第 42 卷 第 19 期/2022 年 10 月/光学学报 研究论文

光学学报

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

沙浩，刘越*，王涌天，卢晨光，赵梦泽

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

摘要 提出了编码器到解码器结构的深度卷积神经网络，并基于二维层面和三维层面共同约束网络从单目图像中学习深度。在二维图像层面，为了均衡网络提取到的浅层细节特征和深层语义特征，引入通道注意力机制，在相同尺度上为编码器特征与解码器特征添加权重连接；为了得到边缘细节信息更丰富的深度图，构建了尺度不变损失和基于图像金字塔的多尺度边缘损失。在三维几何层面，为了提高点云之间的几何一致性，基于空间中坐标点的局部和全局几何关系，构建了深度的全局几何约束损失和局部几何约束损失。在NYU Depth-v2数据集上将所提方法的结果与其他方法进行定量定性比较。结果表明本文方法可以估计出准确度和细节上表现更好的室内场景深度，实现了更为准确和平滑的单张图像三维重建效果。

关键词 成像系统；深度估计；卷积神经网络；单目三维重建；几何约束

中图分类号 TP301.6 **文献标志码** A **DOI**：10.3788/AOS202242.1911001

Monocular Indoor Depth Estimation Method Based on Neural Networks with Constraints on Two-Dimensional Images and Three-Dimensional Geometry

Sha Hao, Liu Yue*, Wang Yongtian, Lu Chenguang, Zhao Mengze

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

Abstract This paper proposes a deep convolutional neural network with an encoder-to-decoder structure and constrains the network's in-depth learning from the monocular image at both two-dimensional (2D) and three-dimensional (3D) levels. At the 2D image level, an attention mechanism of channels is introduced to connect encoder features with decoder features with weights at the same scale, so as to balance the shallow detail features and deep semantic features extracted by the network. In addition, a scale-invariant loss and a multi-scale edge loss based on image pyramids are designed to obtain a depth map with rich edge detail information. At the 3D geometric level, a global geometric constraint loss and a local geometric constraint loss of depth are designed based on the local and global geometric relationships of coordinate points in space, in a bid to enhance the geometric consistency between point clouds. Furthermore, the results obtained through the proposed method are quantitatively and qualitatively compared with that obtained through other methods from the NYU Depth-v2 dataset, and it is shown that the proposed method can estimate indoor scene depth with higher accuracy and detail representation, obtaining accurate and smooth 3D reconstruction results on a single image.

Key words imaging systems; depth estimation; convolutional neural network; monocular three-dimensional reconstruction; geometric constraint

收稿日期：2021-11-22；**修回日期**：2021-12-20；**录用日期**：2021-12-24

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)、广东省重点领域研发计划(2019B010139004)

通信作者：*liuyue@bit.edu.cn

1 引 言

深度是决定场景图像外观的重要本征属性之一，代表着场景的几何关系，因此获取深度已成为许多场景感知相关工作的基石，在三维重建[1-2]、障碍检测[3-4]、增强现实[5]等领域有着广泛的应用。传统方法通常采用激光雷达等硬件设备获取高精度的深度[6]，但受限于成本过高，无法得到广泛应用。相比之下，基于图像的估计方法只需要输入RGB图像就能得到所需的深度信息，无需价格高昂的硬件设备，因此有着更为广阔的应用前景。

根据输入图像的数量不同，基于图像估计场景深度的方法可分为多目深度估计和单目深度估计，其中多目深度估计方法大多基于图像之间的特征匹配和三角测量进行深度估计[7-8]，这些方法需要成对的图像或图像序列，不适用于只有单张图像输入的情况，因此更具有应用潜力的单目深度估计方法成为了研究热点。早期的单目深度估计大多以提取深度线索为手段[9-12]，精度较低。近年来，随着卷积神经网络在视觉任务上的广泛应用[13-16]，大量利用卷积神经网络进行单目深度估计的工作涌现出来[17]，大大提高了单目深度估计的精度。2014年Eigen等[18]提出了利用卷积神经网络恢复单目图像深度的方法，设计了由粗到细的两个子网络结构，其中粗网络聚焦学习深度的大致范围，细网络在粗网络结果的基础上改善边缘细节信息。他们还针对深度的尺度问题提出尺度不变误差损失函数，加速了网络的拟合。之后Laina等[19]将残差网络引入深度估计，提出逐级增大深度图分辨率的快速上采样块以增强网络对特征信息的提取能力，提高深度图的恢复准确率；Laina等还通过引入berHu损失函数提高了图像中深度较小区域的训练效率。但这些已有工作[18-22]大多只在二维图像的层面关注网络学习的好坏，更注重在网络结构、二维损失函数等方面的改进，忽略了深度信息本身包含的几何意义。

深度图中每个像素值的本质是三维坐标点位置在二维图像上的投影，因此基于深度值形成的几何关系，许多工作[23-27]为深度估计添加辅助先验信息，以获取更准确的深度。2015年Eigen等[24]首先将法线估计和深度估计相结合，在网络训练时对二者进行共享权重计算，再将估计出的深度和法线作为语义分割任务的输入，得出场景的分割结果。Eigen等的工作利用共享权重计算和更改输入的方式在隐式空间内加强了不同信息之间的浅约束，却忽略了这些信息在显式空间的强约束。Qi等[26]对深度和法线进行初步估计后，引入深度到法线和法线到深度的显式生成方法，让网络估计的粗深度和粗法线在局部范围内互相约束，进而得到第二阶段的优化结果。他们还在二维层面设计了边缘优化的第三阶段，增强最后结果的细节。上述这些工作虽然都以不同方式在网络训练时添加三维形式的先验信息，但都依赖于局部平面约束等三维坐标点中的某种不完全几何关系，忽略了其他难以满足其约束条件的位置，存在一定的局限性。

为了解决上述问题，本文提出了编码器到解码器结构的深度卷积神经网络，并在二维层面和三维层面共同约束网络从单目图像中学习深度。首先介绍了场景中三维坐标点的局部几何关系和全局几何关系，根据此关系设计了基于局部几何关系约束和基于全局几何关系约束的损失函数，与现有文献中的方法相比，本文设计的损失函数不仅在局部和全局范围内对不同点集之间的几何关系进行约束，而且计算速度更快，更易集成至现有深度学习框架[28]中。本文还在二维图像方面关注网络学习的质量，添加了基于图像金字塔的多尺度边缘损失函数；通过引入基于深浅层通道注意力的跳层连接模块，在将不同层次特征连接的同时均衡了二者的权重，增加了最后恢复结果的边缘细节。在NYU Depth v2测试数据集[29]上定量和定性的评估结果表明，与大多数现有方法相比，本文方法能得到细节更好、准确度更高的室内深度信息，实现了更为平滑的三维重建效果。

