第 43 卷 第 14 期/2023 年 7 月/光学学报 研究论文

光学学报

基于平面系数表示的自适应深度分布单目深度估计方法

王家骏，刘越*，吴宇晖，沙浩，王涌天

北京理工大学光电学院北京市混合现实与新型显示工程技术中心，北京 100081

摘要 提出了一种针对室内场景的轻量化端到端单目深度估计神经网络。首先设计了新型自适应深度分布估计模块，可针对不同输入图像估计差异化的深度范围，使得网络更好地预测室内物体的相对位置关系，恢复出的深度图像能获得更接近真实值的像素分布。其次，在深度估计的过程中，通过基于平面系数的深度间接表示形式加入平面隐式约束，可以在场景的平面区域得到更平滑的深度估计结果。在NYU Depth-v2数据集上的多项实验结果表明，提出方法能满足较高分辨率下的实时性要求，同时能以更少的参数恢复出质量更高、更完整的室内深度图像，有助于实现更加准确的三维重建效果。

关键词 机器视觉；深度估计；单目三维重建；神经网络

中图分类号 TP301.6 **文献标志码** A **DOI**：10.3788/AOS230468

1 引 言

获取场景深度是三维重建[1]、自动驾驶[2-4]、虚拟现实[5]等任务的重要基础。已有的基于激光雷达或飞行时间(ToF)相机获取深度的方法成本较高，无法大规模推广应用。相比之下，仅通过单张彩色(RGB)图像获取场景深度信息的方式成本低、适用性强，有着更为广阔的应用前景。

人类一般可通过透视关系、熟悉物体大小的缩放关系和物体之间的阴影及遮挡等线索进行推理，合理地估计出单目图像的场景深度。对计算机而言，从单张RGB图像中估计深度是一个不适定问题，传统算法难以对该问题进行建模处理。受近年来深度学习方法在各种不适定问题中成功应用的启发[6-8]，基于卷积神经网络的许多工作通过提取物体和场景的特征联合估计出合理精确的单目深度[9-12]，基于深度学习的单目深度估计方法逐渐成为主流。

单目深度估计任务旨在基于训练模型拟合RGB图像与深度图像之间的端到端映射关系，并通过所训练模型直接推断出场景图像的逐像素深度图。2014年，文献[13]提出一个端到端的卷积神经网络，通过由粗到细的两个子网络分别学习不同细节的单目场景深度。此后，研究人员分别从以下几个方面开展了研究工作：1)提出更加先进的骨架网络模型，如ResNet[14]、DenseNet[15]、Transformer[16]等结构以使骨干网络获得更优秀的特征提取能力；2)提出空洞空间卷积池化金字塔[17](ASPP)、局部平面引导层[18]、自注意力机制[16,19]等模块帮助网络提取融合特征；3)设计更贴合深度估计任务的损失函数，如尺度不变误差[13]、每像素最小重投影损失[20]等以帮助网络更好地拟合映射关系；4)改进训练方法为自监督[20-22]或无监督[23-25]等，通过减少对数据集的人工处理以使网络能在更大的数据集上拟合出更好的效果；5)与语义标签[26]、法线[27-28]以及平面检测[29]等任务进行深度联合学习以从多角度多任务中提取所需特征。

综上所述，已有研究[9-12,20-28]大多集中在如何增强网络的特征提取能力上，忽略了对图像本身深度像素值分布的研究。通过估计图像的像素深度分布范围不仅能够获得更加准确的深度关系表示，提高推理精度，而且也能改善深度预测图像的整体深度分布，使重建恢复出的三维图像与真实情况更加相符。因此本文提出了一种新型自适应深度分布模块，使得模型在训练过程中可以差异化地为每张图像预测不同的深度分布区间，从深度分布的层面指导模型学习。结合轻量化的骨架网络[30]和解码器结构设计，本文方法与已有方法相比推理速度更快，且推理出的二维深度图像与实际分布更加相符，恢复出的三维几何场景更加合理完整，具有较高的应用价值。

收稿日期：2023-01-12；**修回日期**：2023-02-24；**录用日期**：2023-03-20；**网络首发日期**：2023-05-08

基金项目：国家自然科学基金(61960206007)、高等学校学科创新引智计划(B18005)

通信作者：*liuyue@bit.edu.cn

2 算法原理

2.1 基于平面系数的深度表示方法

在相机坐标系中，一个空间中的三维平面可通过如下公式来定义：

$$aX+bY+cZ+d=0, \tag{1}$$

式中：(a,b,c)为平面法向量；d为距相机中心的距离。基于相机小孔成像模型，空间中的三维点$P=(X,Y,Z)^{\mathrm{T}}$可通过如下公式投影到平面中对应点$p=(u,v)^{\mathrm{T}}$：

$$\begin{cases} u=f_x\dfrac{X}{Z}+u_0 \\ v=f_y\dfrac{Y}{Z}+v_0 \end{cases}, \tag{2}$$

式中：Z为像素坐标(u,v)对应的深度$D(u,v)$；(f_x,f_y)为相机的焦距；(u_0,v_0)为像主点。将式(2)代入式(1)中可得

$$\frac{1}{Z}=-\frac{a}{f_xd}u-\frac{b}{f_yd}v+\frac{1}{d}\left(\frac{a}{f_x}u_0+\frac{b}{f_y}v_0-c\right)。 \tag{3}$$

在此引入三个新参数$\hat{p}$、$\hat{q}$、$\hat{r}$作为平面系数，将式(3)重写为

$$Z=\frac{1}{\hat{p}u+\hat{q}v+\hat{r}}, \tag{4}$$

式中：$\hat{p}=-\dfrac{a}{f_xd}$；$\hat{q}=-\dfrac{b}{f_yd}$；$\hat{r}=\dfrac{1}{d}\left(\dfrac{a}{f_x}u_0+\dfrac{b}{f_y}v_0-c\right)$。对于室内场景来说，全局的尺度是不确定的，因此在上述公式中引入尺度因子s并对平面系数$\hat{p}$、$\hat{q}$、$\hat{r}$实行归一化：

$$s=\sqrt{\hat{p}^2+\hat{q}^2+\hat{r}^2}, \tag{5}$$

归一化之后的平面系数为尺度无关的系数，最后基于平面系数的间接深度可表示为

$$Z=\frac{1}{(pu+qv+r)s}。 \tag{6}$$

平面系数在对深度图像的表示上为平面区域增加了隐式约束，使得图像中的平面区域成为一个内在的整体。当输入图像中的平面部分有着较强的遮挡或者角度太大时，该表示方法能够通过内在的隐式约束输出平滑的深度平面区域。基于平面的约束引导，在三维空间的重建中图像扭曲失真的情况也会有较好的改善。

2.2 基于图像深度分布的预测模块

不同RGB图像对应的深度分布在很大程度上是不同的，如图1(a)所示，室内场景中例如走廊等最大深度无约束的图像，其像素分布会更加均匀。而对于如图1(b)、1(c)所示的最大深度有约束的图像或室内物体特写图像，像素值会分布在更靠近相机的区域内。针对室内图像中存在的深度区间变化的问题，一些研究工作[11-13,20]在对图像预处理时会将图像数值范围规范为[0,1]区间，以加快网络拟合速度，但这些工作均只使用相同的操作对整个数据集进行操作，没有针对每张图像的像素分布对其进行单独的区间调整。为此，本文提出自适应图像深度分布预测模块，在训练中对每张图像进行差异化的深度分布预测，使得网络在提取特征的同时能学习到像素分布层面上的信息，让网络能够获得和原始深度分布更加一致的预测结果。

