融合扩张卷积网络与 SLAM 的无监督单目深度估计

戴仁月，方志军*，高永彬
上海工程技术大学电子电气工程学院，上海 201600

摘要 针对由一般卷积神经网络预测的粗糙特征生成的深度图质量低、监督学习处理任务限制数据量等问题，提出一种新颖的融合扩张卷积神经网络和同时定位与建图(SLAM)的无监督单目深度估计方法。该方法采用视图重构的思想估计深度，利用光学一致性误差约束网络训练，扩大感受野，考虑图片细节特征。同时采用 SLAM 算法优化相机姿态，并将其嵌入视图重构框架中，实现单目图片与其深度图的直接映射。利用该方法在公开的 KITTI 数据集上进行实验，结果表明，与经典的 sfmlearner 方法相比，误差度量指标绝对差、平方差、均方差和对数均方差分别降低了 0.032、0.634、1.095 和 0.026；准确率度量指标 δ_1、δ_2 和 δ_3 分别提升了 3.8%、2.6%和 0.9%。该模型的可用性与稳健性得到验证。

关键词 图像处理；扩张卷积神经网络；同时定位与建图；无监督学习；单目视觉；深度估计
中图分类号 TP391 **文献标志码** A **doi**：10.3788/LOP57.061007

Unsupervised Monocular Depth Estimation by Fusing Dilated Convolutional Network and SLAM

Dai Renyue, Fang Zhijun*, Gao Yongbin
School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201600, *China*

Abstract The quality of a depth map generated by coarse features which are predicted by convolutional neural networks (CNNs) is low. Meanwhile, strong-supervised methods strictly limit the data volume due to lack of labeling. To address these problems, an unsupervised monocular depth estimation method by fusing dilated convolutional neural network and simultaneous localization and mapping (SLAM) is proposed. This method adopts the idea of view reconstruction to estimate depth. Photo-consistency error is utilized in the method to constrain training, expand the field of view, and concern the image details. Traditional SLAM algorithm functions to globally optimize the camera pose and incorporate it into the reconstruction framework. Finally the straight correspondence between the input monocular image and its depth map is exploited. The method is evaluated on the public KITTI dataset. The evaluation results show that, compared with the classical sfmlearner method, the error indicators, including absolute relative difference, squared relative difference, root mean squared error, and log root mean squared error, decrease by 0.032, 0.634, 1.095, and 0.026 respectively, and the accuracy indicators, δ_1, δ_2 and δ_3, increase by 3.8%, 2.6%, and 0.9% respectively. The availability and robustness of the proposed method are verified.

Key words image processing; dilated convolutional neural network; simultaneous localization and mapping; unsupervised learning; monocular vision; depth estimation
OCIS codes 100.2000; 110.4153; 150.5670

1 引 言

理解三维场景几何是视觉感知的基本主题。大部分物体是规则的并存在一定的相关性，获取场景的深度信息可以了解场景的三维结构，从而结构化地了解世界，继而完成许多经典的计算机视觉任务，

收稿日期：2019-07-04；**修回日期**：2019-08-03；**录用日期**：2019-08-28
基金项目：国家自然科学基金(61772328，61802253，61831018)、上海晨光人才计划(17CG59)、江西省经济犯罪调查和预防技术协作创新中心(JXJZXTCX-027)

* **E-mail**：zjfang@foxmail.com

如深度恢复、视觉测距和三维重建等。这些技术在工业上具有广泛的应用，可以推动自动驾驶[1]、三维重建技术[2]和交互式机器人[3]等人工智能技术的发展。近几年，卷积神经网络(CNN)在深度估计中的重要性不言而喻，文献[4-10]显示，采用 CNN 估计深度已经取得了很好的效果。然而监督方式的 CNN 学习通常需要大量的数据，在标注真实值(ground truth)时成本较高，很大程度上限制了数据量[5-7]。由于监督式学习的深度估计方法存在泛化能力低、成本昂贵等问题，许多研究者将目光转向无监督学习的深度估计方法，利用几何约束来优化深度图，致力于恢复场景的运动结构。然而，当前主流的无监督方法[4,8,11-13]训练的模型在精度上往往不能达到预期的效果，且深度图像分辨率低的问题也一直被广泛关注[14]。相比之下，传统的深度估计方法需要先进传感器如激光雷达，既昂贵又不灵活。此外，传感器生成的深度图分辨率低，边缘信息不准确。RGB 相机因其体积小、功耗低而被广泛应用于深度估计领域。综合近几年各种相关工作，本文采取一种无监督视图重建的方式获取单目图片的深度。受文献[15]对图像进行语义分割时保留完全尺度上的特征图的启发，本文对现有的深度估计网络(DispNet)进行改进，引入扩张卷积神经网络，扩大感受野的同时，保留更大尺度上的深度特征，提高深度图的视觉质量。同时，针对神经网络回归相机姿态准确率低等问题，采用传统的基于定向的 FAST 特征和旋转的 BRIEF 特征的 SLAM(Simultaneous Localization and Mapping)算法(ORB-SLAM 算法)取代神经网络对相机姿态进行全局优化，利用视图重建[4,8,11-13]的方式训练网络模型，通过最小化渲染重建视图与原始图片的差值，促使网络生成更准确的深度图。与现有的深度学习的深度估计方法相比，本文方法在视觉感知上可获取更高质量的深度图，同时获得了更高的准确率。

