	.0899	.7020	.9373	.9919
+md+CMFA		.1616	.0793	.7431	.9546	.9945
深度损失+梯度损失		.1578	.0767	.7488	.9551	.9943
+mg+CMFA	.3420	.1533	.0738	.7672	.9586	.9943
本文方法	总损失	.1493	.0697	.7757	.9644	.9945
	.3134	.1455	.0668	.7859	.9685	.9956
	.3029					

损失函数	RMSE	Abs Rel	δ_1
深度损失	.3185	.1509	.7708
深度损失+梯度损失	.3139	.1499	.7733
总损失	**.3029**	**.1455**	**.7859**

表 2　在 LFSD 数据集上的消融实验结果

图 7　采用不同损失函数训练的网络在 DUT-LFDD 数据集上的深度图

表 3　采用不同损失函数训练的网络在 DUT-LFDD 数据集上的结果

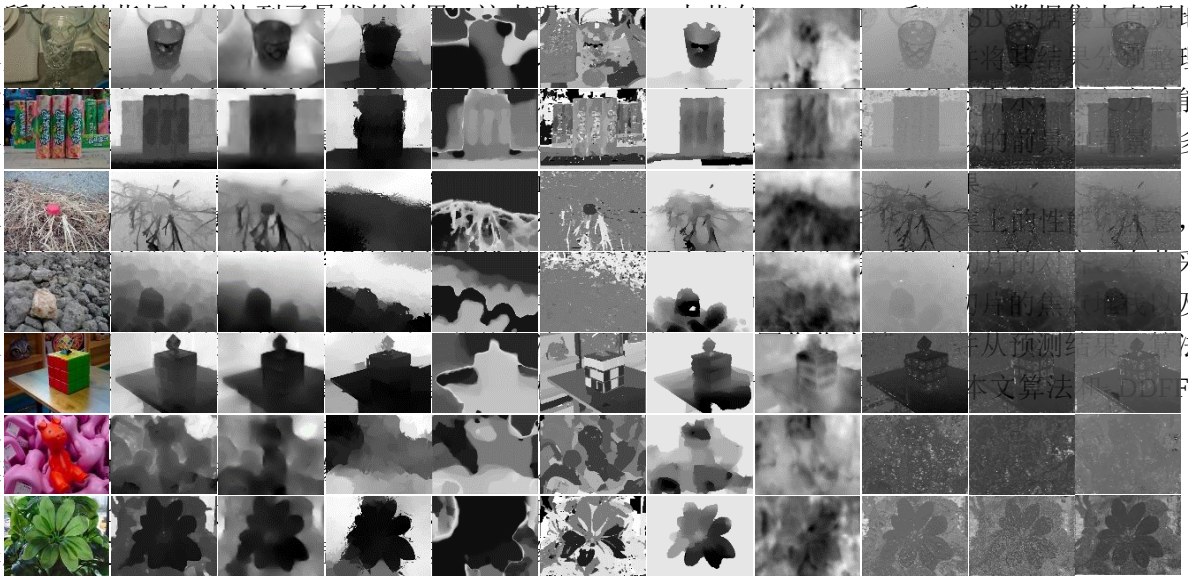
4.4　本文方法与其他方法的比较

本节将本文方法与其他方法进行比较，这些方法涵盖了基于深度学习的方法（DDFF[9]，EPINet[24]，3DConv[35]，MANet[36]）和以*标记的基于非深度学习的方法（PADMM*[25]，LF_OCC*[32]，VDFF*[50]，LF*[51]）。为实现公平比较，在复现对比方法时，本文使用作者提供的参数并对其做出相应调整以适应不同的据集。而且，由于部分方法的代码没有公布，故无法与其进行实验对比。