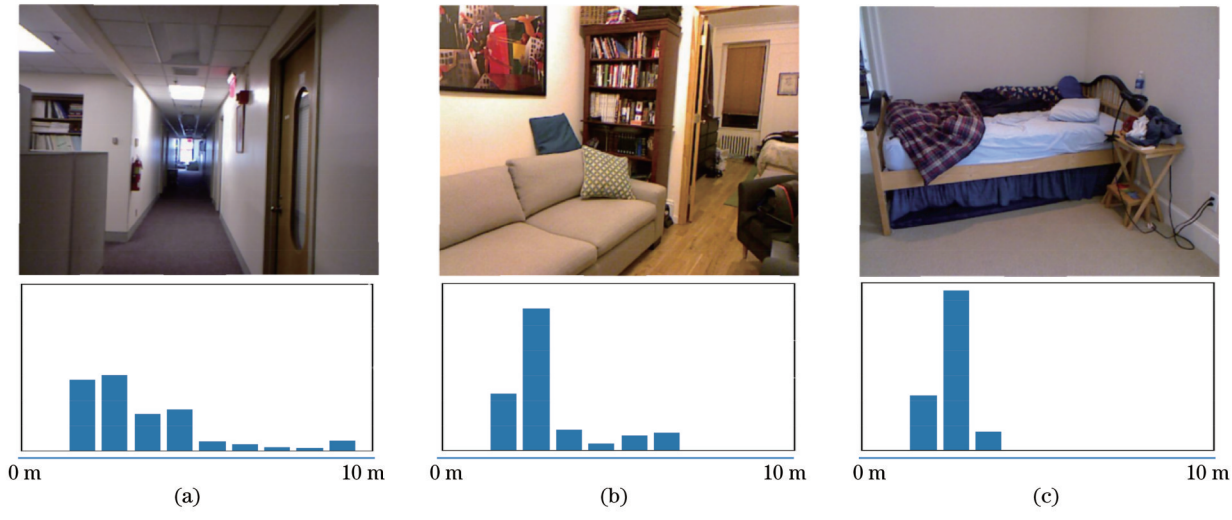

图1 室内场景中不同图像像素深度分布示意图。(a)最大深度无约束图像；(b)最大深度有约束图像；(c)家具特写图像

Fig. 1 Schematic diagram of pixel depth distribution of different images in indoor scene. (a) Image with maximum depth unconstrained; (b) image with maximum depth constrained; (c) close-up image of furniture

基于图像深度分布的预测模块如图2所示，其核心思想旨在让网络学习到每张不同图像的像素值分布区间。为了尽可能降低网络模型的参数大小以加快推理速度，首先将卷积网络输出的特征x_{F}依次经过卷积

层、批量归一化(BN)层[31]和激活函数层，得到特征通道减小为2、分辨率减小为$(H/8, W/8)$的特征x_{sq}，然后经过延展函数F_{st}将参数特征延展为一维特征参数向量x_s，上述过程可表述为

$$x_s = F_{st}(x_{sq}) = F_{st}\{\sigma[\mathrm{BN}(W_0 x_F)]\}, \quad (7)$$

式中：BN( )为批量归一化操作；σ( )为ReLU激活函数；$W_0 \in \mathbb{R}^{c_1 \times c_0}$为可学习的网络参数。压缩后的特征$x_s$最终通过两层线性函数层，获得分别代表深度分布区间最大值和最小值的二维参数。可表述为

$$(P_{\min}, P_{\max}) = W_2 F_D[\sigma(W_1 x_s + b_1)] + b_2, \quad (8)$$

式中：F_D( )表示Dropout操作；$W_1 \in \mathbb{R}^{c_2 \times c_1}$、$W_2 \in \mathbb{R}^{2 \times c_2}$、$b_1 \in \mathbb{R}^{c_2}$、$b_2 \in \mathbb{R}^2$为可学习参数。获得深度区间范围$(P_{\min}, P_{\max})$后，可通过如下公式将初始深度估计图$D_{init}$深度范围自适应调整为最终深度估计图$\hat{D}$：

$$\hat{D} = D_{init}(P_{\max} - P_{\min}) + P_{\min}。 \quad (9)$$

基于图像深度分布的预测模块在训练时允许网络在像素分布上提取信息，能够通过学习到的范围参数$(P_{\min}, P_{\max})$在深度像素分布的层面上对每张图像进行差异化调整。这些操作有助于网络更好地把握整体，估计出分布更加符合实际场景的深度图。

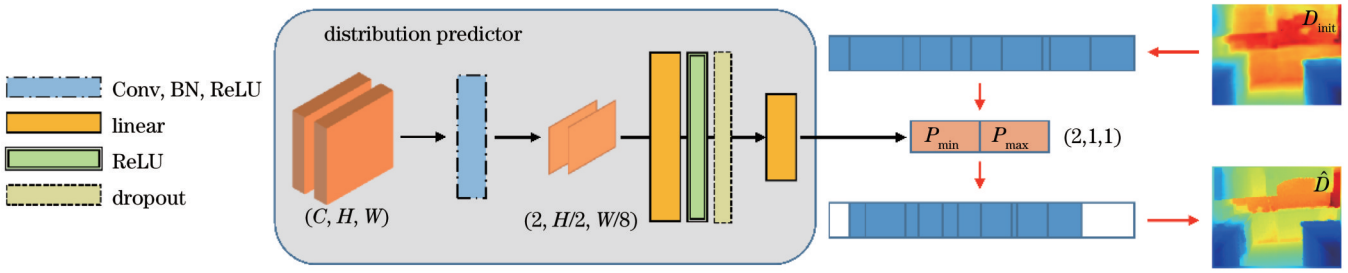

图2　图像深度分布预测模块结构图

Fig. 2　Structure diagram of image depth distribution prediction module

2.3　单目估计方法架构

为实现单目图像的深度估计，需要网络学习到一个映射关系：

$$f_\theta : I(u, v) \rightarrow D(u, v), \quad (10)$$

式中：I为输入网络的RGB图像，其分辨率大小为$H \times W$；D为其对应分辨率下的深度图像；(u, v)是图像像素坐标系下的坐标；θ是映射关系f中的参数。为使网络能够快速学习和预测推理，本文以轻量化的结构设计为前提设计了如下的网络结构。