2 本文算法

2.1 深度满足的三维几何对应关系

在实践中常用针孔相机模型表示单张图像的成像原理。设图像中像素 i 的二维平面坐标为 (u_i, v_i)，将其投影至三维空间的点云坐标为 (x_i, y_i, z_i)，根据透视投影的几何原理，可以得到二者的关系为

$$\begin{cases} x_i = (u_i - c_x) z_i / f_x \\ y_i = (v_i - c_y) z_i / f_y \end{cases}, \tag{1}$$

式中：f_x 和 f_y 分别表示相机在水平和竖直方向的焦距；(c_x, c_y) 是主点坐标。表面法线代表着场景几何表面的方向，由处于同一表面的三维坐标点决定。要计算目标像素点 i 的表面法线，首先要根据公式计算像素点 i 的三维空间坐标，再找到像素点 i 所在三维空间的几何平面，进而通过目标平面计算法向量。对于二维图像中的大多数像素点，目标像素点 i 与邻近像素点同处一个几何平面，基于近邻像素点和目标像素点的位置就可计算出目标像素点 i 所在三维几何平面的法线，因此将三维坐标点的表面法线记为场景中近邻坐标点之间的局部几何关系。通过深度图计算表面法线图的具体步骤如下。

首先建立像素点 i 与近邻像素点 j 投影至三维空间的坐标点集：

$$\mathcal{N}_i = \left\{ (x_j, y_j, z_j) \middle| |u_i - u_j| < \beta, |v_i - v_j| < \beta, |z_i - z_j| < \gamma z_i \right\}, \tag{2}$$

式中：β 是图像平面上选取近邻像素范围的核大小；γ 是三维空间内选取近邻像素的深度范围超参数。设像素点 i 处的几何表面法线 $\boldsymbol{n} = [n_x, n_y, n_z]$，则像素点 i 所处平面的方程为

$$F(x, y, z) = n_x x + n_y y + n_z z - b = 0, \tag{3}$$

式中：b 是常数。根据上述理论，将像素点 i 与近邻像素点的三维坐标代入平面方程并经整理后得到超定线性方程组：

$$\begin{cases} \boldsymbol{A}\boldsymbol{n}=\boldsymbol{b} \\ \|\boldsymbol{n}\|_2^2=1 \end{cases}, \tag{4}$$

式中：$\boldsymbol{A}$ 是点集的矩阵表示，$\boldsymbol{A}=\begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \vdots & \vdots & \vdots \\ x_K & y_K & z_K \end{pmatrix}\in\mathbb{R}^{K\times 3}$，$K$ 是选取点集的大小；$\boldsymbol{b}\in\mathbb{R}^{K\times 1}$ 是元素均为 b 的常数向量。利用最小二乘法求解 $\|\boldsymbol{A}\boldsymbol{n}-\boldsymbol{b}\|^2$ 即可得到像素点 i 处的几何表面法线。在计算单张场景深度图像的表面法线时，采用最小二乘法会耗费极大的计算资源和较多时间，难以将其作为约束训练卷积神经网络，为此文献[26]将法线的计算简化为

$$\boldsymbol{n}=\frac{\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{1}}{\left\|\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{1}\right\|_2}, \tag{5}$$

式中：$\boldsymbol{1}\in\mathbb{R}^{K\times 1}$ 是所有元素都为 1 的常数向量。但从实验中发现利用式(5)计算表面法线存在两个问题：1)对每个像素点集矩阵的求逆运算量过大，如在 Pytorch[28] 等没有集成广播矩阵求逆计算的深度学习框架中计算单张分辨率为 480×640 图像的法线图需要耗时 2000 s 以上；2)该计算过程对非平滑深度图像的噪声鲁棒性差。由于该方法需要考虑所有近邻点集的位置，在不平滑的 NYU Depth v2 深度数据上计算得到的法线会出现许多噪声。

针对上述问题，本文设计了基于近邻像素点采样的法线图转换方法。仍然遵从近邻像素共处同一平面的假设，将像素点 i 投影至三维空间后建立近邻点集，如图 1 所示，只选取点集中距离像素 i 点为 r 的 4 个近邻像素点 A、B、C、D 计算向量 $\overrightarrow{BA}$ 和 $\overrightarrow{DC}$，则像素点 i 处的表面法线为

$$\boldsymbol{n}=\frac{\overrightarrow{BA}\times\overrightarrow{DC}}{\left\|\overrightarrow{BA}\times\overrightarrow{DC}\right\|}。 \tag{6}$$

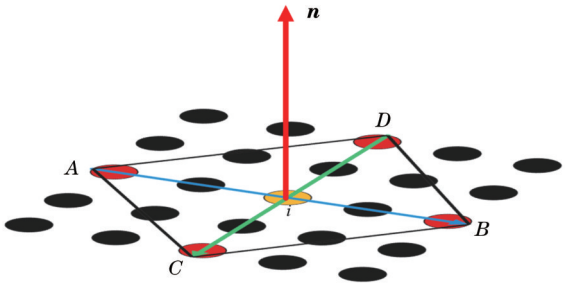

图 1　近邻点采样法的法线计算原理

Fig. 1　Principle of calculating the normal of nearest neighbor point sampling method

基于近邻点采样方法转换的法线图只考虑了近邻范围内关键点采样的位置，对于非平滑深度数据来说，生成法线图的准确度更高，对噪声的鲁棒性更强。在算法效率上，本文方法的时间复杂度只有 $O(1)$，在相同的运算条件下计算单张分辨率为 480×640 的图像的法线图只需要 0.02 s。

表面法线的计算只考虑了局部点集的相对位置，忽略了场景中非近邻点的全局位置关系。为解决这一问题，本文借鉴了文献[25]的方法，通过引入全局相关的虚拟法线将三维坐标点的虚拟法线记为场景中非近邻点集之间的全局几何关系。具体步骤为：首先随机从深度图中选取 N 组点集，每组点集包含三个点 P_{Ai}, P_{Bi}, P_{Ci}，则这些被选中点可以表示为

$$P=\left\{\left\{P_{A1},P_{B1},P_{C1}\right\},\cdots,\left\{P_{Ai},P_{Bi},P_{Ci}\right\},\cdots,\left\{P_{AN},P_{BN},P_{CN}\right\}\right\},\ i=0,1,\cdots,N。 \tag{7}$$

为保证每组选中的点不共线，则必须满足

$$\begin{aligned} \partial&\geqslant\angle\left(\overrightarrow{P_AP_B},\ \overrightarrow{P_AP_C}\right)\geqslant\alpha, \\ \partial&\geqslant\angle\left(\overrightarrow{P_BP_C},\ \overrightarrow{P_BP_A}\right)\geqslant\alpha\left\{P_{\mathrm{A/B/C}}\in P\right\}, \end{aligned} \tag{8}$$