方法	误差			准确率		
	RMSE	Abs Rel	SquRel	δ_1	δ_2	δ_3
RGB 流	.4637	.2098	.1286	.6013	.8855	.9797
焦点堆栈流	.4106	.1821	.1000	.6757	.9198	.9885
基线网络	.4029	.1791	.0957	.6785	.9327	.9899
+CRU	.3727	.1660	.0833	.7136	.9420	.9951
+CMFA	.3647	.1622	.0807	.7257	.9412	.9937
+md+CMFA	.3503	.1566	.0753	.7513	.9546	.9947
+mg+CMFA	.3316	.1490	.0669	.7718	.9658	.9964
本文方法	.3167	.1426	.0627	.7814	.9686	.9964

4.4.1 定量比较

本小节将本文方法在DUT-LFDD和LFSD数据集上与其他深度估计方法进行了比较，实验结果如表 4 所示。表 4 中的结果可以分为两部分，首先是本文方法与其他深度估计方法在 DUT-LFDD 数据集上的比较结果，其次是在 LFSD 数据集上的比较结果。由于 LFSD 数据集中缺少多视角图像，故部分方法在此数据集上没有结果展示。如表 4 所示，与其他方法相比，在 DUT-LFDD 和 LFSD 数据集上本文方法在[illegible]地比了本[illegible]理到的方[illegible]能够使得[illegible]多物体[illegible]，在分利[illegible]采用突出[illegible]及对级上[illegible]法复预测[illegible]方比较对模且相其深[illegible]



RGB 图像 深度真值 本文方法 PADMM* VDFF* LF_OCC* LF* DDFF EPINet 3DConv MANet

的优化将是我们将来所要开展的重点工作。

4.4.3 在手机数据集上的结果

为证明本文方法的实用性，本节在 Intel core i7-8700@3.20GHz 的单核 CPU 主机上评估了本文与 DDFF 方法相比，本文方法成功引入 RGB 图像以辅助焦点堆栈，从而显著改善了深度图质量。得益于提出的上下文推理单元和注意力引导的跨模态融合模块，本文方法获得了比 EPINet 更准确的深度预测结果图。而且，虽然 DUT-LFDD 数据集的深度范围分布较大且不均衡，但是相比较 3DConv 和 MANet，本文方法在此数据集上的结果仍然占据优势，获得了更准确的深度。

4.4.2 定性比较

类型	方法	误差			准确率		
		RMSE	Abs Rel	SquRel	δ_1	δ_2	δ_3
DUT-LFDD	DDFF	.5255	.2666	.1834	.4944	.8202	.9667
	3DConv	.6379	.2817	.2056	.3663	.6747	.9297
	MANet	.4607	.1922	.1044	.5709	.9274	.9947
	EPINet	.4974	.2324	.1434	.5010	.8375	.9837
	VDFF*	.7192	.3887	.3808	.4040	.6593	.8505
	PADMM*	.4730	.2253	.1509	.5891	.8560	.9577
	LF*	.6897	.3835	.3790	.4913	.7549	.8783
	LF_OCC*	.6233	.3109	.2510	.4524	.7464	.9127
	本文方法	**.3029**	**.1455**	**.0668**	**.7859**	**.9685**	**.9956**
LFSD	DDFF	.4255	.2128	.1204	.6185	.8916	.9860
	VDFF*	.7547	.3320	.2660	.4730	.7823	.9359
	PADMM*	.4238	.2153	.1336	.6536	.8880	.9770
	本文方法	**.3167**	**.1426**	**.0627**	**.7814**	**.9686**	**.9964**