如图3所示，本文提出的网络模型采用类似于U-Net网络[32]的跳层连接结构，主要由编码器和解码器两个部分组成。其中编码器采用基于ResNet[14]网络优化的ResNeXt50[30]作为骨架网络。在解码器部分，本文出于轻量化考虑在网络瓶颈处和其他部分设计了两种不同的子网络结构，如图3(a)、3(b)所示，主要由卷积层、BN层和SEAttention[33]等模块组成。图3(a)所示的子网络结构只存在于网络瓶颈处，考虑到该部分的特征往往维数较高，且具有较高的语义属性，因此该结构中特征需先经过一次卷积操作以提炼特征，然后依次通过跳层连接模块、SEAttention模块进行特征权重的放缩，以选择对结果更有价值的特征进行重点学习。剩余部分的解码器均采用如图3(b)所示的子网络结构，由于当前层骨干网络的特征分辨率还较高，语义信息提取得不够充分，因此需要先通过卷积操作进行特征的压缩和提取，然后和上采样后的深层特征通过SEAttention模块进行基于通道注意力机制的挤压和缩放，最后经过卷积和残差跳层连接操作后输出结果。在解码器中，特征依次通过子网络结构后均需要进行上采样操作以匹配上层特征的分辨率，最后所有尺度下的子网络输出层在局部平面引导[18]模块恢复到640×480的初始分辨率，并进行多尺度的输出特征与初始图像的维度聚合操作。基于模型的轻量化考虑，本文仅设计了4层的模型深度，以简化模型的参数和加快模型的推理速度。在局部平面引导层聚合后的特征依次经过卷积层、BN层和激活函数等最终获得6维的预测参数。其中，后4层参数作为深度图像平面表示的参数，经过式(6)的系数转换后输出图像的初始深度图D_{init}，前两层参数经过2.2节中的深度预测分布模块后，输出网络对于当前图像的差异化分布预测。最后通过式(9)将初始深度图D_{init}进行深度的差异化调整，输出最终深度图$\hat{D}$。

2.4　损失函数设计

合理的损失函数设计对于网络的迅速拟合和收敛至关重要[20]。考虑到深度图像本身在维度上隐含的三维空间特征，本文分别基于二维像素和三维空间点两个层次设计了不同的损失函数，在不同的维度上约束网络的拟合优化。

1)基于二维像素关系的损失函数。该损失函数主要用来衡量模型的估计深度图像$\hat{D}$和真值深度图像D^*在二维空间中的差异。考虑到深度估计任务的尺度不确定性，本文选择文献[13]中提出的尺度不变误差的缩放版本作为模型在二维空间中像素间关系的损失函数：

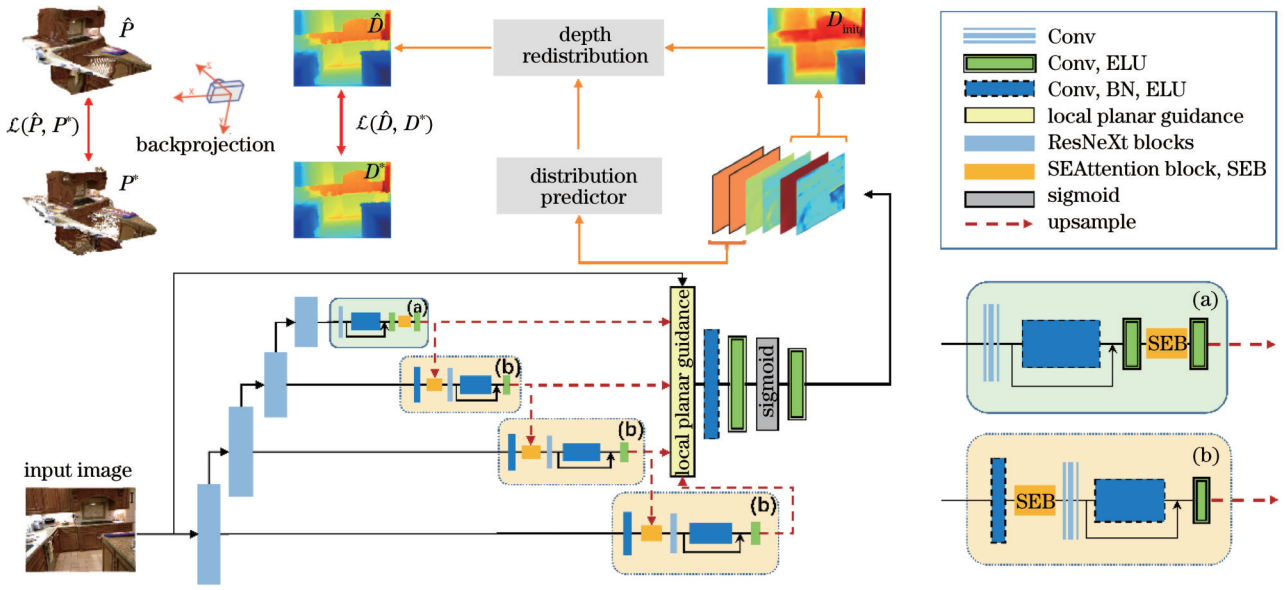

图 3　单目深度估计网络整体结构图。(a)、(b)子网络结构图

Fig. 3　Overall architecture of monocular depth estimation network. (a), (b) Subnetwork structure

$$L_{\text{pixel}}=\mu\sqrt{\frac{1}{N}\sum_i d_i^2-\frac{\lambda}{N^2}\left(\sum_i d_i\right)^2},\tag{11}$$

式中：$d_i=\ln\hat{D}_i-\ln D_i^*$；图像中的总像素个数 N 和像素位置 i 均表示估计深度图像 $\hat{D}_i$ 和真值深度图像 D_i^* 中的深度有效值。在整个训练过程中，本文使用文献[13]中的超参数定义 $\lambda=0.85$，$\mu=10$。基于深度图中边界位置深度变化的不连续性，本文使用 Sobel 算子计算两者的梯度变化，并代入 L_2 损失函数作为模型在二维空间中边界关系的损失函数，强化预测图像与真值图像之间的边缘界限：

$$L_{\text{edge}}=\frac{1}{N}\sum_{i=1}^{N}\left[\left(\nabla_x d_i\right)^2+\left(\nabla_y d_i\right)^2\right],\tag{12}$$

式中：$d_i=\ln\hat{D}_i-\ln D_i^*$；$\nabla_x$、$\nabla_y$ 分别表示 Sobel 算子在图像 x 方向和 y 方向上计算得到的梯度图像。

2)基于三维空间点关系的损失函数。根据式(2)，可通过如下过程将估计深度图像 $\hat{D}$ 和真值深度图像 D^* 分别投影到三维空间中，获得其三维空间点云集 $\hat{P}$ 和 P^*：

$$P=(X,Y,Z)^{\mathrm{T}}=\left[\frac{(u-u_0)}{f_x(pu+qv+r)s},\frac{(v-v_0)}{f_y(pu+qv+r)s},\frac{1}{(pu+qv+r)s}\right]^{\mathrm{T}},\tag{13}$$

式中：p、q、r、s 分别表示 2.1 节中的平面系数；(u,v) 表示像素坐标；(f_x,f_y) 为相机的焦距；(u_0,v_0) 为像主点。为了在三维空间中约束预测图像，本文使用 Chamfer Distance[34]计算 $\hat{P}$ 中每一点到 P^* 中的最近距离，以使估计的深度图像与真值深度图像在三维空间中的点云尽可能接近，可表示为

$$L_{\text{dis}}=\sum_{p_1\in\hat{P}}\min_{p_2\in P^*}\left(p_1-p_2\right)+\sum_{p_2\in P^*}\min_{p_1\in\hat{P}}\left(p_2-p_1\right),\tag{14}$$

