文章编号:1001-5078(2024)09-01373-07

·激光应用技术·

图像与稀疏激光点融合的单目深度估计

蔡文靖,刘　鑫,王礼贺,纪宇航
(华北光电技术研究所,北京 100015)

摘　要:近年来,随着深度学习的快速发展,涌现出大量单目深度估计算法。但由于缺乏视差等几何约束,限制了算法深度预测精度的进一步提升,无法满足实际应用的需求。因此本文提出了一个二维图像与稀疏激光点融合的深度估计网络,通过实时输入少量激光点的高精度测距结果,提高深度预测精度;其次,为解决自采集数据激光雷达点分布不均匀问题,在有监督网络基础上,加入相对位姿估计网络与深度估计网络联合训练,同时增加光度一致性、深度重投影两个损失函数;最终,利用自采集数据进行实验分析,实验结果表明,当使用 160 个激光点时,即可将深度预测绝对相对误差由 10.1 % 降至 7.6 %,当使用 1280 个激光点时,深度预测绝对相对误差变化趋于平稳,降至 4.1 %。

关键词:单目深度估计;稀疏激光点;残差神经网络
中图分类号:TN958.98;TP73　　**文献标识码:**A　　**DOI:**10.3969/j.issn.1001-5078.2024.09.006

Monocular depth estimation based on image and sparse laser point fusion

CAI Wen-jing, LIU Xin, WANG Li-he, JI Yu-hang
(North China Research Institute of Electro-Optics, Beijing 100015, China)

Abstract: In recent years, with the rapid development of deep learning, a large number of monocular depth estimation algorithms have emerged. However, the lack of geometric constraints such as disparity, the depth prediction accuracy limits the further improvement of the depth prediction accuracy of the algorithm and fails to meet the needs of practical applications, soa depth estimation network that integrates images with sparse laser points is proposed in this paper. Firstly, the depth prediction accuracy is improved by inputting the high-precision ranging results of a small number of laser points in real time. Secondly, in order to solve the problem of uneven distribution of LiDAR points from self-collected data, on the basis of the supervised network, the relative position estimation network is added to be trained jointly with the depth estimation network. And two loss functions of luminance consistency and depth reprojection are added at the same time. Finally, the self-collected data are utilized to conduct the experimental analysis, and the experimental results show that when 160 laser points areused, the absolute relative error of depth prediction can be reduced from 10.1 % to 7.6 %, and when 1280 laser points are used, the change of the absolute relative error of depth prediction tends to stabilize to 4.1 %.

Keywords: monocular depth estimation; sparse laser points; residual neural network

1　引　言

目前视频图像在自动驾驶、安防监控、教育医疗等场景都有广泛应用,但相机成像基于小孔成像原理,成像过程可以看做将三维空间的场景投影到二维平面上,在投影过程中丢失了三维场景的深度信息,如果可以获得图像对应的深度数据,将大大提升图像在环境感知、增强现实、三维重建等领域的能力。

收稿日期:2023-12-20

基于图像的深度估计主要分为单目、双目、多目深度估计方法，双目、多目对相机内参、基线等要求较高，相较之下单目深度估计对硬件的需求简单，已有大量相关研究。早期基于图像深度线索的深度估计，主要包括消失点[1]、聚焦[2]、阴影[3]、运动恢复结构(SFM)[4]等，但其存在一些问题无法广泛应用，如精度较低，对场景要求较高，依赖特征提取与匹配，在场景缺乏细节信息时效果不好，计算量大无法实时运行等等。近年来，随着深度学习的快速发展，大量学者开始研究基于深度学习的单目深度估计算法[5-7]。

Eigen 等人[8]首次将深度卷积神经网络引入单目深度估计中，提出通过多尺度 CNN 网络预测图像的深度，该方法通过一个全局粗网络，根据全局信息对整幅图像的深度进行初步预测，之后通过一个局部精网络，根据局部信息对初始深度进行优化，但该方法的边界比较模糊，且只适用于室内场景。针对室外场景的深度估计，Kuznietsov 等人[9]提出有监督学习与无监督学习结合的方法，由于，室外场景很难获取与图像匹配的稠密深度真值，因此在有稀疏激光点的区域进行有监督学习，在没有真值的区域利用双目内参、基线的参数进行无监督学习，有监督与无监督结合可解决局部最优解问题，同时提高有监督学习的速度。Li 等人[10]提出学习深度与深度梯度的双流框架，通过深度与梯度的融合，提升模型的泛化能力，得到更丰富的细节信息。该方法包含两个分支，均采用 VGG-16 网络结构，同时也可用 VGG-19、ResNet 等卷积神经网络替换。