式中：∂ 和 α 是控制向量夹角范围的超参数，本文将 ∂ 和 α 分别设为 120°和 30°。为保证每组选取点的相对位置范围更大，还要满足

$$\left\{\left|\overrightarrow{P_kP_m}\right|>\theta\middle|\ k,m\in[A,B,C],P_{\mathrm{A/B/C}}\in P\right\}, \tag{9}$$

式中：θ 是控制选中点距离的超参数，本文设为 0.6 m。对于选中的每组点集，可以确定一个平面，将基于此平面计算得到的法线称为虚拟法线，最终单张深度图像的虚拟法线 $\boldsymbol{N}_{\mathrm{v}}$ 可表示为

$$\boldsymbol{N}_{\mathrm{v}}=\left\{\boldsymbol{n}_i=\frac{\overrightarrow{P_{Ai}P_{Bi}}\times\overrightarrow{P_{Ai}P_{Ci}}}{\left|\overrightarrow{P_{Ai}P_{Bi}}\times\overrightarrow{P_{Ai}P_{Ci}}\right|}\middle|\ \left\{P_{Ai},P_{Bi},P_{Ci}\right\}\in P,i=0,\cdots,N\right\}。 \tag{10}$$

虚拟法线选取点集的条件与表面法线选取点集的条件相反，弥补了表面法线计算忽略的全局几何位置关系，因此虚拟法线和表面法线统称为深度的三维几何对应关系。

2.2　损失函数设计

以图像形式表现的深度不仅有着二维图像属性，还有着大多数普通图像不具有的三维几何属性。为了在三维层次和二维层次共同约束卷积神经网络拟合，本文设计了基于三维几何的损失函数和基于二维图像的损失函数。

1)基于三维几何的损失函数。根据 2.1 节描述的局部几何关系和全局几何关系，将深度 d 与表面法线 $\boldsymbol{n}$ 的转换过程记为：$\boldsymbol{n}=\mathcal{T}_{\mathrm{L}}(d)$，深度 d 与虚拟法线 $\boldsymbol{n}_{\mathrm{v}}$ 的转换过程记为：$\boldsymbol{n}_{\mathrm{v}}=\mathcal{T}_{\mathrm{G}}(d)$。局部三维约束损失函数 L_{L} 在三维空间的近邻点范围内衡量三维坐标点的几何一致性，可表示为

$$L_{\mathrm{L}}=\frac{1}{N}\sum_{i=0}^{N}\left[\mathcal{T}_{\mathrm{L}}\left(d_i^{\mathrm{pred}}\right)-\mathcal{T}_{\mathrm{L}}\left(d_i^{\mathrm{gt}}\right)\right]^2,\tag{11}$$

式中：N是像素个数；d_i^{pred}和d_i^{gt}分别代表在像素i所处深度的预测值和真实值。全局三维约束损失函数L_{G}在非近邻点超大范围内衡量三维坐标点的几何一致性，可表示为

$$L_{\mathrm{G}}=\frac{1}{N_1}\sum_{i=0}^{N_1}\left(\boldsymbol{n}_{vi}^{\mathrm{pred}}-\boldsymbol{n}_{vi}^{\mathrm{gt}}\right)^2,\tag{12}$$

式中：N_1是计算虚拟法线采样得到点集的个数；$\boldsymbol{n}_{vi}^{\mathrm{pred}}$和$\boldsymbol{n}_{vi}^{\mathrm{gt}}$分别是预测深度图和真实深度图的虚拟法线。

2）基于二维图像的损失函数。该函数主要用来衡量预测图像和真实图像之间在二维层次上的差异，包括像素的值、图像的边缘信息等。由于深度存在全局尺度模糊问题，本文引入与文献[18]中类似的尺度不变误差损失作为深度估计任务中的主损失函数：

$$L_{\mathrm{SCV}}=\frac{1}{N}\sum_{i=0}^{N}\left(d_i^{\mathrm{pred}}-d_i^{\mathrm{gt}}\right)^2-\frac{1}{2N^2}\left[\sum_{i=0}^{N}\left(d_i^{\mathrm{pred}}-d_i^{\mathrm{gt}}\right)\right]^2。\tag{13}$$

为了获取边缘细节更清晰的结果，本文还设计了基于图像金字塔[30]的多尺度边缘损失函数，在不同的二维尺度空间衡量预测值和真实值边缘的差异，将其表示为

$$L_{\mathrm{GD}}=\sum_{l=0}^{L}\left[\frac{1}{N_l}\sum_{i=1}^{N_l}\left(\nabla\boldsymbol{d}_{l,i}^{\mathrm{pred}}-\nabla\boldsymbol{d}_{l,i}^{\mathrm{gt}}\right)^2\right],\tag{14}$$

式中：$\boldsymbol{d}_{l,i}^{\mathrm{gt}}$和$\boldsymbol{d}_{l,i}^{\mathrm{pred}}$分别表示在像素$i$和图像金字塔的第$l$尺度处深度的真实值和预测值，$\nabla(\cdot)$为图像边缘计算符；$N_l$是第$l$尺度图像的有效像素数，当$l=1$时，$N_l$为原始尺度图像的有效像素数，本文在4个图像尺度下计算边缘损失函数。

2.3　基于深浅层通道注意力机制的特征连接

浅层神经网络提取的图像特征包含更多与原图像更接近的细节信息，深层神经网络提取到的图像特征则以最终任务为导向，包含更多语义信息。因此，对于逐像素的密集预测任务，添加类似于U-net[31]的相同尺度跳层对称连接不仅会加速网络的拟合，还会改善预测结果的边缘细节。但大多数研究工作只单纯用级联的方式合并深、浅层特征，没有考虑不同层特征对最终结果所占的比重大小。为此，本文设计了基于深浅层通道注意力机制[32]的特征连接模块（SE_Concat_Block），在网络末端的深层特征处添加了类似U-net结构中浅层特征的跳跃连接，并为来自不同层合并后的图像特征添加可学习权重，以使网络在学习中有选择性地利用来自不同层次的图像特征。

基于深浅层通道注意力机制的特征连接模块如图2所示，相同尺度的浅层特征X_1和深层特征X_2首先通过级联的方式进行通道合并，合并后的特征为Y，然后利用挤压函数$F_{\mathrm{sq}}(\cdot)$对Y进行计算以将全局信息转换为通道描述子$\boldsymbol{z}\in\mathbb{R}^C$，上标$C$表示全局特征通道，$F_{\mathrm{sq}}(\cdot)$代表相应通道特征的全局平均池化，第$c$个通道的描述子$z_c$可表示为

$$z_c=F_{\mathrm{sq}}(Y_c)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}Y_c(i,j),\tag{15}$$

式中：H和W分别为特征图像的纵向像素数和横向像素数；Y_c为特征Y在第c个通道的特征分量。

为了利用挤压后通道描述子的全局信息，引入了激励函数$F_{\mathrm{ex}}(\cdot,W)$对描述子进行自适应再校准，进而获取通道的权重$\boldsymbol{s}$，可表示为

$$\boldsymbol{s}=F_{\mathrm{ex}}(\boldsymbol{z},W)=\sigma\left[W_2\delta(W_1\boldsymbol{z})\right],\tag{16}$$