式中，p_1 和 p_2 分别表示三维空间点云集 $\hat{P}$ 和 P^* 中的任意一点。由于真值深度噪声的存在，通过投影操作，通常会恢复出凹凸不平的三维空间点平面。为了减少噪声对网络三维空间拟合效果的影响，本文引入文献[28]中的基于深度的虚拟法线转换策略 H_{v}，使网络在三维空间中更加关注物体之间的全局相对位置关系，具体可表示为

$$L_{\text{vir}}=\frac{1}{N_{\text{v}}}\sum_{i=0}^{N_{\text{v}}}\left(\hat{n}_i-n_i^*\right)^2,\tag{15}$$

式中：N_{v} 表示虚拟法线转换策略 H_{v} 中选择的三维空间点集的数量；$\hat{n}_i$、n_i^* 分别表示估计深度图像 $\hat{D}_i$ 和真值深度图像 D_i^* 通过 $H_{\text{v}}(\hat{D}_i)$、$H_{\text{v}}(D_i^*)$ 获得的对应虚拟法线。

综上所述，本文使用的损失函数可分为二维像素关系的损失 L_{pixel}、L_{edge} 和三维空间点关系的损失 L_{dis}、L_{vir}，具体可表示为

$$L=\alpha(\lambda_1 L_{\text{pixel}}+\lambda_2 L_{\text{edge}})+\beta(\lambda_3 L_{\text{dis}}+\lambda_4 L_{\text{vir}}),\tag{16}$$

式中：α、β 为网络在不同训练阶段的权重常数；λ_1、λ_2、λ_3、λ_4 分别为各损失函数的权重常数。在网络训练时，通过计算在二维像素和三维空间点上估计深度与真值深度之间的误差损失，并将误差总和反向传播至网络

以迭代优化参数θ，进而不断降低损失函数，具体过程可定义为

$$\min_{\theta}\sum_{(I,D^*)\in T}L\left[f_\theta(I),D^*\right],\tag{17}$$

式中：T为输入图像I和对应的真值深度图像D^*的训练对；L为估计深度图像$f_\theta(I)$和真值深度图像D^*之间差距的损失函数。本文采用两阶段的训练策略，在训练时首先利用二维像素关系对网络进行预训练，当损失不再下降时加入三维空间点关系进行联合训练，直至网络收敛。当网络完成训练后，输入二维图像即可完成对该场景深度图像的估计。训练过程的具体细节见3.2节。

3 实验与结果分析

3.1 数据集

本文使用NYU Depth-v2数据集[35]对网络进行训练。该数据集由Microsoft Kinect[36]设备采集的RGB和深度图像的视频序列组成，有效深度区间为0～10 m，具体包括在464个室内场景下约40万张RGB和深度图像对。在实验中，本文依据官方提供的划分方式按场景对数据集进行划分[13]，同时使用官方提供的深度补全工具[37]对缺失的深度信息进行补全处理。在测试时，本文采用文献[13]中定义的中心裁剪方法对测试图像进行预处理以统一对比标准。最终本文所用的训练集包含约5万张图像对，测试集为官方处理好的654张图像对。

3.2 实验细节

本文方法主要基于PyTorch深度学习框架实现。计算平台搭载的CPU型号为i9-7920X，在Ubuntu 20.04.2 LTS的环境下使用4块GeForce RTX 2080Ti显卡进行训练。在整个训练期间使用的优化器为Adam优化器，其中$(\beta_1,\beta_2)=(0.9,0.999)$，批量大小(batch size)为32，损失函数中的超参数设置为$\lambda_1=0.9$、$\lambda_2=0.5$、$\lambda_3=0.7$、$\lambda_4=0.9$。本文采用两阶段的训练策略，在第一阶段时首先加载骨干网络在ImageNet[38]上进行预训练后的权重，设置超参数$\alpha=1$、$\beta=0$，使用的初始学习率为1×10^{-4}，共训练30轮。当第一阶段完成训练后，加入三维关系的损失函数对网络进行联合训练，设置第二阶段的超参数$\alpha=0.5$、$\beta=0.8$，使用的初始学习率为5×10^{-6}，共训练25轮。在训练过程中，当均方根误差(RMSE)指标连续3轮没有下降时，学习率会以0.5的比率降低。训练完成后，在官方提供的测试集[34]上对模型推理结果进行定性定量的评估。

3.3 评价指标

本文使用6项标准指标[41]进行模型精度验证，包括RMSE、lg误差、绝对值相对误差(REL)和阈值内准确比率($\delta<X_{\mathrm{thr}}$)，分别定义为

$$V_{\mathrm{RMSE}}=\sqrt{\frac{1}{N}\sum_{i=0}^{N}\left(\hat{D}_i-D_i^*\right)^2},\tag{18}$$

$$V_{\lg}=\frac{1}{N}\sum_{i=0}^{N}\left|\lg\hat{D}_i-\lg D_i^*\right|,\tag{19}$$

$$V_{\mathrm{REL}}=\frac{1}{N}\sum_{i=0}^{N}\frac{\left|\hat{D}_i-D_i^*\right|}{D_i^*},\tag{20}$$

$$V_{\mathrm{TH}}=\max\left(\frac{\hat{D}_i}{D_i^*},\frac{D_i^*}{\hat{D}_i}\right)=\delta<X_{\mathrm{thr}},\tag{21}$$

式中，图像中的总像素个数N和像素位置i均表示在估计图像$\hat{D}_i$和真值图像D_i^*中的深度有效值。在这6项指标中，RMSE、lg误差和REL用于衡量估计深度和真实深度值之间的绝对和相对误差，其数值越小，准确度越高。而阈值内准确比率$\delta<X_{\mathrm{thr}}$用于衡量估计深度和真实深度在一定范围内的误差比率大小，其数值越大，准确度越高[39]，通常取$X_{\mathrm{thr}}=1.25^i$，$i=1,2,3$。

为直观地对比各模型之间的深度分布预测效果，本文基于IoU[40](intersection over union)设计模型分布的重叠度，定义为

$$D_{\mathrm{IoU}}=\frac{\sum_{n=1}^{100}\min\left(U_{\mathrm{gt}},U_{\mathrm{pred}}\right)}{\sum_{n=1}^{100}\max\left(U_{\mathrm{gt}},U_{\mathrm{pred}}\right)},\tag{22}$$

式中：n为划分的深度区间；U_{gt}、U_{pred}分别为图像的真实值和模型的预测值在n内的像素数量；$0\leqslant D_{\mathrm{IoU}}\leqslant1$。模型的预测结果与真实值之间深度分布的重叠度越高，意味着其与真实深度分布越接近，效果越好。

3.4 实验结果与分析

本文方法与其他文献的方法的定性定量比较结果如表1和图4所示。出于轻量化考虑，本文将模型简化设计为4层的网络深度，并在模块的连接部分加入卷积层以强化模型的特征提取能力。从表1可以看出，本文所提出的简化网络模型在参数量仅为46 M的情况下，指标超越了所列的大多数方法，证明了本文中设计的模型整体结构简练有效。其中，文献[18]和文献[28]虽然在部分指标上比本文略好，但其模型的参数量远远大于本文所提方法，证明本文设计的模型结构在有限参数的条件下特征提取能力远超其他方法。由图4的可视化比较中可以观察到，本文方法在深度图的预测上准确度较高，尤其在物体的平面区域能获得更为平滑和准确的连续深度预测，且在几何区域较为复杂的情况下也能获得较为清晰的边缘。