基于单目图像的深度估计问题，本身是一个不适定问题[11]，由于缺乏视差等几何约束，现有算法的深度预测精度有限，在 KITTI 数据集[12]中的绝对相对误差一般在 10 % 左右，这限制了算法在实际中的应用。为进一步提高单目深度估计算法的精度，使其可以在实际中广泛应用，本文提出通过引入少量激光点提升算法精度。首先，提出一个二维图像与稀疏激光点融合的深度估计网络，将图像与激光点分别通过残差神经网络(ResNet)进行编码，特征图融合后进行解码得到深度图；其次，为解决自采集数据激光雷达分布不均匀问题，增加相对位姿估计网络，以及光度一致性、深度重投影误差两个损失函数；最后，通过融合不同数量的稀疏激光点对比实验，为激光与相机结合的硬件设计提供思路。在自采集数据中的实验结果表明，当引入 160 个激光点时，即可将深度预测绝对相对误差从无激光点的 10.1 % 降低至 7.6 %，当引入 1280 个激光点时，深度预测绝对相对误差的降低趋于平稳达到 4.1 %。

2 网络模型

2.1 网络结构

本文采用双分支编码器与混合解码器的结构，采用 ResNet[13]作为图像与稀疏激光点的编码器，分别提取图像与激光点的特征图并融合，之后使用多尺度跳跃连接结构的网络解码器对融合特征解码，得到深度图。整体网络结构如图 1 所示。

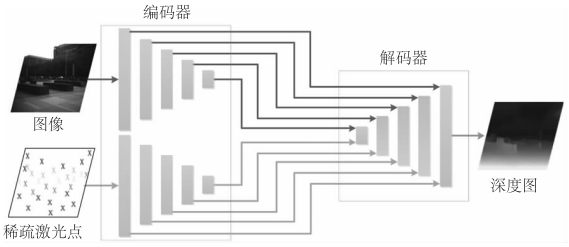


图 1 网络结构图

Fig. 1 The structure diagram of neural network

2.1.1 残差神经网络

本文采用残差神经网络 ResNet18 作为两个分支的编码器，ResNet18 包括 17 个卷基层和 1 个全连接层。网络结构主要包括 4 个残差块，每个残差块由两个 building block 组成，每个 building block 包括 2 个 3×3 卷积、2 个批量归一化(Batch Normalization，BN)、以及 2 个激活函数。其中，残差块 1～4 中的卷积厚度分别为 64，128，256，512，激活函数统一采用 ReLU 激活函数。具体结构如图 2 所示。

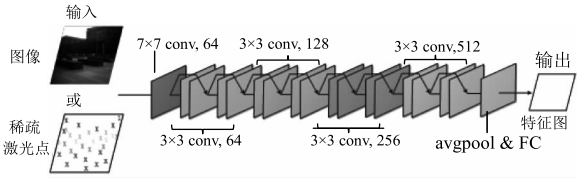


图 2 ResNet18 结构图

Fig. 2 The structure diagram of ResNet18

解码器由 5 个反卷积模块组成，每个模块分别由 3×3 卷积 – 激活函数 ReLU – 3×3 卷积 – 激活函数 ReLU – 上采样构成。其中每个模块将特征图进行 2 倍的上采样，逐步将特征图恢复至原始图像的分辨率。此外，为利用编码器中的底层图片特征，

编解码之间采用跳跃连接的方式，每个反卷积模块的第一个卷积－激活函数输出，与编码器对应分辨率的特征图进行 concatenate 连接，之后一起输入至第二个卷积中。第 2 ~ 5 个反卷积模块的输出特征图通过一个带 sigmoid 激活函数的 3 × 3 卷积层，可得到 4 个不同分辨率的深度图。训练时，对 4 个深度图均进行监督；系统实际运行时，直接使用最后输出的最高分辨率的深度图。