图 8 本文方法和其他方法在 DUT-LFDD 数据集上的预测深度图

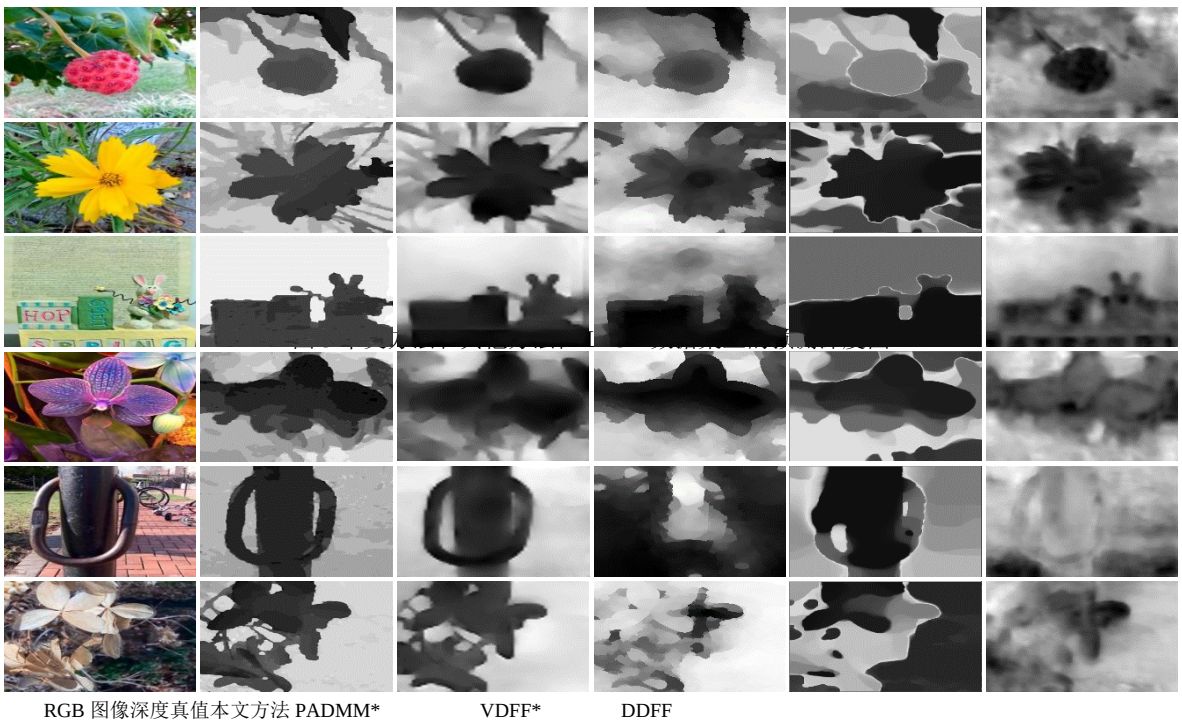


图 10 本文方法和 DDFF 方法在手机数据集上的预测深度

为方便比较，我们将每一场景的所有焦点切片和其RGB 图像视作 1 帧。表 5 展示了本文方法和 DDFF 方法在手机数据集下的帧率和网络参数。图 10 展示了本文方法与 DDFF 方法在此数据集上的预测深度图。显然，相比 DDFF 方法，本文方法具有更小的帧率。同时，图 10 的可视化结果也表明，本文方法相比于 DDFF 方法获得了更优的预测结果，捕获了更多的细节信息。这证明了本文方法在手机数据集上的适用性以及相对其他方法的优越性，为其在日常生活中的实际应用开辟了道路。为进一步提升本文方法的实时性，我们将在后期研究中重点考虑网络结构的优化问题，实现该方法在手机端的应用。

表 5　本文方法和 DDFFF 方法在手机数据集上的帧率和网络参数

方法	帧率	模型参数
DDFF	6 fps	304M
本文方法	5 fps	496M

5 结论

本文设计了一种结合深度学习和焦点堆栈的光场深度估计方法。此方法提出了两个有效模块，并将其嵌入到现有的网络模型中，充分利用焦点堆栈和 RGB 图像，通过端到端的网络训练获得鲁棒的预测深度图。本文方法的创新之处具体表现在以下两个方面：（1）提出了有效的上下文推理单元（CRU）以有效提取上下文信息，探索焦点堆栈和RGB 图像的内部空间相关性；(2) 提出了注意力引导的跨模态融合模块（CMFA）以有效融合从焦点堆栈和 RGB 图像中提取的空间相关性信息，从而有效解决了焦点切片散焦模糊带来的细节丢失问题。本文在两个光场数据集 DUT-LFDD 和 LFSD 上进行了大量实验，实验结果证明了所提模块的有效性。而且，相比现有的方法，本文方法在光场数据集上和手机数据集均实现了最优的性能，为其在日常生活中的实际应用开辟了道路。