本文提出采用平面系数的隐式表达形式来还原单目图像的深度信息。与其他使用端到端的像素预测方法相比，本文的深度表达形式对平面结构有着更为良好的约束，恢复出的三维图像更为平滑和准确。如图5所示，文献[45]中的方法未在训练中添加三维空间内的约束损失，在三维重建后图像的空洞较多，过渡不

表 1　提出方法在 NYU Depth-v2 数据集上与其他方法的定量比较

Table 1　Quantitative comparison of proposed method with other methods on NYU Depth-v2 dataset

Method	RMSE	REL	lg	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$	Params /M
Eigen et al. [13]	0.641	0.158	—	0.769	0.950	0.988	141
Laina et al. [42]	0.573	0.127	0.055	0.811	0.953	0.988	64
Hao et al. [43]	0.555	0.127	0.053	0.841	0.966	0.991	60
Fu et al. [44]	0.509	0.115	0.051	0.828	0.965	0.992	110
Hu et al. [45]	0.530	0.115	0.050	0.866	0.975	0.993	157
Raman et al. [46]	0.495	0.139	0.047	0.888	0.979	0.995	80.4
Lee et al. [18]	0.419	0.119	0.051	0.865	0.975	0.993	49.5
Yin et al. [28]	0.416	0.108	0.048	0.878	0.977	0.994	114.2
Proposed	0.416	0.121	0.050	0.864	0.974	0.995	46

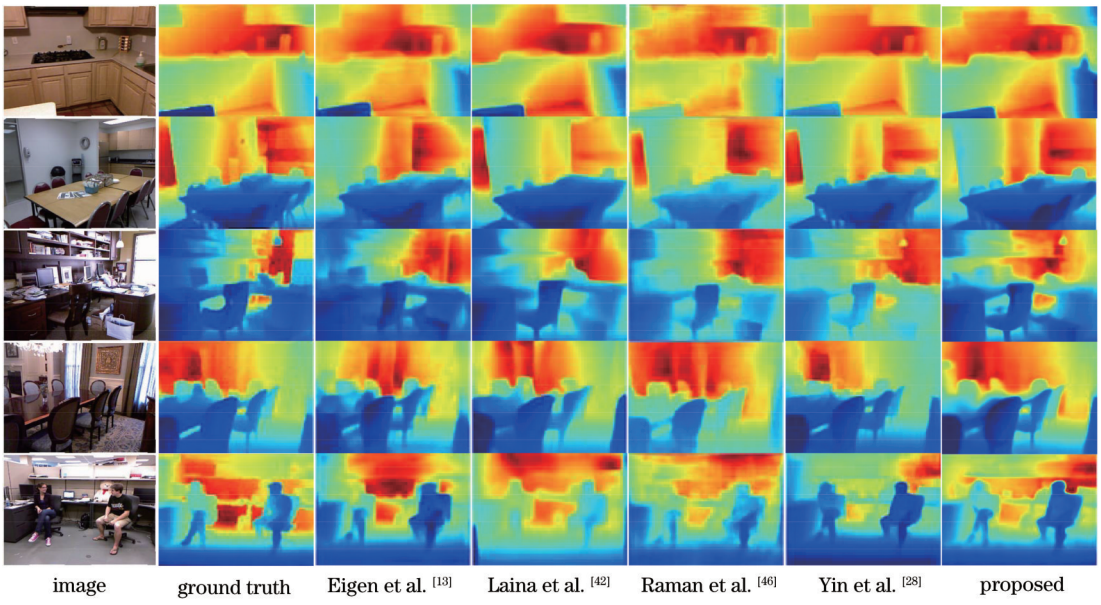

图 4　提出方法在 NYU Depth-v2 数据集上与其他方法的可视化比较

Fig. 4　Visual comparison of proposed method with other methods on NYU Depth-v2 dataset

够平滑，且在桌面、地面等位置的噪声较大。文献[28]中的方法在训练时加入了虚拟法线的约束损失，但是没有使用平面系数的深度表示形式，在被部分遮挡或缺少信息的平面区域中存在扭曲和失真的情况。相比之下，本文使用基于平面系数的间接深度表示方法，并在三维点云空间设计损失函数进行约束，即使在缺少信息的平面区域也可以给出合理的预测结果。

本文所提方法与其他文献方法的模型推理速度定量比较如表 2 所示，表中包含的文献模型均在单块 GeForce RTX 3060 显卡上进行测试。表 2 的各项指标中最好的数据使用粗体标出，次好的数据使用下划线标出。由表 2 可得，本文所提出的模型在 640 × 480 的图像分辨率下运行速度可达 33 frame/s，可以满足实时的单目深度估计要求，且本文模型在表 2 所含文献的方法中具有最小的参数量，RMSE 指标最好，推理时间和帧率均为次好指标，证明本文所提出的模型方法在速度和精度上达到了平衡，可以在保证模型轻量化和推理速度的同时，获得较好的预测质量，具有较高的应用价值。

为验证本文提出的自适应深度分布模块的有效性，本文将 NYU Depth-v2 数据集的深度划分为 100 个深度区间（0.1 m/interval），并在官方提供的 654 张测试图像[35]上对模型的深度分布预测效果进行研究。本文所提方法与其他文献的方法在官方测试集上的深度分布预测定性定量对比结果如图 6 和表 3 所示。由深度分布曲线和 D_{IoU} 值的计算结果可得，本文所提出的预测方法使得模型的预测深度分布与真实值的深度分布在曲线的走势上更加贴合，且与真实值分布的重叠度更高，证明本文所提出的模型能够估计出更加符合真实场景深度分布的图像。

3.5　消融实验

本文在 NYU Depth-v2 数据集上针对所提网络结构的有效性进行了消融实验。为保证实验变量的一致性，去除相应结构模块后的网络在训练策略上与初始

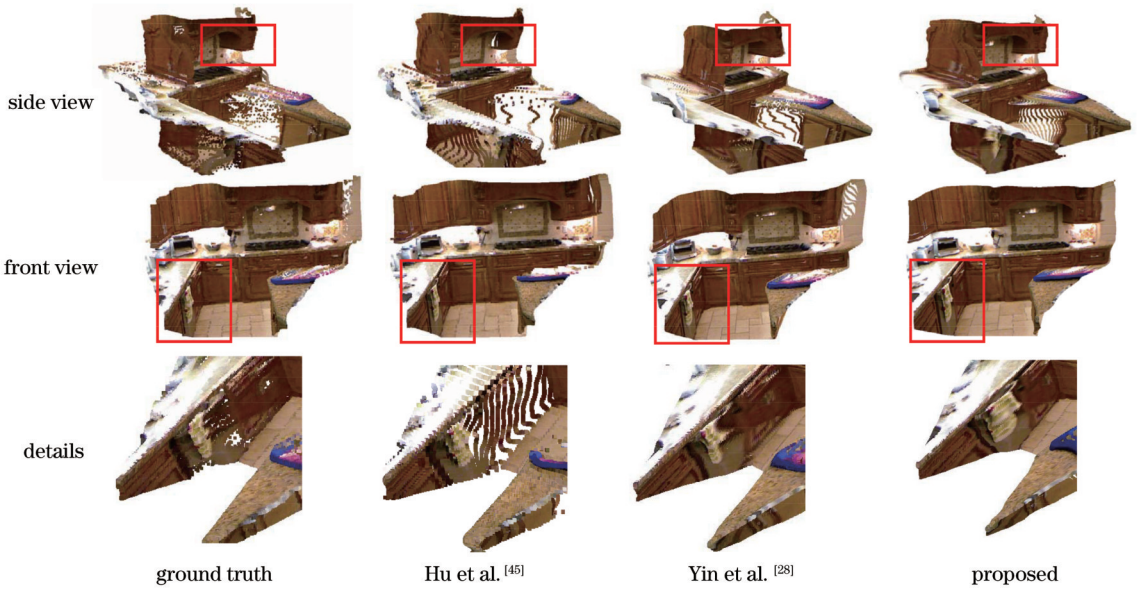

图 5　提出方法与其他方法的三维重建结果可视化比较

Fig. 5　Visual comparison of three-dimensional reconstruction of proposed method with other methods