本文方法创新点如下：1)采用扩张卷积神经网络保留完全尺度上的特征图，考虑了更大范围的图片特征，保留了更多的细节位置信息，提高了深度图的视觉质量；2)利用传统 ORB-SLAM 算法的追踪线程计算并优化相机姿态，将优化后的相机姿态嵌入视图重构深度估计网络框架中，促使改进的 DispNet 生成更准确的深度图。

2 视图重构理论分析

视图重构的无监督学习法以目标帧的重建误差作为监控信号。给定目标帧的深度及其相邻帧的相机运动姿态，利用仿射变换进行渲染能够重建目标帧。其中，重建误差作为网络训练的约束项和评估深度准确率的度量方式。采用的视图重构框架中，对于每个训练样本，输入 $t-1$，t 和 $t+1$ 时刻的三张图片经预处理后的单张图片(参考文献[4,11])，三张图片分别表示为 F_{t-1}，F_t 和 F_{t+1}。其中，F_{t-1}，F_{t+1}分别为左侧和右侧参考视图，F_t 为目标视图，F'_t为重构得到的目标视图，将〈F_1，F_2，…，F_N〉视为框架的视频训练集。采用扩张卷积改进的 DispNet 提取图片的深度特征并采用 ORB-SLAM 算法分别优化左/右视图与目标视图之间的全局相机姿态。目标帧的深度和目标帧与相邻帧之间的相机姿态用于从相邻参考帧中重建目标帧。利用目标帧的深度和相邻帧的运动信息，基于几何推理可以重构目标帧，优化重构视图与目标视图之间的光学误差从而优化深度图。重构过程包含两个可微分运算，允许梯度传播训练神经网络。重构得到的两个目标视图为

$$F'_{t-1}=r(F_{t-1},\boldsymbol{K},\boldsymbol{T}_{t\to t-1},D_t), \tag{1}$$

$$F'_{t+1}=r(F_{t+1},\boldsymbol{K},\boldsymbol{T}_{t\to t+1},D_t), \tag{2}$$

式中：$r(\cdot)$为重构过程；$\boldsymbol{K}$ 为相机内参；$\boldsymbol{T}_{t\to t-1}$、$\boldsymbol{T}_{t\to t+1}$分别为左、右参考视图和目标视图之间的相对相机姿态；D_t 为目标视图的深度。与文献[4,13]相同的是，$r(\cdot)$使用极线几何变换和变形定义左视图和右视图之间的像素关联，并使用此相关信息合成目标视图。引入扩张卷积神经网络提取大尺度下的特征并直接映射第 t 帧的深度值 D_t。同时，采用传统的 ORB-SLAM 算法优化全局相机姿态 $\boldsymbol{T}_{t\to t-1}$ 和 $\boldsymbol{T}_{t\to t+1}$。利用线性几何原理进行视图重构，重构的主要过程表示为

$$r(\boldsymbol{K},\boldsymbol{T}_{t\to t-1},D_t,F_{t-1})=\boldsymbol{K}\boldsymbol{T}_{t\to t-1}D_t\boldsymbol{K}^{-1}F_{t-1}。 \tag{3}$$

3 无监督单目深度估计扩张卷积神经网络

3.1 网络架构

基于视图重构是无监督深度估计的典型方法。通过训练卷积编码器估计目标视图的深度，利用相邻帧之间的相机位姿重构目标视图，进而优化深度图，这一方法被广泛使用。本文提出了一种融合扩张卷积网络与视觉同时定位与建图(SLAM)的方法估计单目图片的深度，该方法通过扩张卷积增大特

征提取的感受野，考虑更多的细节特征，并采用传统的 SLAM 算法对相机姿态进行全局优化，利用线性几何原理重构得到目标视图。在整体框架中，具有时序信息的视频帧经过预处理后的单张图片作为输入图像，通过扩张卷积网络获取初始深度图，并结合传统 SLAM 算法优化后的相机位姿，得到重构视图。通过最小化重构视图与目标视图之间的差值优化深度图，从而提高深度图的视觉质量和准确率。详细的网络架构如图 1 所示，包含两部分：1）扩张卷积深度估计网络，主要由 3 个扩张卷积层、7 个标准卷积层和 7 个反卷积层组成；2）ORB-SLAM算法优化相机姿态模块，对优化后的相机姿态与扩张卷积深度图网络提取的深度进行几何变换以重构输入图片，从而促使网络生成更高视觉质量的深度图。图 1 中 $\boldsymbol{R}$ 和 $\boldsymbol{t}$ 分别为旋转矩阵和平移矩阵。