致　谢　感谢国家自然科学基金(No.61976035)和大连市科技创新基金(No.2019J12GX034)的资助。感谢审稿专家和编辑在百忙之中审阅本文!

参 考 文 献

[1] LIN Jin-Hua, YAO Yu, WANG Ying. Scene restoration and semantic classification network using depth map and discrete pooling technology. ACTA AUTOMATICA SINICA, 2019, 45(11): 2178-2186(in Chinese)
(林金花, 姚禹, 王莹. 基于深度图及分离池化技术的场景复原语义分类网络. 自动化学报, 2019, 45(11): 2178-2186)

[2] Qian Yin-Zhong, Shen Yi-Fan. Hybrid of pose feature and depth feature for action recognition in static Image. ACTA AUTOMATICA SINICA, 2019, 45(3): 626-636(in Chinese)
(钱银中, 沈一帆. 姿态特征与深度特征在图像动作识别中的混合应用. 自动化学报, 2019, 45(3): 626-636)

[3] Herbort S, Wöhler C. An introduction to image-based 3d surface reconstruction and a survey of photometric stereo methods. 3D Research, 2011, 2(3): 1-17

[4] Piao Y, Li J, Ji W, Zhang M, et al. Memory-oriented decoder for light field salient object detection//Proceedings of the 33rd Neural Information Processing Systems. Vancouver,Canada, 2019: 1-11

[5] Xu D, Ricci E, Ouyang W, et al. Monocular depth estimation using multi-scale continuous crfs as sequential deep networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018, PP(99):1-1

[6] Fu H, Gong M, Wang C, et al. Deep ordinal regression network for monocular depth estimation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 2002-2011

[7] Mahjourian R, Wicke M, Angelova A. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints// Proceedings of the IEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 5667-5675

[8] Anwar S, Hayder Z, Porikli F. Depth estimation and blur removal from a single out-of-focus image//Proceedings of British Machine Vision Conference. London, UK，2017: 1-13

[9] Hazirbas C, Soyer S G, Staab M C, et al. Deep depth from focus// Proceedings of Asian Conference on Computer Vision. Perth, Australia, 2018: 525-541

[10] Kim, M J, Oh T H, Kweon I S. Cost-aware depth map estimation for lytro camera//Proceedings of the IEEE International Conference on Image Processing.Paris, France, 2014: 36-40

[11] Heber S, Yu W, Pock T. Neural epi-volume networks for shape from light field//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2252-2260

[12] Tomioka T, Mishiba K, OyamadaY et al. Depth map estimation using census transform for light field cameras. IEICE Transactions on Information and Systems, 2017, 100(11): 2711-2720

[13] Jin J, Hou J H, Yuan H, et al. Learning light field angular super-resolution via a geometry-aware Network//Proceedings of AAAI Conference on Artificial Intelligence. New York, USA, 2020: 11141-11148

[14] Zhou W, Liang L, Zhang H, et al. Scale and orientation aware

epi-patch learning for light field depth estimation//Proceedings of International Conference on Pattern Recognition. Beijing, China, 2018: 2362-2367

[15] Leistner T, Schilling h, R Mackowiak, et al. Learning to think outside the box: Wide-baseline light field depth estimation with epi-shift//Proceedings of the International Conference on 3D Vision. Quebec City, Canada, 2019: 1-11