表 2　提出方法与其他方法的推理速度比较

Table 2　Inference speed comparison of proposed method with other methods

Method	Running time /ms	Frame rate /(frame·s^{-1})	RMSE	Params /M
Laina et al. [42]	36	27	0.573	64
Fu et al. [44]	35	28	0.509	110
Hu et al. [45]	91	11	0.530	157
Raman et al. [46]	**21**	**41**	0.495	80.4
Lee et al. [18]	60	16	<u>0.419</u>	<u>49.5</u>
Yin et al. [28]	36	27	**0.416**	114.2
Proposed	<u>30</u>	<u>33</u>	**0.416**	**46**

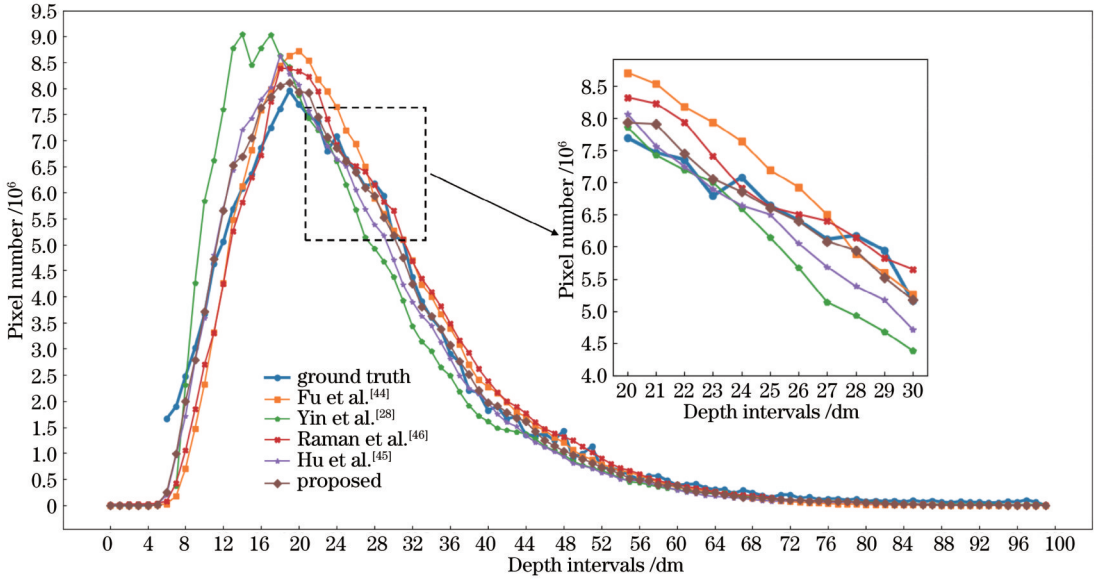

图 6　提出方法与其他方法的深度分布定性定量比较

Fig. 6　Qualitative and quantitative comparison of depth distribution of proposed method with other methods

表 3 提出方法与其他方法在测试集的 IoU 计算结果比较

Table 3 Comparison of IoU results of proposed method with other methods on test set

Method	D_{IoU}
Fu et al.[44]	0.863
Yin et al.[28]	0.805
Raman et al.[46]	0.883
Hu et al.[45]	0.887
Proposed	0.914

网络均保持一致。在实验中首先去除网络的深度分布预测模块，直接使用真值的深度数据对预测图像进行反归一化。然后将本文中使用的基于深度平面表示间接深度预测方法修改为对图像像素深度的直接预测方法，并对网络进行训练直至网络收敛。由表 4 可以看出，使用基于深度平面的间接深度预测方法和本文提出的深度分布预测模块在 RMSE 等多项指标上均有较大的提升，表明了在模型中该种预测方法和预测模块的有效性。

表 4 基于网络结构消融实验的定量结果

Table 4 Quantitative results of ablation experiments based on network architecture

Method	RMSE	REL	lg	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Depth distribution predictor	0.430	0.127	0.052	0.853	0.971	0.994
Plane coefficient representations	0.420	0.123	0.051	0.862	0.975	0.995
Baseline	0.416	0.121	0.050	0.864	0.974	0.995

本文同时针对所选用的损失函数进行消融实验，将只使用 L_{pixel} 进行训练的网络作为参考基线。实验中使用的训练策略与初始网络均保持一致。为保证实验变量的一致性，在仅使用二维像素关系和使用二维像素关系及三维空间点关系的损失函数进行训练时，其训练的总轮数保持一致，最终实验结果如表 5 所示。由表 5 可得，基于基线分别添加 L_{edge}、L_{vir} 和 L_{dis} 均能使网络模型的准确率获得进一步的提升，且网络模型在经过二维像素关系的损失函数训练并收敛后，再添加三维空间点关系进行训练，其各项指标均会获得进一步的提升，证明了本文选择的各项损失函数和训练策略的有效性。

表 5 基于网络损失函数消融实验的定量结果

Table 5 Quantitative results of ablation experiments based on network loss function

Method	RMSE	REL	lg	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Baseline	0.431	0.125	0.052	0.857	0.971	0.994
With L_{edge}	0.428	0.124	0.051	0.858	0.973	0.994
With L_{edge} and L_{vir}	0.426	0.123	0.051	0.859	0.973	0.994
With L_{edge} and L_{dis}	0.418	0.122	0.050	0.862	0.974	0.995
With L_{edge}, L_{dis}, and L_{vir}	0.416	0.121	0.050	0.864	0.974	0.995

3.6 限制性分析

本文以轻量化为前提设计了单目深度估计网络，且设计的自适应深度分布模块使网络在对单张图像深度分布的整体把控上有着较为良好的效果。但是如图 7 中玻璃墙上的栏杆、椅子腿和床头上的栏杆等部分所示，其对图像局部细节还是缺少还原的能力。同时如图 7 中玻璃墙、镜子以及透明材质的磨砂门等部分所示，由于单目深度估计仅凭视觉线索对图像进行