2.1.2　双分支编码

为了进一步提高深度预测的精度，本文融合少量高精度激光测距数据，将激光雷达的激光点投影到图像二维平面上，随机选取 N 个点作为网络输入的稀疏激光点，$N>0$，在第 3 节中将介绍 N 取不同值时对深度预测的优化效果。

然而激光点投影到二维图像上非常稀疏，且激光点是随机选取的，所以图像中的位置不固定，如果将投影的稀疏深度图作为图像的新通道，则稀疏的深度数据会被网络忽略。为了更多地保留和利用稀疏深度值的信息，本文采用双分支编码器与混合解码器的结构，具体结构如图 1 所示。构建 2 个 ResNet18 编码器，分别提取图像和稀疏激光点的特征图，融合 2 个特征图后进行解码得到最终的深度预测。通过双分支和特征后融合的设计，充分利用稀疏激光点的深度信息，同时提升模型抗噪能力。

由此训练的网络具有一定的深度估计能力，在仅二维图像输入时绝对相对误差为 9.3 %。但本论文数据为自采集数据，真值通过 80 线激光雷达采集，现有激光雷达为适配自动驾驶，中间部分线数较密，上下两侧的线数非常稀疏，因此点云投影到图像上后，上下 2 个部分的激光点非常稀疏，呈条纹状分布，如图 3 所示。

(a) 原图像　　(b)激光雷达数据投影到图像平面后与原图像叠加

图 3　激光雷达数据

Fig. 3 Lidar data

使用分布不均匀的数据训练后，得到的深度图中也具有明显的条纹，如图 4 所示。从图中可以看出，图像上下 2 个部分中 2 个激光条纹之间的误差较大，虽然这些区域没有激光真值，不会统计进误差之中，但极大的影响了深度图的视觉效果和可用性。为了去除条纹问题，本文引入相对位姿估计网络、光度一致性损失函数和深度重投影损失函数，通过帧间信息给网络添加更多的约束和指导。

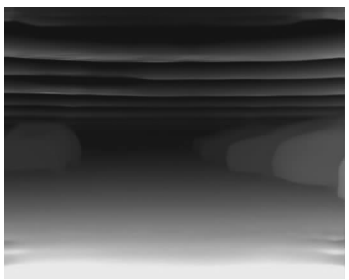


图 4　深度图上下部分具有明显条纹

Fig. 4 The upper and lower parts of the depth map have stripes

2.2　帧间信息约束

本文首先采用常规的 L_1 损失函数度量深度预测值与真值的误差，其中真值采用 80 线激光雷达采集的数据：

$$L_1 = \sum |D_{\text{lidar}} - D_{\text{predicted}}| \tag{1}$$

其中，D_{lidar} 和 $D_{\text{predicted}}$ 分别为激光雷达采集的深度真值与深度预测值。

由于第 2.1 节中提到的条纹问题，引入无监督深度估计[11]中的相对位姿估计网络和重投影误差，训练时，根据相邻帧之间的重投影误差，实现神经网络的反向传播。

2.2.1　相对位姿估计网络

相对位姿估计网络的输入为 2 帧相邻图像，输出为这 2 帧图像对应的相机位姿之间的变换矩阵 T。T 是一个 4 × 4 的矩阵，包含旋转分量 R 和平移分量 t。通过变换矩阵 T，即可将 2 帧图像对应的三维点进行转换。本文继续采用 ResNet18 网络预测相对位姿，该网络与深度估计网络耦合在一起联合训练，具体结构如图 5 所示。

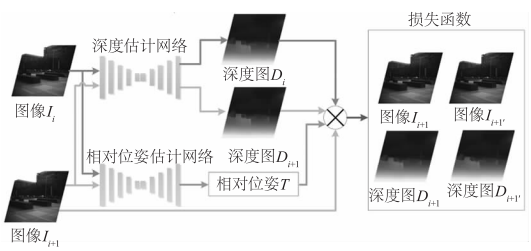


图 5　相对位姿估计网络与深度估计网络联合训练

Fig. 5 The joint training of pose estimation network and depth estimation network