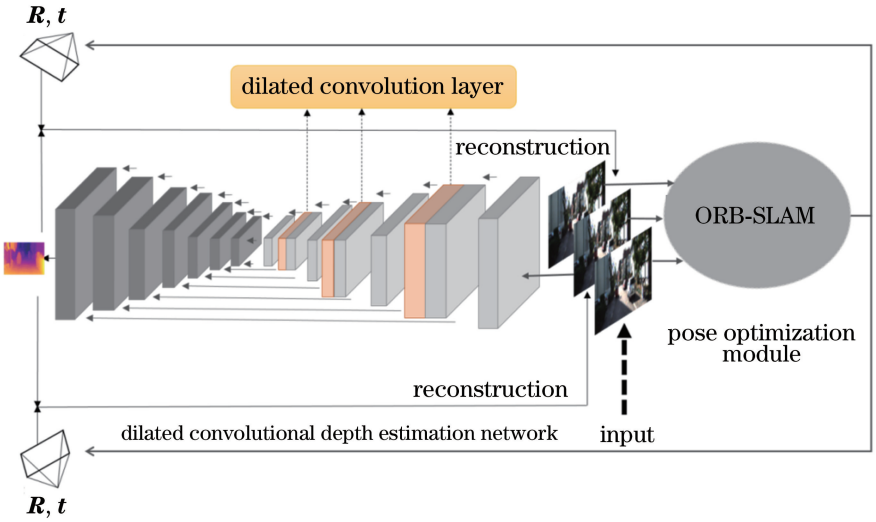

图 1 网络框架示意图

Fig. 1 Illustration of the network framework

3.2 扩张卷积深度图网络

采用扩张卷积神经网络提取图像的深度信息，基于主流的 DispNet[12] 进行改进。在视图重建损失的约束下，通过迭代训练得到单目图片深度估计模型。扩张卷积的引入源于特征提取的过程中，下采样操作常丢弃具有小尺寸的局部特征。而扩张卷积正是下采样层的替代方案[15]，不仅增加了感受野，保留更多的局部信息，还保持了特征图的空间维度，实现了局部与全局的双重优化。

扩张卷积在非零滤波器中插入零点对特征图进行采样，加快了感受野的动态速率，保持了特征图的空间维度而不增加计算复杂度[15]。与标准卷积的区别在于（假设卷积核大小为 3×3），标准大小的卷积核的感受野大小为 3×3；扩张率为 n 的扩张卷积的感受野扩大为 $(2n+1)\times(2n+1)$，在扩张卷积的过程中，卷积核大小保持不变。具体的卷积对比如图 2 所示。

扩张卷积与标准卷积的可视化操作对比如图 3 所示。借鉴文献[15]，扩张率的设置需要满足两个

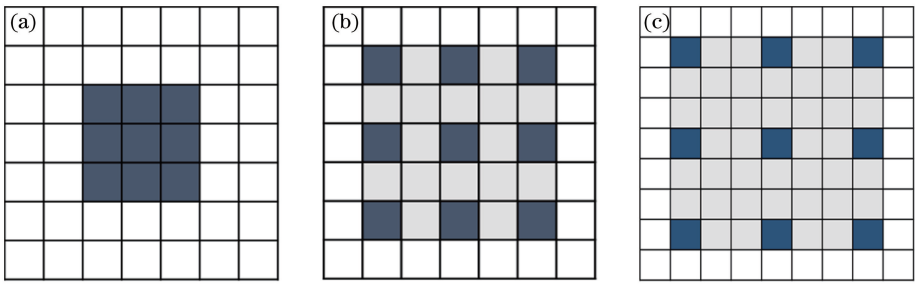

图 2 标准卷积与扩张卷积滤波器对比图。(a)标准卷积滤波器；(b)扩张率为 2 的扩张卷积滤波器；(c)扩张率为 3 的扩张卷积滤波器

Fig. 2 Comparison of standard convolution and dilated convolution filters. (a) Standard convolution filter; (b) dilated convolution filter with dilation ratio of 2; (c) dilated convolution filter with dilation ratio of 3

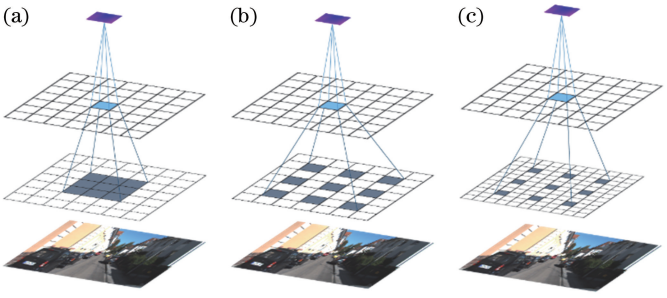

图 3 扩张卷积与标准卷积的可视化过程对比图。(a)标准卷积可视化过程;(b)扩张率为 2 的扩张卷积可视化过程;(c)扩张率为 3 的扩张卷积可视化过程

Fig. 3 Visualization process comparison of dilated convolution and standard convolution. (a) Visualization process of standard convolution; (b) visualization process of dilated convolution with dilation ratio of 2; (c) visualization process of dilated convolution with dilation ratio of 3

条件:1)使用尽可能少的扩张卷积层最大化感受野,从而节省内存使用量;2)完全覆盖感受野而不丢失任何输入信息。遵循上述条件,实验发现,提出的深度特征提取模型在特征提取的过程中,使用扩张率分别为 3,9,15 的卷积核对原始图像进行特征提取效果最优,最终由全连接层将不同空间位置的信息进行线性融合。然而,内存不足与扩张率设置难度[15]给扩张卷积带来了巨大的挑战,限制了扩张卷积的实际应用。

4 ORB-SLAM 算法优化相机姿态

4.1 ORB-SLAM 算法优化模块

采用 ORB-SLAM 算法全局优化相机姿态。选定初始帧后,使用随机抽样一致性(RANSAC)算法[16-17]进行图像的特征点匹配。只有满足前后帧匹配点对超过 100,才认为当前两帧可以进行初始化并利用两帧的匹配关系进行后续工作。继而开始计算两帧之间的变换矩阵 $\boldsymbol{T}$(由旋转矩阵 $\boldsymbol{R}$ 和平移矩阵 $\boldsymbol{t}$ 组成)。ORB-SLAM 算法详细流程如图 4 所示。