式中：$\sigma(\cdot)$为Sigmod激活函数；$\delta(\cdot)$为ReLU激活函数；$W_1\in\mathbb{R}^{\frac{C}{r}\times C}$和$W_2\in\mathbb{R}^{C\times\frac{C}{r}}$是可学习参数。自适应再校准操作的本质是将描述子依次输入一个降维的全连接层、一个ReLU层和一个升维的全连接层，输出C维的通道权重$\boldsymbol{s}$。最后，引入缩放函数$F_{\mathrm{scale}}(\cdot,\cdot)$对合并后的特征$Y$进行缩放，得到最终的输出特征：

$$\tilde{Y}_c=F_{\mathrm{scale}}(Y_c,\boldsymbol{s}_c)=\boldsymbol{s}_cY_c,\tag{17}$$

式中：$\tilde{Y}_c$是特征$\tilde{Y}$在第c个通道的特征分量；$\boldsymbol{s}_c$是权重参数$\boldsymbol{s}$在第c个通道的分量值；$F_{\mathrm{scale}}(\cdot,\cdot)$的本质是各个通道的权重参数标量与相应通道特征的逐像素乘积。深浅层特征通道的注意力机制相当于为合并后的深浅层特征引入以不同通道统计信息为条件的权重，帮助网络更有侧重地提取不同意义的图像特征。

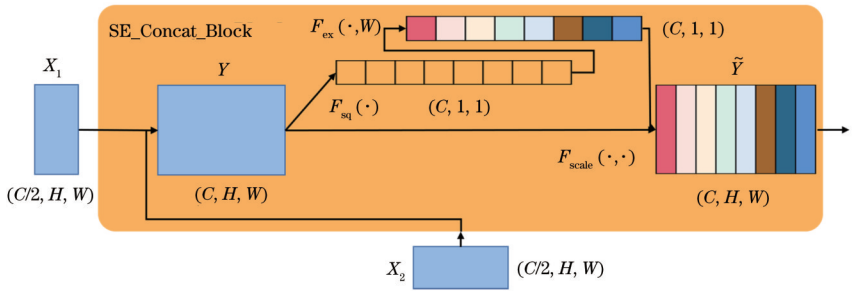

图2　基于深浅层通道注意力机制的特征连接模块

Fig. 2　Feature connection module based on depth channel attention mechanism

2.4　单目深度估计方法整体架构

本文方法的整体架构如图3所示，将单目图像输入神经网络的编解码器，输出网络估计的深度，然后将深度投影至三维空间形成点云，进而通过计算真实深度图与估计深度图之间的二维图像损失、真实点云与估计点云之间的三维几何损失并将总误差反向传播至

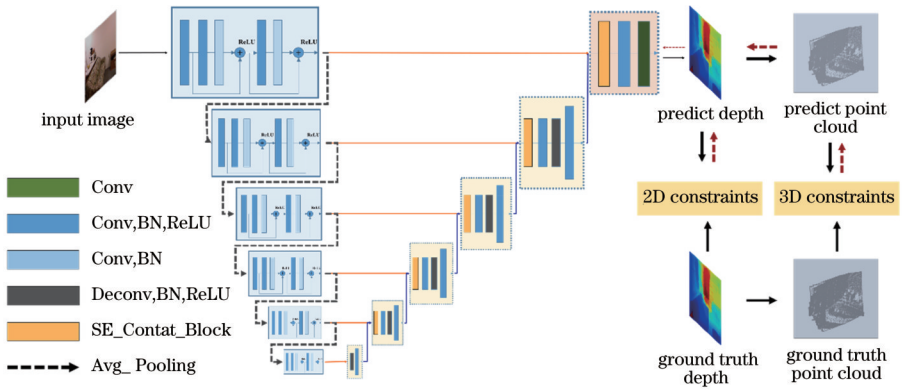

图 3 单目深度估计方法整体架构

Fig. 3 Overall architecture of monocular depth estimation method

网络以不断调整网络参数，迭代此过程直至网络估计深度与真实深度之间的误差最小，完成训练。在测试时，输入单目图像，网络即可估计出场景深度。

本文设计的卷积神经网络主要由编码器和解码器组成，输入图像经过编码器输出 6 个不同尺度的浅层图像特征，编码器每个尺度的子网络结构如图 4(a)所示，主要由残差结构的卷积层组成，每个卷积层的卷积核大小均为 3×3，第 1、2、4 个卷积层后紧跟批量归一化(BN)层[33]和 ReLU 激活函数，第 3、5 个卷积层紧跟 BN 层，随后与隔层的特征相加再进行 ReLU 激活，最后通过平均池化下采样至下个尺度。解码器不同尺度的子网络结构如图 4(b)～(d)所示，主要由卷积层、上采样层和基于深浅层通道注意力机制的连接块组成，除最小尺度的子解码器[图 4(b)]外，其他子解码器均接收来自上层子解码器的深层特征和同尺度编码器的浅层特征，并将依次经过基于深浅层通道注意力机制的连接块和通道压缩一半的卷积层作为本尺度的深度特征。除最大尺度的子解码器[图 4(d)]外，深度特征会经过上采样层作为上个尺度子解码器的部分输入。解码器中卷积层的卷积核大小与编码器相同，除最后一层只由单层卷积构成外，其余均紧跟 BN 层和 ReLU 激活函数，上采样层依次由反卷积层[34]、BN 层和 ReLU 激活函数构成。

根据本文 2.2 节所设计的损失函数，二维图像损失 L_{2D} 和三维几何损失 L_{3D} 分别可表示为

$$\begin{cases} L_{2D}=\lambda_1 L_{SCV}+\lambda_2 L_{GD} \\ L_{3D}=\lambda_3 L_G+\lambda_4 L_L \end{cases}, \tag{18}$$

式中：λ_1、λ_2、λ_3 和 λ_4 分别是不同损失函数的权重常数。本文方法的整体训练步骤为在网络初始化后首先设置 $\lambda_1=1$，$\lambda_2=0.7$，利用二维图像损失 L_{2D} 对本文所设计的神经网络进行预训练，当网络不再收敛时完成预训练；其次载入预训练参数，设置 $\lambda_1=1$，$\lambda_2=0.7$，$\lambda_3=0.5$，$\lambda_4=0.9$，添加三维损失 L_{3D}，其与 L_{2D} 一起对网络进行联合迁移训练，直至完成。