根据预测的图像 I_i 与相邻帧图像 I_{i+1} 之间的相对位姿,以及预测的图像 I_i 的深度 D_i ,将 I_i 投影到 I_{i+1} 位姿的相机平面,得到合成的相邻帧图像 I_{i+1}' ,最小化 I_{i+1}' 与 I_{i+1} 之间的光度误差;同时,根据预测的图像 I_i 与相邻帧图像 I_{i+1} 之间的相对位姿,以及预测的图像 I_i 的深度 D_i ,合成相邻帧图像 I_{i+1} 的深度 D_{i+1}' ,最小化 D_{i+1}' 与预测出的图像 I_{i+1} 的深度 D_{i+1} 的深度图重投影误差。通过光度一致性损失和深度图重投影损失,将相邻帧间的几何约束作为监督信号,实现深度和相对位姿估计网络的无监督学习。

2.2.2 光度一致性损失函数

光度一致性假设:对于一个静止的场景,场景中同一个三维点 P 投影在两帧图像上,形成像素点 p_1 和 p_2 ,如图 6 所示, p_1 和 p_2 对应的像素值是一致的。

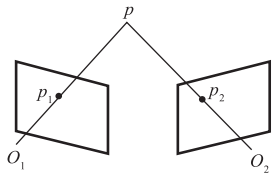


图 6 光度一致性

Fig. 6 Photometric consistency

对两帧相邻图像 I_i 和 I_{i+1} ,深度估计网络预测得到 I_i 的深度图 D_1 ,则可以将图像像素点逆投影到三维空间中,得到 I_i 中所有像素点在 I_i 相机坐标系下的三维坐标:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = D_1 \cdot K^{-1} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} \tag{2}$$

其中, $\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ 为三维空间坐标, K 为相机内参。 $\begin{pmatrix} u \\ v \\ 1 \end{pmatrix}$ 为像素点的齐次二维图像坐标。

根据相对位姿估计网络预测得到的两帧相对位姿 T, 可以得到三维点在 I_{i+1} 相机坐标系下的三维坐标:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = R \begin{pmatrix} x \\ y \\ z \end{pmatrix} + t = T \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{3}$$

其中, $T = [R \mid t]$,R 和 t 分别为旋转、平移矩阵。

将 I_{i+1} 坐标系下的三维点投影到图像坐标系下,可得到该点的像素坐标:

$$\begin{pmatrix} u' \\ v' \end{pmatrix} = \pi\left(K \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix}\right) \tag{4}$$

其中, π 代表齐次化操作,即 $\pi \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \frac{x}{z} \\ \frac{y}{z} \end{pmatrix}$。

通过上述计算可以得到图像 I_i 中像素点 (u,v) 在图像 I_{i+1} 中对应的像素点 (u',v') 。根据光度一致性假设,如果相对位姿估计网络预测的相机变换矩阵 T,和深度估计网络预测的深度图 D_i 完全准确,则在光度一致性假设下, I_i 中 (u,v) 和 I_{i+1} 中 (u',v') 的像素值是相等的。由此,在不考虑遮挡的情况下, I_i 中所有像素点都可以找到其在 I_{i+1} 中对应像素点的位置,则可以用 I_i 的像素值构建 I_{i+1} 得到 I_{i+1}' ,当 T 和 D_i 预测准确时, I_{i+1} 与重建的 I'_{i+1} 是一致的,则可以得到光度一致性误差。

$$L_{\text{photo}} = \left| I_{i+1} - I_{i+1}' \right| \tag{5}$$

通过计算重建图像与原始图像的差距,体现出像素对应关系的不准确性,进一步体现出深度预测值 D 与相机变换矩阵 T 的预测误差。因此,光度一致性损失函数对深度估计网络和相对位姿估计网络都有影响,是一个联合训练的过程。

光度一致性误差对每一个像素的深度预测都施加了损失约束,因此相比与稀疏激光雷达的 L1 损失监督,不会产生条纹状效应。结合这两个损失函数,可以在保持整体预测精度的同时,初步减轻条纹效果。但由于图像中存在部分相似区域或者低纹理区域,这些区域的像素值基本一致,即使预测的深度或位姿变换矩阵的误差较大,光度一致性误差也难以体现,因此,进一步引入深度图重投影误差。