用于优化相机姿态的 ORB-SLAM 算法主要采用 Bundle Adjustment[17](BA)优化算法,即采用最小化重投影误差来减小漂移误差。重投影误差指真实三维空间点在图像平面上的投影和重投影的差值,最小化差值的和是整个过程优化的目标。将这个问题转化成一个最小二乘法的问题,可获取最优的相机位姿参数[18]及三维空间点的坐标。

此外,BA 算法是一个图优化模型,由节点和边组成。该图模型的节点由相机 C_t 和三维空间点 X^k 构成,若点 X^k 投影到相机 C_t 的图像上,则将这两个节点连接起来。BA 可以化为稀疏矩阵的形式,从而减小计算量。联合最小化所有相机和点的重投影误差,优化本质图得到最终的相机位姿。具体的投影约束过程如图 5 所示。

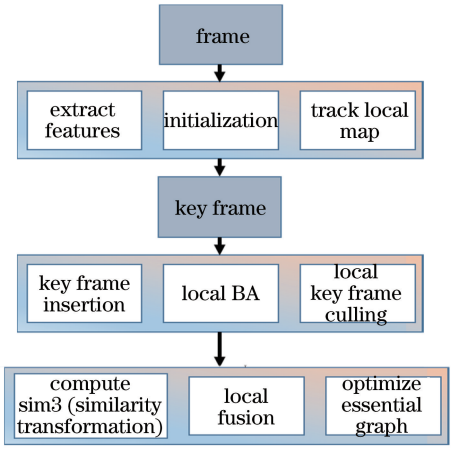

图 4 ORB-SLAM 算法优化全局相机姿态总体流程

Fig. 4 Flow chart of optimizing global camera pose by ORB-SLAM algorithm

姿势优化的目标函数表示为

$$\boldsymbol{T}_t=\operatorname{argmin}\sum_{C_t,X^k}\|f(X^k,C_t)-x_{k,t}\|^2=\begin{pmatrix}\boldsymbol{R}_{t,t-1} & \boldsymbol{t}_{t,t-1}\\ 0 & 1\end{pmatrix},\tag{4}$$

式中:$\boldsymbol{T}_t$ 为 t 时刻的变换矩阵;$x_{k,t}$ 为 t 时刻第 k 帧的三维坐标点的真实值;$f(\cdot)$ 代表重投影过程;$\boldsymbol{R}_{t,t-1}$ 与 $\boldsymbol{t}_{t,t-1}$ 分别表示 t 时刻与 $t-1$ 时刻的旋转矩阵和平移向量。

4.2 误差度量方式

4.2.1 视图重建损失

记 F_{tar} 为原始目标图像,其对应的重建图像为 F_{rec}。由上文可知,给定目标图像的深度图和目标视图与相邻视图之间的相机运动姿态,利用仿射变换可以重建目标视图。视图重建过程的损失

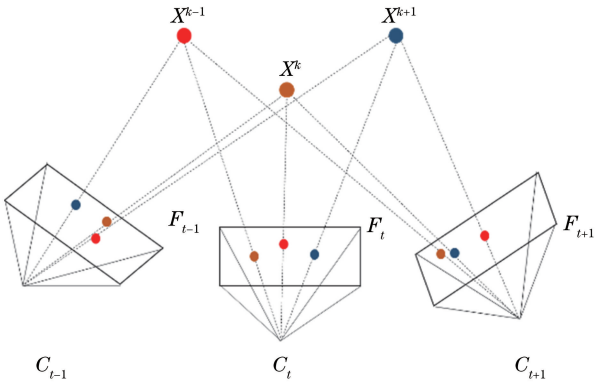

图 5 三维空间点在图像平面上的投影过程

Fig. 5 Projection process of three-dimensional space points onto the image plane

(L_{rec})为

$$L_{\mathrm{rec}}=\sum_{\langle F_1,F_2,\cdots,F_N\rangle}\sum_{p}\left|F_{\mathrm{rec}}(p)-F_{\mathrm{tar}}(p)\right|,\quad(5)$$

式中：p 为像素。

4.2.2 深度平滑损失

CNN 学习的深度特征具有空间细节、边缘和噪声，图像中的梯度差异很大。为了减小梯度差异并使深度图看起来更真实，视差梯度由图像域的边缘感知平滑项进行加权，通过边缘感知平滑项使得深度图添加鲁棒的空间平滑度惩罚。深度预测平滑损失(L_{smooth})为[9]

$$\begin{aligned}L_{\mathrm{smooth}}=\sum_{i,j}&\left|\theta_x D_{ij}\right|\exp\left(-\left|\theta_x F_{ij}\right|\right)+\\&\left|\theta_y D_{ij}\right|\exp\left(-\left|\theta_y F_{ij}\right|\right)/N,\end{aligned}\quad(6)$$

式中：N 为像素总数；θ_x 和 θ_y 分别为水平方向和垂直方向的梯度；F_{ij} 为像素 i，j 处的原始图像；D_{ij} 为其深度。

4.2.3 总体损失

结合 4.2.1 节和 4.2.2 节的两个损失函数，记 λ_{rec} 和 $\lambda_{\mathrm{smooth}}$ 为每个损失项的权重。总体损失函数(L_{final})为

$$L_{\mathrm{final}}=\lambda_{\mathrm{rec}}L_{\mathrm{rec}}+\lambda_{\mathrm{smooth}}L_{\mathrm{smooth}}。\quad(7)$$