[16] Anisimov Y, Wasenmuller O, Stricker D. A compact light field camera for real-time depth estimation//Proceedings of the 18th International Conference on Computer Analysis of Images and Patterns. Salerno, Italy, 2019:1-12

[17] Jeon H G, Park J, Choe G, et al. Depth from a light field image with learning-based matching costs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2):297-310

[18] Peng J, Xiong Z, Wang Y, et al. Zero-shot depth estimation from light field using A convolutional neural network. IEEE Transactions on Computational Imaging, 2020, 6:682-696

[19] Song G, Lee K M. Depth estimation network for dual defocused images with different depth-of-field//Proceedings of the IEEE International Conference on Image Processing. Athens, Greece, 2018:1563-1567

[20] Zhou W, Zhou E, Yan Y, et al. Learning depth cues from focal stack for light field depth estimation//Proceedings of the IEEE International Conference on Image Processing. Taibei, China, 2019: 1074-1078

[21] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848

[22] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks//Proceedings of the International Conference on Learning Representations. Toulon, France, 2017:1-14

[23] Li N, Ye J, Ji Y, et al. Saliency detection on light field//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, Ohio, 2014: 2806-2813

[24] Shin C, Jeon H G, Yoon Y, et al. Epinet: A fully convolutional neural network using epipolar geometry for depth from light field images//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4748-4757

[25] Javidnia, H, Corcoran P. Application of preconditioned alternating direction method of multipliers in depth from focal stack. Journal of Electronic Imaging, 2018, 27(2): 019-023

[26] Suwajanakorn S, Hernandez C, Seitz S M. Depth from focus with your mobile phone//Proceedings of the IEEE Conference Computer Vision and Pattern Recognition. Boston, USA, 2015: 3497-350

[27] Wanner S, Goldluecke B. Variational light field analysis for disparity estimation and super-resolution. IEEE transactions on pattern analysis and machine intelligence, 2013, 36(3): 606-619

[28] Tao, M W, Hadap, S, Malik, J, et al. Depth from combining defocus and correspondence using light-field cameras//Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia, 2013:673-680

[29] Tao M W, Srinivasan P P, Malik J, et al. Depth from shading, defocus, and correspondence using light-field angular coherence//Proceedingsof the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1940-1948

[30] Anisimov Y, Oliver W, Stricker D. Rapid light field depth estimation with semi-global matching//Proceedings of International Conference on Intelligent Computer Communication and Processing. Tokyo, Japan, 2019: 1-8

[31] Zhang Shun, Gong Yi-Hong, Wang Jin-Jun. The development of deep convolutional neural network and its applications on computer vision. Chinese Journal of Computers, 2019, 42(3): 453-482 (in Chinese) (张顺，龚怡宏，王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用. 计算机学报, 2019, 42(3): 453-482)

[32] Johannsen O, Sulc A, Goldluecke B. What sparse light field coding reveals about scene structure//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3262-3270

[33] Luo Y, Zhou W, Fang J, et al. Epi-patch based convolutional neural network for depth estimation on 4d light field//Proceedings of International Conference on Neural Information Processing.Guangzhou, China, 2017: 642-652

[34] Heber S, Pock T. Convolutional networks for shape from light field//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3746-3754

[35] Faluvegi A, Bolsee Q, Nedevschi S, et al. A 3D convolutional neural network for light field depth Estimation//Proceedings of the International Conference on 3D Immersion. Kunming, China, 2019: 1-5

[36] Li Y, Zhang L, Wang Q, et al. Manet: Multi-scale aggregated network for light field depth estimation//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.Barcelona, Spain, 2020: 1998-2002

[37] Kinoshita T, Ono S. Depth estimation from 4D light field videos//Proceedings of International Workshop on Advanced Image Technology. Kagoshima, Japan, 2021: 301-306

[38] Chandra S, Usunier N, Kokkinos. Dense and lowrankgaussiancrfs using deep embeddings//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017:1-11