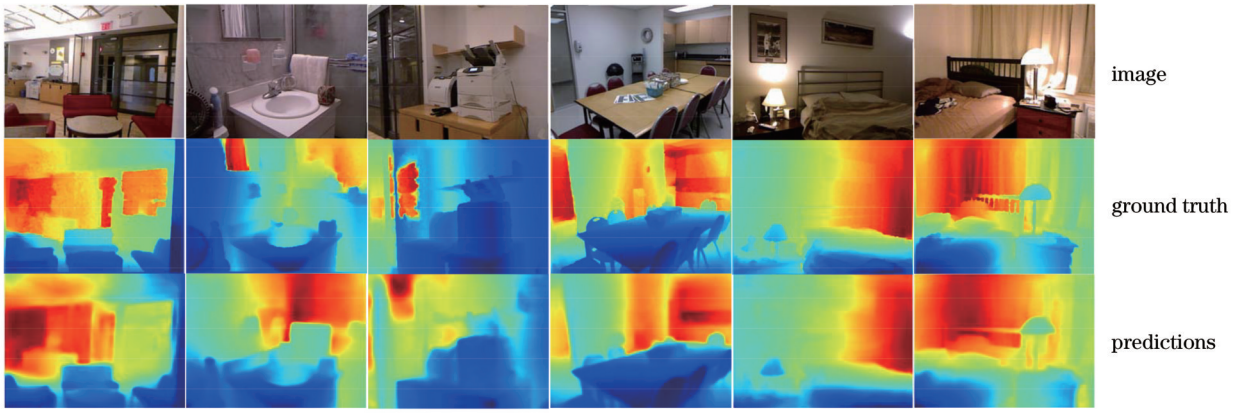

图 7 提出方法在 NYU Depth-v2 数据集上的估计失误图像

Fig. 7 Failure predictions of proposed method on NYU Depth-v2 dataset

处理，其在对玻璃、镜面等特殊材料的处理上较容易失误。因此，如何设计网络使其在保证模型预测实时性的同时能恢复出更多的深度细节，或针对玻璃、镜面等特殊材料进行专项的学习将成为后续工作开展的重点。

4 结 论

针对单目图像深度估计任务，本文引入了基于平面参数的间接深度表示方式，同时设计了深度分布预测模块，提出了基于平面系数表示的自适应深度分布模型。在NYU Depth-v2数据集上的定性定量结果和多项实验结果表明，本文采用的基于平面系数的间接深度表示方法对于图像中遮挡较为明显或可视角度较小的平面区域能够获得合理的预测结果；同时本文中提出的深度分布预测模块能够为每张图像提供差异化的像素分布优化，获得更加贴近真值图像的像素深度分布预测结果。在已知相机参数的情况下，本文提出的模型恢复出的三维场景更加完整合理，同时由于基于轻量化的设计，本文方法能够实现推理速度和推理精度的平衡，在室内虚拟现实、人机交互等需要实时性和准确性的实际场景中具有较高的应用价值。

参 考 文 献

[1] Han X F, Laga H, Bennamoun M. Image-based 3D object reconstruction: state-of-the-art and trends in the deep learning era [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(5): 1578-1604.

[2] Rasouli A, Tsotsos J K. Autonomous vehicles that interact with pedestrians: a survey of theory and practice[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 21(3): 900-918.

[3] Hussain R, Zeadally S. Autonomous cars: research results, issues, and future challenges[J]. IEEE Communications Surveys & Tutorials, 2019, 21(2): 1275-1313.

[4] 丁萌, 姜欣言. 先进驾驶辅助系统中基于单目视觉的场景深度估计方法[J]. 光学学报, 2020, 40(17): 1715001.
Ding M, Jiang X Y. Scene depth estimation based on monocular vision in advanced driving assistance system[J]. Acta Optica Sinica, 2020, 40(17): 1715001.

[5] Beever L, John N W. LevelEd SR: a substitutional reality level design workflow[C]//2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), March 12-16, 2022, Christchurch, New Zealand. New York: IEEE Press, 2022: 130-138.

[6] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 5686-5696.

[7] Hu H, Gu J Y, Zhang Z, et al. Relation networks for object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 3588-3597.

[8] He A F, Luo C, Tian X M, et al. A twofold Siamese network for real-time object tracking[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4834-4843.

[9] 刘佳涛, 张亚萍, 杨雨薇. 基于迁移学习的高效单目图像深度估计[J]. 激光与光电子学进展, 2022, 59(16): 1611002.
Liu J T, Zhang Y P, Yang Y W. Efficient monocular image depth estimation based on transfer learning[J]. Laser & Optoelectronics Progress, 2022, 59(16): 1611002.

[10] Zhang W D, Zhang W, Zhang Y D. GeoLayout: geometry driven room layout estimation based on depth maps of planes [M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12361: 632-648.

[11] Jun J, Lee J H, Lee C, et al. Depth map decomposition for monocular depth estimation[M]//Computer vision-ECCV 2022. Lecture notes in computer science. Cham: Springer, 2022, 13662: 18-34.

[12] Li Z Y, Chen Z H, Liu X M, et al. DepthFormer: exploiting long-range correlation and local information for accurate monocular depth estimation[EB/OL]. (2022-03-27) [2022-11-09]. https://arxiv.org/abs/2203.14211.

[13] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems, December 8-13, 2014, Montreal, Canada. Cambridge: MIT Press, 2014: 2366-2374.

[14] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[15] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2261-2269.

[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4-9, 2017, Long Beach, USA. Cambridge: MIT Press, 2017: 6000-6010.

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

[18] Lee J H, Han M K, Ko D W, et al. From big to small: multi-scale local planar guidance for monocular depth estimation[EB/OL]. (2019-07-24) [2022-11-09]. https://arxiv. org/abs/1907.10326.

[19] 杨蕙同, 雷亮, 林永春. 基于多尺度注意力特征融合的双目深度估计算法[J]. 激光与光电子学进展, 2022, 59(18): 1815005.
Yang H T, Lei L, Lin Y C. Binocular depth estimation algorithm based on multi-scale attention feature fusion[J]. Laser & Optoelectronics Progress, 2022, 59(18): 1815005.

[20] Godard C, Mac Aodha O, Firman M, et al. Digging into self-supervised monocular depth estimation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2020: 3827-3837.

[21] Watson J, Mac Aodha O, Prisacariu V, et al. The temporal opportunist: self-supervised multi-frame monocular depth[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 1164-1174.

[22] Song C Q, Niu M L, Liu Z P, et al. Spatial-temporal 3D dependency matching with self-supervised deep learning for monocular visual sensing[J]. Neurocomputing, 2022, 481: 11-21.

[23] Zhou T H, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-

26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6612-6619.

[24] Bian J W, Zhan H Y, Wang N Y, et al. Auto-rectify network for unsupervised indoor depth estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 9802-9813.

[25] Hui T W. RM-depth: unsupervised learning of recurrent monocular depth in dynamic scenes[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 1665-1674.

[26] Zhu S J, Brazil G, Liu X M. The edge of depth: explicit constraints between segmentation and depth[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 13113-13122.

[27] Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2016: 2650-2658.

[28] Yin W, Liu Y F, Shen C H, et al. Enforcing geometric constraints of virtual normal for depth prediction[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2020: 5683-5692.

[29] Patil V, Sakaridis C, Liniger A, et al. P3Depth: monocular depth estimation with a piecewise planarity prior[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 19-24, 2022, New Orleans, USA. New York: IEEE Press, 2022: 1600-1611.

[30] Xie S N, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5987-5995.

[31] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning, July 6-11, 2015, Lille, France. New York: ACM Press, 2015: 448-456.

[32] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[M]//Navab N, Hornegger J, Wells W M, et al. Medical image computing and computer-assisted intervention-MICCAI 2015. Lecture notes in computer science. Cham: Springer, 2015, 9351: 234-241.