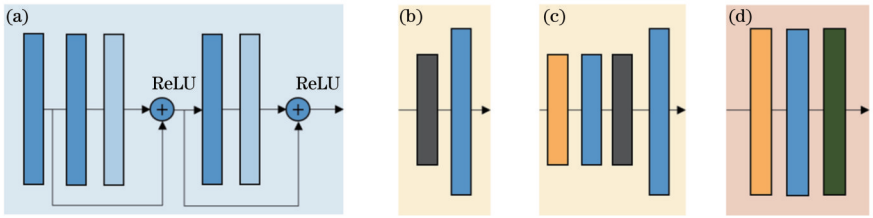

图 4 编解码器子网络结构。(a)编码器的子网络结构；(b)～(d)解码器的子网络结构

Fig. 4 Architecture of encoder and decoder sub-networks. (a) Sub-network structure of encoder; (b)-(d) subnetwork structures of decoder

3 实验结果及分析

3.1 实施细节

本文方法主要在 PyTorch 深度学习框架下实现，计算平台搭载的 CPU 型号为 i9-10850k，GPU 型号为 NVIDIA GeForce GTX TITANX。NYU Depth v2[29]是利用 Kinect 深度相机拍摄的单目深度数据集，总共包含 464 个室内场景，实验根据 NYU Depth v2 官方的训练集与测试集的划分方式使用 249 个场景数据对网络进行训练，使用 215 个场景数据对本文算法进行测试，训练集的深度数据来自训练场景的原始帧采样图。对于彩色图像，其原始分辨率为 640×480，将其分辨率缩至 256×192 作为网络输入。对于原始训练数据缺失的深度值，采用官方提供的工具包[35]填补空缺位置。本文实验所用训练集包含 30000 张图像对，测试集为官方处理好的 694 张图像对。

实验采用 Adam 优化器[36]对本文设计网络进行训练，在第一次预训练时设置批量大小(batch size)为 16，初始学习率为 0.0001，用二维约束训练 50 个周期。预训练完成后，加载预训练网络模型参数，添加三维约

束对网络进行迁移训练，设置批量大小为8，初始学习率为0.0001，共训练25个周期。最后，在测试集上对训练好的模型进行定性定量的评估。

3.2 在NYU Depth v2数据集上的实验结果与分析

为了定量评估结果的好坏，本文首先引入如下常用评价指标。

1）均方根误差（RMSE）：

$$V_{\mathrm{RMSE}}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i-y_i^*\right)^2},\tag{19}$$

式中：y_i和y_i^*分别代表深度的估计值和真实值；N是像素个数。

2）绝对值相对误差（REL）：

$$V_{\mathrm{REL}}=\frac{1}{N}\sum_{i=1}^{N}\frac{\left|y_i-y_i^*\right|}{y_i^*}。\tag{20}$$

3）阈值内准确比率（TH）（$\delta<X_{\mathrm{thr}}$）：

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

阈值内准确比率是图像中满足$\delta<X_{\mathrm{thr}}$条件的像素点比率，通常取X_{thr}为1.25、1.25^2和1.25^3。RMSE和REL用于衡量估计深度和真实深度之间的绝对误差和相对误差，其数值越小，准确度越高；TH用于衡量估计深度和真实深度在一定范围内误差的比率大小，其数值越大，准确度越高。

本文方法与其他文献方法的定量定性比较结果如表1和图5所示。在室内数据集上本文方法的RMSE超过了所列的所有方法，这证明本文方法可以从单目室内图像中估计出平均准确度很高的深度值。从图5中可以看出：本文方法估计的深度相比其他文献中的结果具有更清晰的边缘，深度值也更加平滑，尤其在一些几何形状比较复杂的区域，本文方法可以恢复出其他方法很难区分的深度细节。

表1 所提方法与其他不同方法在NYU Depth v2数据集上的定量比较

Table 1 Quantitative comparison between proposed method and other different methods on NYU Depth v2 dataset

Method	RMSE	REL	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Ref. [18]	0.907	0.215	0.611	0.887	0.971
Ref. [37]	0.824	0.230	0.614	0.883	0.971
Ref. [38]	0.620	0.149	0.806	0.883	0.987
Ref. [39]	0.635	0.143	0.788	0.958	0.991
Ref. [40]	0.819	0.232	0.646	0.892	0.968
Ref. [24]	0.641	0.158	0.769	0.950	0.988
Ref. [19]	0.573	0.127	0.811	0.953	0.988
Ref. [22]	0.586	0.121	0.811	0.954	0.987
Ref. [26]	0.600	0.144	0.791	0.960	0.991
Ref. [41]	0.572	0.139	0.815	0.963	0.991
Ref. [27]	0.599	0.159	0.772	0.942	0.984
Ref. [37]	0.555	0.126	0.843	0.968	0.991
This paper	0.552	0.164	0.768	0.940	0.984

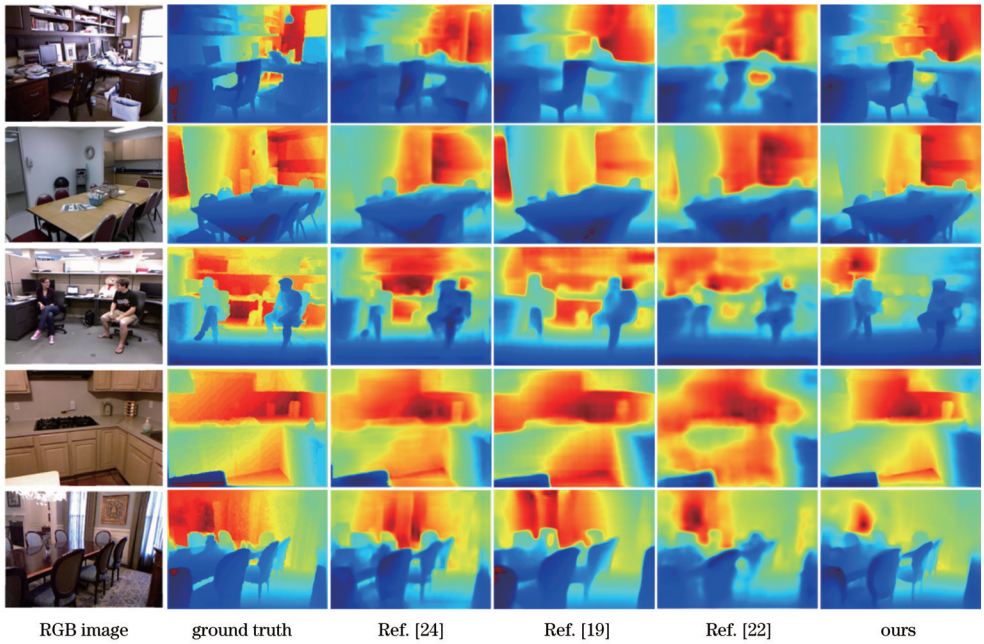

图5 不同方法在NYU Depth v2数据集上的深度估计结果

Fig. 5 Depth prediction results of different methods on NYU Depth v2 dataset

本文设计了基于三维几何的约束方法。与其他基于二维约束的方法相比，本文方法的预测结果的几何一致性与原场景更接近，为此基于单目深度重建出了三维的室内场景图像。如图6所示：文献[18]的重建场景出现了明显的畸变，尤其是在几何形状较复杂的沙发区域；文献[19]的重建场景虽然远好于文献[18]，但与本文方法的三维重建结果相比，文献[19]的重建场景与原场景空间的几何一致性略低；本文方法采用