5 实验结果与分析

将扩张卷积模型预测的深度图效果与现有方法进行比较，同时也利用标准的评估工具对模型性能进行评估与对比。

5.1 KITTI 数据集实验

5.1.1 实验环境与平台

实验环境包括一台配有 NVIDIA GTX 1070 显卡和 8 GB 内存的计算机硬件设备、Pycharm 专业版以及 Kdevelop4 IDE 软件开发工具。采用 TensorFlow 框架，在官方 KITTI Odometry 数据集上进行训练，并利用 KITTI raw 数据集进行评估。实验过程中，选用 KITTI Odometry 数据集中00-10的左视频序列作为训练集和验证集，数据集中的图片大小均设置为 128 × 416。训练阶段采用 Adam[19]优化器，通过不断进行实验与调参发现，超参数[β_1，β_2]=[0.9，0.999]效果最佳。借鉴文献[4]，将 λ_{rec} 和 $\lambda_{\mathrm{smooth}}$ 分别设为 1.0 和 0.5 来训练网络，以达到最优性能。训练初始学习率设置为 0.0001。重建损失变化、平滑损失变化及总体损失变化情况如图 6 所示。随着迭代次数的增加，总体损失逐渐减小，网络迭代 1.8×10^5 次时收敛。

5.1.2 相机姿态估计可视化结果比较

ORB-SLAM 算法采用 BA 算法、回环检测和重定位共同对相机姿态进行全局优化。为了更直观地理解 ORB-SLAM 算法计算相机姿态的结果，使用可视化工具显示 00-03、09 和 10 序列的相对姿态轨迹、ground truth 与当前主流方法的对比，如图 7 所示，其中 DVF-T 和 DVF-N 分别表示 Depth-VO-Feat Temporal 和 Depth-VO-Feat FullNYUv2，是文献[13]的预测结果。

从图 7 不难发现，ORB-SLAM 算法优化后的相机姿态轨迹几乎接近真实值，远优于神经网络学习的相机姿态，与当前主流方法[13]相比表现出了更高的准确率。将准确率更高的相机相对姿态嵌入视图重建框架中，可进一步促进扩张卷积网络生成更加准确的深度图。

5.1.3 深度图估计可视化结果比较

模型预测结果如图 8 所示，将本文方法与 Garg 等[11]和 sfmlearner[4]的结果进行可视化比较。由图 8 可以看到，本文方法在目标深度轮廓的保留上显示出更好的性能，与 Garg 等[11]和 sfmlearner[4]

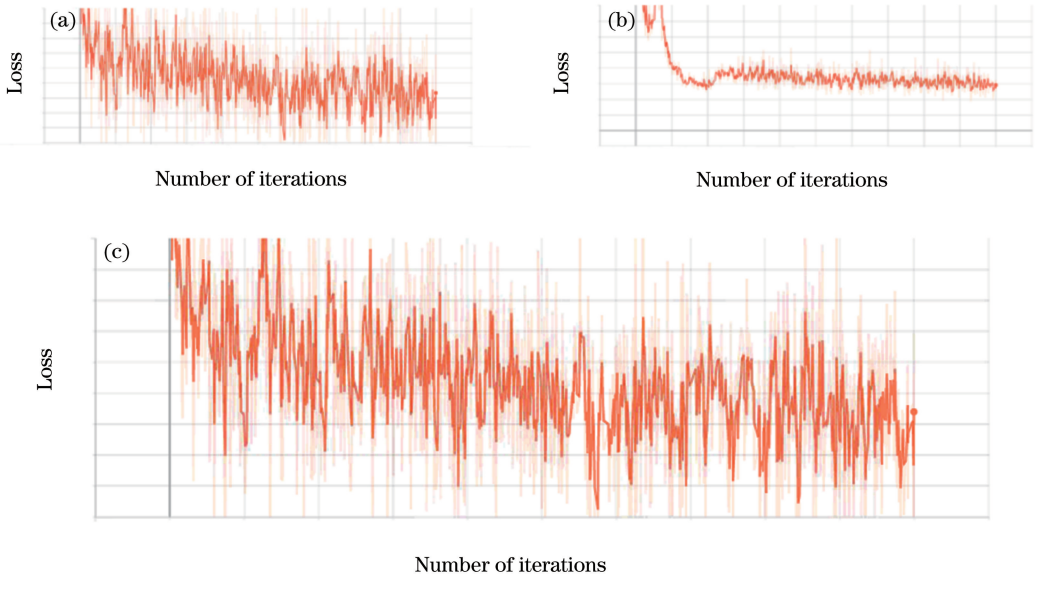

图 6 不同损失变化曲线。(a)重建损失;(b)平滑损失;(c)总体损失

Fig. 6 Curves for different losses. (a) Reconstruction loss; (b) smooth loss; (c) total loss