[33] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7132-7141.

[34] Wu T, Pan L, Zhang J, et al. Balanced chamfer distance as a comprehensive metric for point cloud completion[J]. Advances in Neural Information Processing Systems, 2021, 34: 29088-29100.

[35] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from RGBD images[M]//Fitzgibbon A, Lazebnik S, Perona P, et al. Computer vision-ECCV 2012. Lecture notes in computer science. Heidelberg: Springer, 2012, 7576: 746-760.

[36] Newcombe R A, Izadi S, Hilliges O, et al. KinectFusion: Real-time dense surface mapping and tracking[C]//2011 10th IEEE International Symposium on Mixed and Augmented Reality, October 26-29, 2011, Basel, Switzerland. New York: IEEE Press, 2012: 127-136.

[37] Levin A, Lischinski D, Weiss Y. Colorization using optimization [C]//SIGGRAPH '04: ACM SIGGRAPH 2004 Papers, August 8-12, 2004, Los Angeles, California. New York: ACM Press, 2004: 689-694.

[38] 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, June 20-25, 2009, Miami, FL, USA. New York: IEEE Press, 2009: 248-255.

[39] 沙浩, 刘越, 王涌天, 等. 基于二维图像和三维几何约束神经网络的单目室内深度估计方法[J]. 光学学报, 2022, 42(19): 1911001.
Sha H, Liu Y, Wang Y T, et al. Monocular indoor depth estimation method based on neural networks with constraints on two-dimensional images and three-dimensional geometry[J]. Acta Optica Sinica, 2022, 42(19): 1911001.

[40] Yu J H, Jiang Y N, Wang Z Y, et al. UnitBox: an advanced object detection network[C]//Proceedings of the 24th ACM international conference on Multimedia, October 15-19, 2016, Amsterdam, The Netherlands. New York: ACM Press, 2016: 516-520.

[41] Fang Z C, Chen X R, Chen Y H, et al. Towards good practice for CNN-based monocular depth estimation[C]//2020 IEEE Winter Conference on Applications of Computer Vision (WACV), March 1-5, 2020, Snowmass, CO, USA. New York: IEEE Press, 2020: 1080-1089.

[42] Laina I, Rupprecht C, Belagiannis V, et al. Deeper depth prediction with fully convolutional residual networks[C]//2016 Fourth International Conference on 3D Vision (3DV), October 25-28, 2016, Stanford, CA, USA. New York: IEEE Press, 2016: 239-248.

[43] Hao Z X, Li Y, You S D, et al. Detail preserving depth estimation from a single image using attention guided networks [C]//2018 International Conference on 3D Vision (3DV), September 5-8, 2018, Verona, Italy. New York: IEEE Press, 2018: 304-313.

[44] Fu H, Gong M M, Wang C H, et al. Deep ordinal regression network for monocular depth estimation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 2002-2011.

[45] Hu J J, Ozay M, Zhang Y, et al. Revisiting single image depth estimation: toward higher resolution maps with accurate object boundaries[C]//2019 IEEE Winter Conference on Applications of Computer Vision (WACV), January 7-11, 2019, Waikoloa, HI, USA. New York: IEEE Press, 2019: 1043-1051.

[46] Ramamonjisoa M, Lepetit V. SharpNet: fast and accurate recovery of occluding contours in monocular depth estimation [C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), October 27-28, 2019, Seoul, Korea (South). New York: IEEE Press, 2020: 2109-2118.

Monocular Depth Estimation Method Based on Plane Coefficient Representation with Adaptive Depth Distribution

Wang Jiajun, Liu Yue*, Wu Yuhui, Sha Hao, Wang Yongtian

Beijing Engineering Research Center of Mixed Reality and Advanced Display, School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, *China*

Abstract

Objective Obtaining scene depth is crucial in 3D reconstruction, autonomous driving, and other related tasks. Current methods based on lidar or time of flight (ToF) cameras are not widely applicable due to their high cost. In contrast, only employing a single RGB image to infer scene depth information is more cost-effective, which has broader potential for more applications. Inspired by the successful applications of deep learning methods in various ill-posed problems recently, many researchers tend to adopt convolutional neural networks to estimate reasonable and accurate monocular depths. However, most existing studies based on deep learning focus on how to enhance the feature extraction capability of the network, without attention paid to the distribution of image depths. Estimating the pixel distributions of images can not only improve the inference precision but also make the reconstructed 3D images more consistent with ground truth. Therefore, we propose a new adaptive depth distribution module, which allows the model to predict different depth distributions for each image during the training.

Methods The NYU Depth-v2 dataset created by New York University is employed. Overall, our model is built based on the encoder-decoder structure with skip connections, which has been proven to be able to guide image generation more effectively. An indirect representation of depth maps based on plane coefficient is also introduced to implicitly add the plane constraint in the depth estimation and obtain smoother depth estimation results in the plane region of the scene. Specifically, two sub-networks with different lightweight designs are adopted at the bottleneck and other upsampling stages in the network to enhance the model′s feature extraction capability. In addition, an adaptive depth distribution estimation module is also designed to estimate different depth distributions according to different input images, which makes the pixel distribution of depth maps closer to the ground truth. A two-stage training strategy is employed. In the first stage, we load the pretrained weights on ImageNet into the backbone network and optimize the model using the loss function only at the 2D level. In the second stage, we perform joint training through loss functions at both the 2D and 3D levels.

Results and Discussions Our study employs multiple metrics including root mean square error (RMSE), relative error (REL), and intersection over union (IoU) to qualitatively evaluate the inference ability of the proposed model. As shown in Table 1, the proposed lightweight network model outperforms most of the listed methods with only 46 M parameters, which proves the overall structure of the model is concise and effective. The visual comparison results of 3D depth reconstruction (Fig. 5) demonstrate that the proposed network can output smoother and more continuous depth predictions in planar regions, and reasonable predictions in the partially occluded or missing areas of planar regions. In terms of depth distribution, the carefully designed adaptive depth distribution module can make the predicted distribution fit better with the ground truth in the trend of the curve and can get a higher IoU rate compared with other methods (Fig. 6 and Table 3), thus indicating the effectiveness of the proposed module. Additionally, the lightweight network can balance accuracy and speed in real-time scenarios (Table 2), and yield good inference and reconstruction results. However, the proposed network has some limitations in recovering fine details of the depth predictions (Fig. 7), and thus how to design the network to recover more depth details while ensuring the model′s real-time prediction performance will be the focus of our future work.

Conclusions An innovative model based on plane coefficient representation with adaptive depth distribution for monocular image depth estimation tasks is presented. Qualitative and quantitative results obtained from the NYU Depth-v2 dataset and multiple comparative experiments demonstrate that the proposed method is capable of obtaining reasonable prediction results for planar regions in images with partial occlusions or small viewing angles. Additionally, the proposed depth distribution prediction module provides differentiated pixel distribution optimization for each image, which can make the model achieve pixel depth distribution prediction results closer to the real images. With its lightweight design, this method realizes a balance between inference speed and inference accuracy and is highly applicable in practical scenarios that require accuracy in real time, such as indoor virtual reality and human-computer interaction.

Key words machine vision; depth estimation; monocular three-dimensional reconstruction; neural network