[39] Wang X, Gupta A. Videos as space-time region graphs//Proceedings of the IEEE European Conference on Computer Vision. Munich, Germany, 2018: 1- 21

[40] Chen Y, Rohrbach M, Yan Z, et al. Graph based global reasoningnetworks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 433-442

[41] Fu J, Liang J, Wang Z. Monocular Depth Estimation Based on

Multi-Scale Graph Convolution Networks. IEEE Access, 2020(8): 997-1009

[42] Wang F, Jiang M, Qian C, et al. Residual attention network for image classification. arXiv preprint arXiv: 1704.06904, 2017

[43] Kong S, Fowlkes C. Pixel-wise attentional gating for parsimonious pixel labeling. arXiv preprint arXiv: 1805.01556, 2018

[44] Chen L C, Yang Y, Wang J, et al. Attention to scale: Scale aware semantic image segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3640-3649

[45] Xu D, Wang W, Tang H, et al. Structured attention guided convolutional neural fields for monocular depth estimation//Proceedings of the IEEE Conference on Computer Vision andPattern Recognition. 2018: 3917-3925

[46] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014

[47] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Venice, Italy, 2017: 1-14

[48] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 2366-2374

[49] Hu J, Ozay M, Zhang Y, et al. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries//Proceedingsof the IEEE Winter Conference on Applications of Computer Vision. Hawaii, USA,2019: 1043-1051

[50] Moeller M, Benning M, Cremers D, et al. Variational depth from focus reconstruction. IEEE Transactions on Image Processing 2015, 24(12): 5369-5378

[51] Jeon H G, Park J, et al. Accurate depth map estimation from a lenslet light field camera//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1547-1555

JI Xin-Xin, M. S. candidate. Her research interests include computer vision.

PIAO Yong-Ri ， Ph. D., associate professor. His main research interests include compute vision.

ZHANG Miao,Ph. D., associate professor. Her research interests include computational imaging.

JIA Ling-Yao, Ph. D. candidate. His research interests include compute vision.

LI Pei-Hua, Ph. D., professor, Ph. D. supervisor. His research interests include compute vision.

Background

Image-based depth estimation is a fundamental problem in computer vision, and it plays an important role in a wide range of applications, such as 3D reconstruction, object tracking and so on. Since common two-dimensional cameras only record the intensity of light, the previous methods of depth estimation based on two-dimensional images captured by ordinary cameras to obtain inaccurate depth maps. Different from traditional imaging system, the light field cameras can capture the additional direction information. Therefore, the light field data provides more cues for the depth estimation.

The focal stack, as one important type of light field data, consists of several focal slices which contain the abundant depth cue and allow a human observer to instantly understand the order of objects along the depth in a scene. Some researchers have done many works on the depth estimation from the focal stack based on the non-deep-learning: Ng solved the corresponding depth by comparing the ambiguities of pixels at different focal stacks; Tao combined the defocus and correspondence cues to predict the depth map from the focal stack. Those methods achieve accurate depth maps, but they are usually too dependent on prior knowledge to generalize to other datasets easily. Hazirbas proposed the first method based deep-learning to compute the depth from the focal stack, which can be adapted to other datasets easily. But the loss of detail caused by the defocus blur was ignored.

In this paper, we present a novel accurate and robust method that predict depth map from multi-modal information. In order to effectively extract and fuse the multi-modal features from the foal stack and RGB images, we design the context reasoning unit and attention-guided cross-modal

fusion module. Experimental results show that our method can accurately estimate depth.

Our research group has done a lot of research on the light field. Related works have been published in international journals and conference in the fields of saliency detection, motion estimation, such as IEEE Trans on Image Processing, AAAI, NIPS, IJCAI etc. Recently, we are aiming at depth estimation from light field.

This work is supported by the National Natural Science Foundation of China (NO. 61471082) and the Science and Technology Innovation Foundation of Dalian (No.2019J12GX034).