的三维几何约束对噪声具有鲁棒性，因此本文所得结果比基于真实深度值的重建结果更平滑，在图中具体表现为：基于真实深度重建的墙面和地面具有许多坑注，而基于本文方法的重建结果具有平整的表面，与现实场景更加一致。

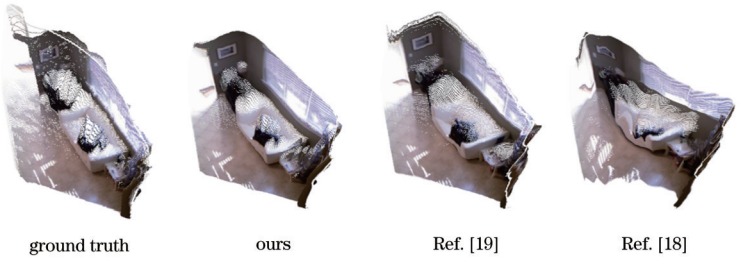

图 6　基于单目深度的三维重建结果

Fig. 6　3D reconstruction results based on monocular depth

在模型运行速度上，本文方法与其他方法的比较如表 2 所示，可以看出：本文模型的运行帧率可以达到 17 frame/s，优于文献[19]和文献[24]中方法；文献[18]中方法的运行速度较快，但其准确率远低于本文方法，证明本文模型可以在速度和精度上达到平衡，可以满足单张图像实时三维重建的需求。

表 2　不同方法的运行速度比较

Table 2　Comparison of running speeds of different methods

Method	Runing time /ms	Frame rate /(frame·s^{-1})	RMSE
Ref. [18]	23	43	0.907
Ref. [19]	237	10	0.604
Ref. [24]	96	6	0.753
This paper	58	17	0.552

3.3　消融实验

为了更好地验证本文方法每个模块的有效性，同样在 NYU Depth v2 数据集上对本文方法的组成部分进行了消融实验。为保证每个消融实验其他变量的一致性，包括学习率、batch size 和训练周期在内的训练细节需同上文保持一致。首先对网络的整体结构进行消融实验，依次去除原网络结构中的 SE_Concat_Block 和跳层连接模块，只使用二维图像损失对网络进行训练。添加类似于 U-net 的跳层连接在改善定量结果的同时极大增加了最终估计结果的细节表现；添加了 SE_Concat_Block 的基线网络得到的定量比较的各项指标都超过了未添加 SE_Concat_Block 的网络，表明了该结构的有效性。为了证明本文整体网络结构的性能，同样只使用二维图像损失对 U-net 网络和以 Resnet-101 为骨干编码器的网络[42]进行训练，定量定性的结果如表 3 和图 7 所示。基线网络结构在各项定量指标和定性结果中都大大超过了经典的 Resnet-101 网络和 U-net 网络，证明了本文整体网络结构在单目室内深度估计任务上的优异性。

表 3　基于网络结构消融实验的定量结果

Table 3　Quantitative results of ablation experiments based on network architecture

Method	RMSE	REL	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Without skip connect	0.727	0.222	0.631	0.885	0.969
Without SE_Concat_Block	0.604	0.177	0.731	0.922	0.976
Baseline	0.586	0.178	0.738	0.932	0.982
U-net	0.647	0.202	0.681	0.915	0.978
Resnet-101	0.628	0.189	0.704	0.921	0.981

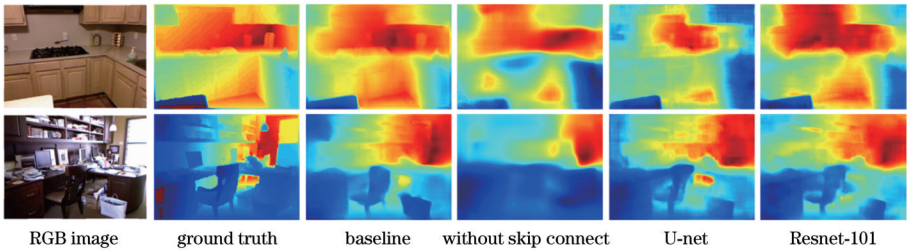

图 7　基于网络结构消融实验的定性结果

Fig. 7　Qualitative results of ablation experiments based on network architecture

对本文设计的二维图像约束和三维几何约束进行消融实验，将只使用尺度不变误差训练的网络作为参考基线，定量定性的结果如表 4 和图 8 所示。可以看出：添加了二维图像约束的网络可以让估计深度值更趋于稳定，不会出现类似图 8 中基线实验在某些位置的坏点，提升了整体结果的准确率；添加三维约束的网络会大大提高深度估计的准确率，具体表现为全局三维几何损失在提升结果的同时会让估计深度变得更平滑，局部三维损失则在提升结果的同时保持更多几何细节。

表 4　基于约束消融实验的定量结果

Table 4　Quantitative results of ablation experiments based on constraints

Method	RMSE	REL	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Baseline	0.594	0.177	0.740	0.926	0.980
With L_{2D}	0.586	0.178	0.738	0.932	0.982
With L_{2D} and L_G	0.561	0.165	0.761	0.935	0.983
With L_{2D}, L_G, and L_L	0.552	0.164	0.768	0.940	0.984

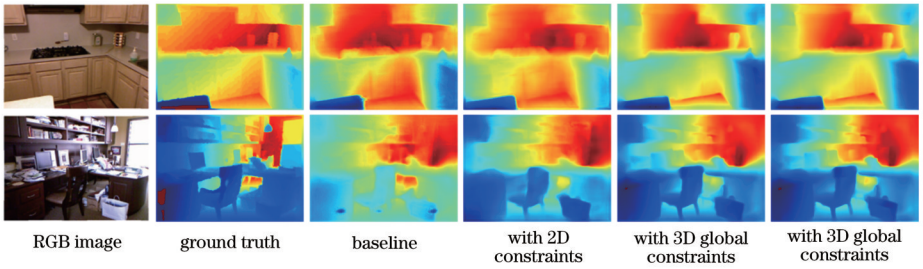

图 8　基于约束消融实验的定性结果

Fig. 8　Qualitative results of ablation experiments based on constraints

3.4　误差分析

为探究各个定量指标表现不均衡和影响各个定量指标质量的原因，本文绘制了测试集不同深度值范围内定量结果的统计图，如图 9 所示。选取了在 RMSE、REL 和 TH1 三个定量指标上分别表现最不佳的各 10 张图像，绘制了不同深度值范围内定量结果的统计图，如图 10(a)～(c)所示。其中 prob 表示所在深度范围内的像素概率，RMSE、REL、TH1、TH2、TH3 分别表示所在深度范围内的各个定量指标，其中 TH1、TH2、TH3 分别为 X_{thr} 为 1.25、1.25^2、1.25^3 对应的 TH。从图 9 可以看出，测试集图像的深度值主要集中在 1～5 m，总体呈长尾分布；各深度范围的 RMSE 分布较为平均，具体在边缘深度范围的 RMSE 低于中间范围，深度值较小范围的 RMSE 低于深度值较大范围；REL 总体随深度的增大而逐渐降低，深度分布概率较大区域的 REL 与正常分布趋势相比明显下降；各深度范围内的三个阈值准确比率分布较为一致，遵循各个范围内的深度概率分布。