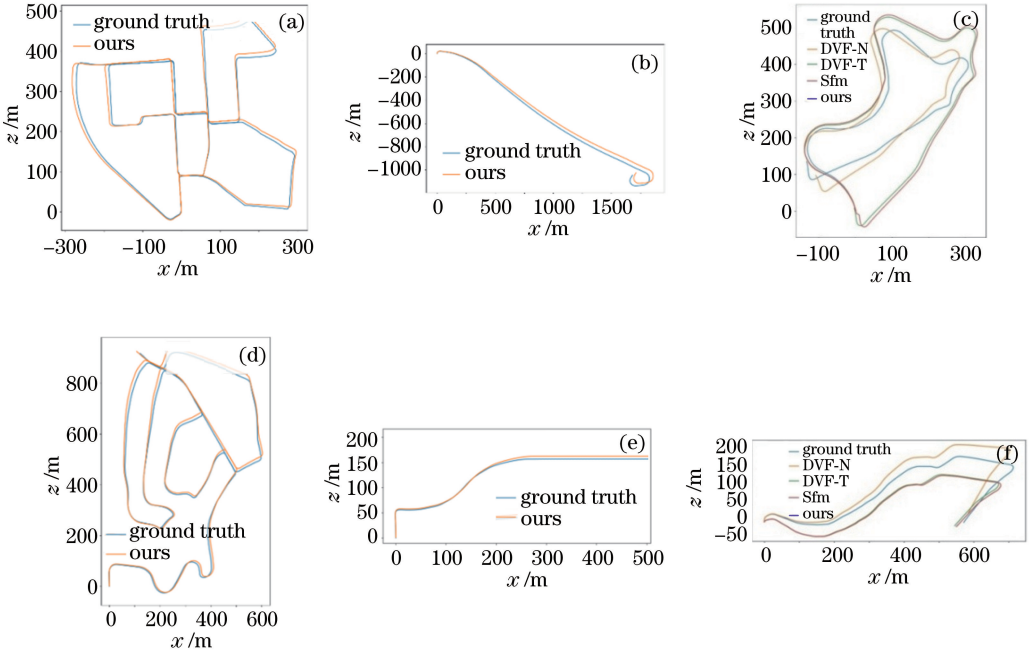

图 7 KITTI Odometry 数据集中不同序列的相对相机姿态轨迹。(a) 00;(b) 01;(c) 09;(d) 02;(e) 03;(f) 10

Fig. 7 Camera pose trajectories for different sequences in the KITTI Odometry dataset. (a) 00; (b) 01; (c) 09; (d) 02; (e) 03; (f) 10

方法相比,本文方法得到的深度图更加接近真实值,视觉上更加真实可靠。最直观的是,本文方法对深度图的视觉质量有很大程度的改进。该结果证明了深度网络使用的扩张卷积扩大感受野的有效性。此外,利用可视化工具对保留的深度细节特征进行视觉展示,如图 9 所示,黑色框标注为细节信息。

5.2 模型性能评估

5.2.1 相机姿态结果评估

采用均方根误差(RMSE)度量方式评估 ORB-SLAM 算法的准确性,并与现有方法进行比较,如表 1 所示。其中,t_{error} 为平移误差,r_{error} 为旋转误差。

由表 1 可以看到,与现有方法相比,本文采用的传统 ORB-SLAM 算法计算得到的相机相对位姿误

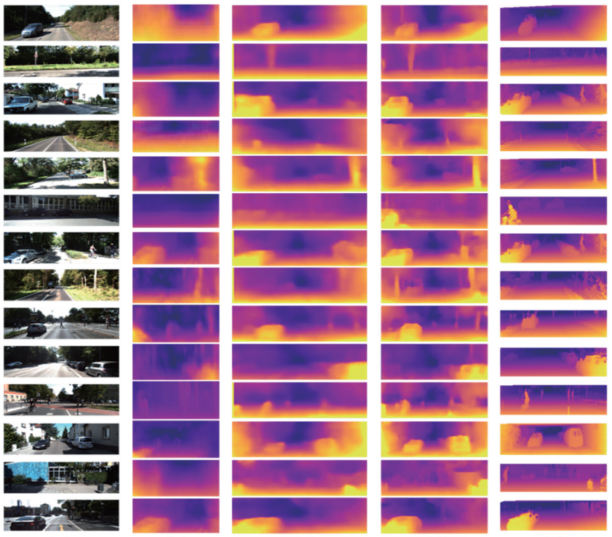

图 8 深度预测的定性比较。(a) RGB 输入图像；(b) Garg等[11]的方法；(c) sfmlearner 方法[4]；(d)本文方法；(e) ground truth

Fig. 8 Qualitative comparison of depth prediction. (a) RGB input image; (b) method of Garg *et al*.[11]; (c) sfmlearner method[4]; (d) our method; (e) ground truth

差最小，效果最佳。

5.2.2 深度估计模型性能评估

使用官方 TUM 评估工具包对 KITTI raw 数据集进行评估实验，与其他方法进行公平比较，具体的度量方式如下：

绝对差为

$$A=\sum_{t\in N}\frac{|D_t-D_t'|}{D_t'}\Big/N,\tag{8}$$

平方差为

$$S=\sum_{t\in N}\frac{\|D_t-D_t'\|^2}{D_t'}\Big/N,\tag{9}$$

RMSE 为

$$R=\sqrt{\sum_{t\in N}\|D_t-D_t'\|^2/N},\tag{10}$$

对数均方差为

$$\lg R=\sqrt{\sum_{t\in N}\|\lg D_t-\lg D_t'\|^2/N},\tag{11}$$

阈值为

$$\text{s.t. }\max(D_t/D_t',D_t'/D_t)=\delta<T_{\text{threshold}},\tag{12}$$

式中：N 为测试集中具有地面实况的像素数；D_t'和 D_t 分别为第 t 个像素的真实和预测深度；δ 取值为 $1.25,1.25^2,1.25^3$；$T_{\text{threshold}}$ 为阈值。利用(8)～(12)式，深度估计模型的 TUM 评估结果如表 2 所示，使用的是 KITTI 数据集。