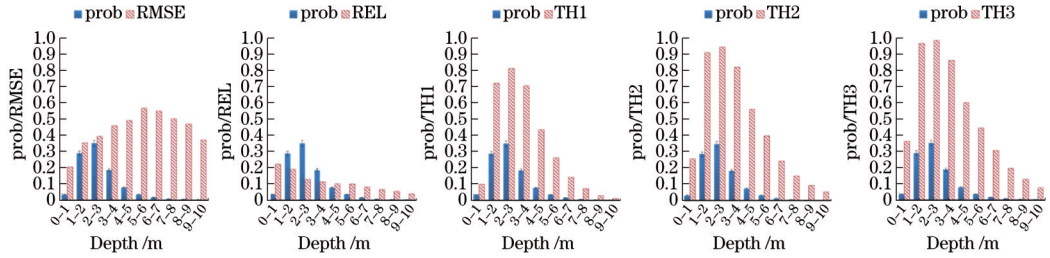

图 9　测试集不同深度值范围内的定量结果

Fig. 9　Quantitative results of test set in range of different depth values

从 RMSE 较高的图像[图 10(a)]来看，其远距离深度值的百分比远大于测试集的平均水平，深度布局也有明显不同，这表明数据之间的分布差异导致了 RMSE 升高；从 REL 较高的图像[图 10(b)]来看，其近距离深度值的百分比较大，同时更近距离范围内的 REL 也更高，这表明近距离场景导致了平均 REL 的升高；从 TH1 较低的图像[图 10(c)]来看，相较于图 9(a)、(c)，深度值分布与平均水平更接近，其他定量指标表现也较好，说明这些图像在各个深度范围的平均误差较大但接近，测试场景的特殊性是导致 TH1 降低的可能原因。总的来看，在遇到与深度分布平均水平较为一致的场景时估计结果表现优异，表明提升数据集的多样性可以显著改善估计结果，因此制作场景更丰富、更均衡的数据集应是未来工作关注的重点。

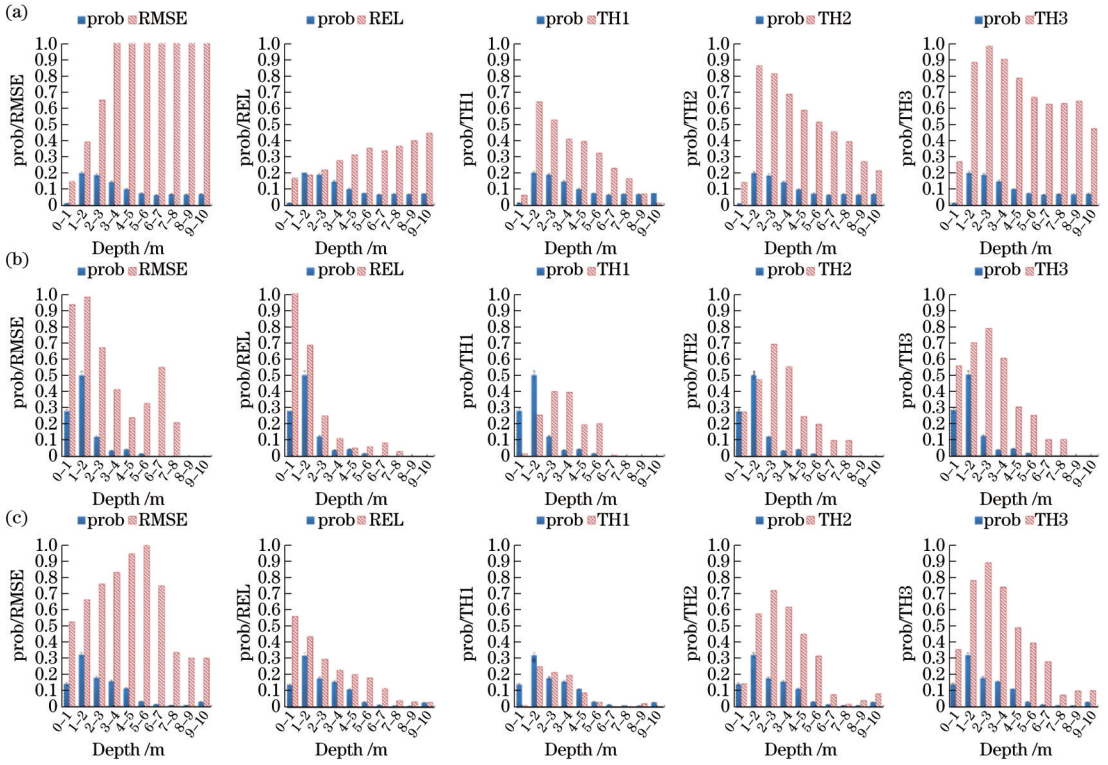

图 10 选定图像不同深度值范围内的定量结果。(a) RMSE最差的10张图像;(b) REL最差的10张图像;(c) TH1最差的10张图像

Fig. 10 Quantitative results of selected images in range of different depth values. (a) 10 images with worst RMSE; (b) 10 images with worst REL; (c) 10 images with worst TH1

4 结 论

针对单目深度估计任务,考虑到深度图的二维图像属性和三维几何属性,提出了基于二维图像约束和三维几何约束卷积神经网络的单目深度估计方法。为均衡浅层细节特征和深层语义特征对估计结果的影响,设计了基于深浅层通道注意力机制的特征连接结构,提高了网络的性能,同时设计了基于特征金字塔的边缘损失、局部三维损失和全局三维损失联合约束网络训练。结果表明:基于特征金字塔的边缘损失可以让网络估计更稳定,全局三维损失可以提升预测结果的平滑性,局部三维损失可以增加预测结果的几何细节。在NYU Depth v2数据集上本文方法与其他方法的定性定量对比表明,本文方法估计出室内深度值的平均准确度、几何一致性和平滑度更高,可以实现高质量的单目室内三维重建效果。最后的分析结果表明,提升数据集的多样性可以显著改善估计结果,制作场景更丰富、更均衡的数据集应是未来工作关注的重点。

参 考 文 献

[1] Izadi S, Kim D, Hilliges O, et al. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera[C]//UIST′11: Proceedings of the 24th annual ACM symposium on User interface software and technology, October 16-19, 2011, Santa Barbara, CA, USA. New York: ACM Press, 2011: 559-568.

[2] 郑太雄, 黄帅, 李永福, 等. 基于视觉的三维重建关键技术研究综述[J]. 自动化学报, 2020, 46(4): 631-652.
Zheng T X, Huang S, Li Y F, et al. Key techniques for vision based 3D reconstruction: a review[J]. Acta Automatica Sinica, 2020, 46(4): 631-652.

[3] 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.

[4] 郭克友, 杨民, 张沫, 等. 基于透视*N*点模型的实时单目深度估计方法[J]. 激光与光电子学进展, 2021, 58(6): 0615005.
Guo K Y, Yang M, Zhang M, et al. Real-time monocular depth estimation method based on perspective *N*-point model[J]. Laser & Optoelectronics Progress, 2021, 58(6): 0615005.

[5] Meka A, Fox G, Zollhöfer M, et al. Live user-guided intrinsic video for static scenes[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(11): 2447-2454.

[6] 陆俣, 何伟基, 邬淼, 等. 光子计数激光雷达的时间相关卡尔曼深度估计[J]. 光子学报, 2021, 50(3): 0311001.
Lu Y, He W J, Wu M, et al. Time-correlated Kalman depth estimation of photon-counting lidar[J]. Acta Photonica Sinica, 2021, 50(3): 0311001.

[7] Palomer A, Ridao P, Forest J, et al. Underwater laser scanner: ray-based model and calibration[J]. IEEE/ASME Transactions on Mechatronics, 2019, 24(5): 1986-1997.

[8] Gu C J, Cong Y, Sun G. Three birds, one stone: unified laser-based 3-D reconstruction across different media[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-12.

[9] Zhang R, Tsai P S, Cryer J E, et al. Shape-from-shading: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21(8): 690-706.

[10] Asada N, Fujiwara H, Matsuyama T. Edge and depth from focus[J]. International Journal of Computer Vision, 1998, 26(2): 153-163.

[11] Favaro P, Soatto S. A geometric approach to shape from defocus[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(3): 406-417.

[12] 刘晓旻, 杜梦珠, 马治邦, 等. 基于遮挡场景的光场图像深度估计方法[J]. 光学学报, 2020, 40(5): 0510002.
Liu X M, Du M Z, Ma Z B, et al. Depth estimation method of light field image based on occlusion scene[J]. Acta Optica Sinica, 2020, 40(5): 0510002.

[13] Wu X W, Sahoo D, Hoi S C H. Recent advances in deep learning for object detection[J]. Neurocomputing, 2020, 396: 39-64.

[14] Taghanaki S A, Abhishek K, Cohen J P, et al. Deep semantic segmentation of natural and medical images: a review[J]. Artificial Intelligence Review, 2021, 54(1): 137-178.

[15] Ciaparrone G, Sánchez F L, Tabik S, et al. Deep learning in video multi-object tracking: a survey[J]. Neurocomputing, 2020, 381: 61-88.

[16] Anwar S, Khan S, Barnes N. A deep journey into super-resolution[J]. ACM Computing Surveys, 2021, 53(3): 1-34.

[17] Zhao C Q, Sun Q Y, Zhang C Z, et al. Monocular depth estimation based on deep learning: an overview[J]. Science China Technological Sciences, 2020, 63(9): 1612-1627.

[18] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network [C]//NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2, December 8-13, 2014, Montreal, Quebec, Canada. Cambridge: MIT Press, 2014: 2366-2374.

[19] 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.

[20] Wang P, Shen X H, Lin Z, et al. Towards unified depth and semantic prediction from a single image[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA. New York: IEEE Press, 2015: 2800-2809.

[21] Liu P, Zhang Z H, Meng Z Z, et al. Monocular depth estimation with joint attention feature distillation and wavelet-based loss function[J]. Sensors, 2020, 21(1): 54.

[22] Xu D, Ricci E, Ouyang W L, et al. Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 161-169.

[23] Liu B, Gould S, Koller D. Single image depth estimation from predicted semantic labels[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 13-18, 2010, San Francisco, CA, USA. New York: IEEE Press, 2010: 1253-1260.

[24] 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, December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 2650-2658.

[25] 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, 2019: 5683-5692.

[26] Qi X J, Liu Z Z, Liao R J, et al. GeoNet: iterative geometric neural network with edge-aware refinement for joint depth and surface normal estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(2): 969-984.

[27] Yu Z H, Jin L, Gao S H. P^2Net: patch-match and plane-regularization for unsupervised indoor depth estimation [M]//Vedaldi A, Bischof H, Brox T, et al. Computer Vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12369: 206-222.

[28] Paszke A, Gross S, Chintala S, et al. Automatic differentiation in pytorch[EB/OL]. (2017-10-28)[2021-11-22]. https://openreview.net/forum?id=BJJsrmfCZ.

[29] Silberman N, Fergus R. Indoor scene segmentation using a structured light sensor[C]//2011 IEEE International Conference on Computer Vision Workshops, November 6-13, 2011, Barcelona, Spain. New York: IEEE Press, 2011: 601-608.

[30] Yang J C, Yu K, Gong Y H, et al. Linear spatial pyramid matching using sparse coding for image classification[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE Press, 2009: 1794-1801.

[31] 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.

[32] 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.

[33] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift

[C]// Proceedings of the 32nd International Conference on Machine Learning, {ICML} 2015, July 6-11, 2015, Lille, France. Cambridge: JMLR, 2015: 448-456.

[34] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[EB/OL]. (2015-11-19)[2021-02-05]. https://arxiv.org/abs/1511.06434.

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

[36] Kingma D P, Ba J. Adam: a method for stochastic optimization[EB/OL]. (2014-12-22)[2021-02-04]. https://arxiv.org/abs/1412.6980.

[37] Liu F Y, Shen C H, Lin G S. Deep convolutional neural fields for depth estimation from a single image[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 5162-5170.

[38] Chakrabarti A, Shao J Y, Shakhnarovich G. Depth from a single image by harmonizing overcomplete local network predictions[EB/OL]. (2016-05-23)[2021-02-05]. https://arxiv.org/abs/1605.07081.

[39] Jun L, Can Y C, Klein R, et al. A two-streamed network for estimating fine-scaled depth maps from single RGB images[J]. Computer Vision and Image Understanding, 2019, 186: 25-36.

[40] Cao Y, Wu Z F, Shen C H. Estimating depth from monocular images as classification using deep fully convolutional residual networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(11): 3174-3182.

[41] Lee J H, Heo M, Kim K R, et al. Single-image depth estimation based on Fourier domain analysis[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: 330-339.

[42] 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, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.