表 1 KITTI Odometry 数据集 09 和 10 序列的 RMSE 比较

Table 1 RMSE comparison of 09 and 10 sequences in the KITTI Odometry dataset

Method	Sequence 09		Sequence 10	
	t_{error}/%	r_{error} per 100 m /(°)	t_{error}/%	r_{error} per 100 m /(°)
Luo *et al*.[20]	3.72	1.60	6.06	2.22
Zhou *et al*.[4]	18.77	3.21	14.33	3.30
Li *et al*.[21]	7.01	3.61	10.63	4.65
Zhan *et al*.[13] (Tem)	11.93	3.91	12.45	3.46
Zhan *et al*.[13] (New York University datasets)	11.92	3.60	12.62	3.43
Ours	1.70	0.50	1.43	0.52

Eigen 等[5-6]使用从粗略到细致的网络预测深度图，在真实深度值的监督下，采用了由局部到全局的思想实现单目图片深度估计。Liu 等[7]使用 VGG-16 的监督方法训练网络以初始化参数。以上几种方法均需利用标注好的真实深度值监督网络

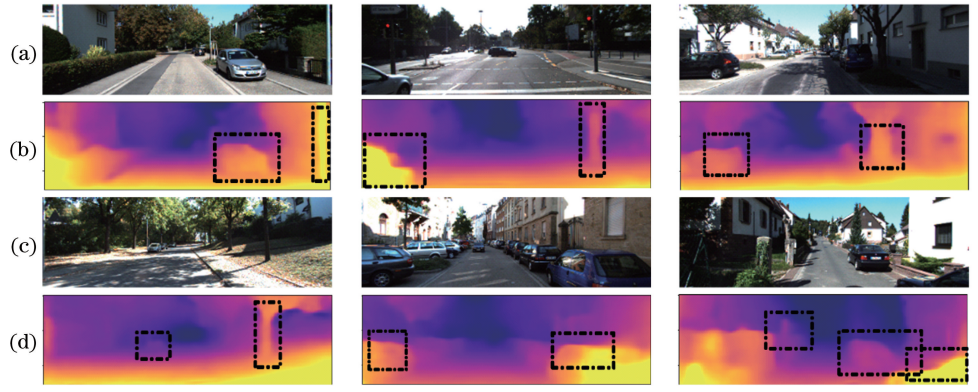

图 9 深度细节可视化比较。(a)(c)输入图像；(b)(d)输出图像

Fig. 9 Visualization comparison of depth details. (a)(c) Input images; (b)(d) output images

表 2 深度估计模型的 TUM 评估结果比较

Table 2 Comparison of TUM evaluation results for depth estimation model

Method	Supervised	Data	Error				Accuracy		
			A	S	R	lg R	δ_1 / %	δ_2 / %	δ_3 / %
Method in Ref. [5]	√	KITTI	0.214	1.605	6.563	0.292	67.3	88.4	95.7
Method in Ref. [6]	√	KITTI	0.203	1.548	6.307	0.282	70.2	89.0	95.8
Method in Ref. [7]	√	KITTI	0.202	1.614	6.523	0.275	67.8	89.5	96.5
Method in Ref. [22] (photo)	×	KITTI	0.211	1.980	6.154	0.264	73.2	89.8	95.9
Method in Ref. [22] (photo+ad)	×	KITTI	0.220	1.976	6.340	0.273	70.8	86.7	93.4
Method in Ref. [4]	×	KITTI	0.208	1.768	6.856	0.283	67.8	88.5	95.7
Method in Ref. [4] (without explainability masks)	×	KITTI	0.221	2.226	7.527	0.294	67.6	88.5	95.4
Ours	×	KITTI	0.189	1.592	6.432	0.268	71.4	91.1	96.3

的训练，大大限制了数据量。Kumar 等[22]采用一种生成对抗式网络估计单目图片的深度信息，取得了不错的效果。Zhou 等[4]采用视图合成的思想进行无监督学习的单目图片深度估计，对不同的数据集进行了实验，并将网络的性能分为有、无可解释性掩模分别进行了定量比较。本文基于文献[4]的方法进行改进，由表 2 中的定量比较数据可以看到，与文献[4]无可解释性掩模的方法相比，本文方法的误差度量指标 A、S、R 和 lg R 分别降低了 0.032、0.634、1.095和 0.026；准确率度量指标 δ_1、δ_2 和 δ_3 分别提升了 3.8%、2.6%和 0.9%。从深度图的视觉对比结果可以看到，本文方法得益于扩张卷积增大感受野的作用，在深度图的细节保留上取得了更好的效果。此外，与现有的深度估计方法相比，在无需真实深度值监督的情况下，提出的无监督学习的单目深度估计效果优于其他无监督学习方法，甚至优于有监督的深度估计方法。

6 结 论

为处理无监督单目图片深度估计的任务，提出了一种新颖的无监督单目图片深度估计方法，利用光学误差的约束进行单目图片的深度估计。网络训练受扩张卷积网络和传统姿态估计的双重约束。其中，深度图网络使用扩张卷积扩大了感受野，考虑了更大尺度的信息，促使特征更加完整。用于视图重建的相机姿态采用传统的算法进行全局优化，即 ORB-SLAM 算法，结果几乎接近真实值，优于大部分现有的姿态估计方法。该网络架构不需要真实的深度值，质量和准确率都得以提高，能够将网络回归的深度图直接应用于三维重建、智能机器人等领域。实验结果表明，扩张卷积网络考虑到了更多的细节特征，传统方法的姿态估计几乎接近实际数值，准确率达到了较高的水平。模型最终估计的深度图不仅在视觉质量上取得了更好的效果，在准确率方面也得到了较大提升。本文方法无需深度真实值的监督，优于大部分无监督方式的深度估计方法，与有监督学习的方法相比性能也有较大的提升。

参 考 文 献

[1] Pan X L, You Y R, Wang Z Y, et al. Virtual to real reinforcement learning for autonomous driving[C] // Proceedings of the British Machine Vision Conference 2017, September 2017, London, UK. UK: BMVA Press, 2017: 11.

[2] Zeng Z P, Zhang J L, Wei Z S, et al. Three-dimensional reconstruction method based on smartphone imaging[J]. Laser & Optoelectronics Progress, 2018, 55(11): 111502.
曾昭鹏, 张江乐, 魏志尚, 等. 一种基于智能手机成像的三维重建方法[J]. 激光与光电子学进展, 2018, 55(11): 111502.

[3] Michalos G, Karagiannis P, Makris S, et al. Augmented reality (AR) applications for supporting human-robot interactive cooperation[J]. Procedia CIRP, 2016, 41: 370-375.

[4] 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, 2017: 6612-6621.

[5] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C] // Advances in Neural Information Processing Systems, December 8-13, 2014, Montreal, Quebec, Canada. Canada: NIPS, 2014: 2366-2374.

[6] 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, 2015: 2650-2658.

[7] 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 (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 5162-5170.

[8] Wang A J, Fang Z J, Gao Y B, et al. Depth estimation of video sequences with perceptual losses [J]. IEEE Access, 2018, 6: 30536-30546.

[9] Tosi F, Aleotti F, Poggi M, et al. Learning monocular depth estimation infusing traditional stereo knowledge[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 16-20, 2019, Long Beach, CA. New York: IEEE, 2019: 9799-9809.

[10] Li Y, Chen X W, Wang Y, et al. Progress in deep learning based monocular image depth estimation[J]. Laser & Optoelectronics Progress, 2019, 56(19): 190001.
李阳, 陈秀万, 王媛, 等. 基于深度学习的单目图像深度估计的研究进展[J]. 激光与光电子学进展, 2019, 56(19): 190001.

[11] Garg R, Vijay K B G, Carneiro G, et al. Unsupervised CNN for single view depth estimation: geometry to the rescue[M] // Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9912: 740-756.

[12] Mayer N, Ilg E, Hausser P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 4040-4048.

[13] Zhan H Y, Garg R, Weerasekera C S, et al. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction[C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 340-349.

[14] Li S M, Lei G Q, Fan R. Depth map super-resolution reconstruction based on convolutional neural networks[J]. Acta Optica Sinica, 2017, 37(12): 1210002.
李素梅, 雷国庆, 范如. 基于卷积神经网络的深度图超分辨率重建[J]. 光学学报, 2017, 37(12): 1210002.

[15] Zhou X Y, Zheng J Q, Yang G Z. Atrous convolutional neural network (ACNN) for biomedical semantic segmentation with dimensionally lossless feature maps [J/OL]. (2019-01-26) [2019-07-03]. https://arxiv.gg363.site/abs/1901.09203.

[16] Chen S, Wang L. An improved SIFT feature matching based on RANSAC algorithm [J]. Information Technology, 2016, 40(12): 39-43.

[17] Shi G J, Xu X Y, Dai Y P. SIFT feature point matching based on improved RANSAC algorithm[C] // 2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics, August 26-27, 2013, Hangzhou, China. New York: IEEE, 2013: 474-477.

[18] Chen J W, Quan S B, Quan Y M, et al. Calibration method of relative position and pose between dual two-dimensional laser radar [J]. Chinese Journal of Lasers, 2017, 44(10): 1004005.
陈健武, 全思博, 全燕鸣, 等. 双二维激光雷达相对位姿的标定方法[J]. 中国激光, 2017, 44(10): 1004005.

[19] Kingma D P, Ba J. Adam: a method for stochastic optimization [J/OL]. (2017-01-30) [2019-07-03]. https://arxiv.org/abs/1412.6980.

[20] Luo C X, Yang Z H, Wang P, et al. Every pixel counts ++: joint learning of geometry and motion with 3D holistic understanding [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019: 1.

[21] Li R H, Wang S, Long Z Q, et al. UnDeepVO: monocular visual odometry through unsupervised deep learning [C] // 2018 IEEE International Conference on Robotics and Automation (ICRA), May 21-25, 2018, Brisbane, QLD, Australia. New York: IEEE, 2018: 7286-7291.

[22] Kumar A R S, Bhandarkar S M, Prasad M. Monocular depth prediction using generative adversarial networks [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 18-22, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 413-421.