2.2.3 深度图重投影损失函数

光度一致性损失误差是在光度一致性假设的基础上,度量重建图像的误差,而深度图重投影误差,是直接度量两帧图像的预测深度值之间的一致性误差。

由 2.2.1 节中的推导可知,对于一个静止场景,通过 I_i 预测的深度图 D_i ,可以得到三维空间点,根据预测的相机变换矩阵 T,可以得到 I_{i+1} 相机坐标系下的三维坐标点,将三维点投影到 I_{i+1} 图像坐标系

下,可以得到 I_{i+1} 坐标系下的像素坐标,同时,可以得到 I_{i+1} 的投影深度图 D'_{i+1} :

$$D'_{i+1} = Z(KTD_1 \cdot K^{-1}\begin{pmatrix} u \\ v \\ 1 \end{pmatrix}) \tag{6}$$

其中, $Z(\cdot)$ 表示取三维坐标的 Z 分量。

通过比较重投影重建的深度图 D'_{i+1} 和深度估计网络预测的 I_{i+1} 深度图 D_{i+1} ,得到深度图重投影误差:

$$L_{\text{reproj}} = |D_{i+1} - D_{i+1}'| \tag{7}$$

该误差同样与深度估计网络、相对位姿估计网络都有关,是一个联合训练过程,但其与图片像素值无关,因此对于低纹理区域也可以实现有效的监督。

综上所述,本文的损失函数由 L_1 损失、光度一致性损失、深度图重投影损失这三部分构成:

$$L = w_1 \cdot L_1 + w_{\text{photo}} \cdot L_{\text{photo}} + w_{\text{reproj}} \cdot L_{\text{reproj}} \tag{8}$$

其中, w 为超参数权重,分别设为 0.1,1.0,0.05。

通过增加后两个损失函数,深度估计网络预测的深度图中,去除了条纹现象,同时保持着监督训练的预测精度。

3　实验及结果分析

3.1　数据集

本文采取公开数据集预训练与真实采集数据集调优的训练策略,以提高网络在实际应用时的精度,即在 KITTI 数据集上训练 10 个 epoch,再在自采集数据集上进行 20 个 epoch 的调优训练。数据采集平台如图 7 所示,相机采用自研可见光、红外共孔径相机,可见光分辨率为 1280 × 1024,红外图像分辨率为 640 × 512;激光雷达采用速腾 RS-Ruby Lite,共采集 4 h 数据,36 万张可见光图像 - 红外图像对。激光雷达与相机的联合标定,通过 matlabLiDAR + Camera 校准工具箱完成。

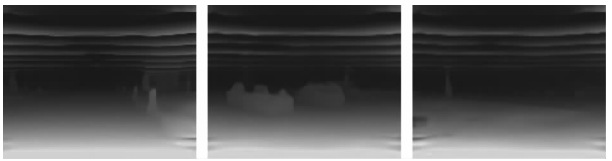
图 7　数据采集平台

Fig. 7 The data collection platform

3.2　实验结果及分析

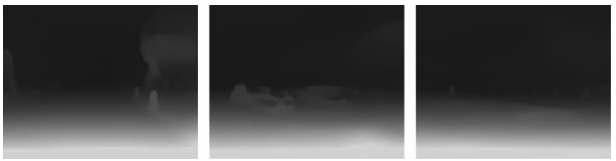
本节对比了上文介绍的单纯有监督训练、加入相对位姿估计网络和光度一致性损失函数训练、增加深度图重投影损失函数训练,以及融合不同数量稀疏激光点的深度估计结果,具体对比效果如图 8 所示。

(a)原始图像



(b)有监督训练效果图(不融合稀疏激光点)

(c)加入相对位姿估计、光度一致性损失函数训练效果图(不融合稀疏激光点)



(d)增加深度图重投影损失函数训练效果图(不融合稀疏激光点)

(e)融合160个稀疏激光点效果图

(f)融合900个稀疏激光点效果图

(g)融合1440个稀疏激光点效果图

图 8　模型优化、以及融合不同数量稀疏激光点后网络预测的深度图效果对比

Fig. 8 Comparison of depth map after model optimization and fusing different numbers of sparse laser points

由图 8(b)、(c)、(d)可以看出,通过深度估计

网络与相对位姿估计网络的联合训练,以及加入光度一致性损失函数和深度图重投影损失函数,可有效去除因自采集数据激光分布不均引发的条纹问题。但如图 8(d)所示,当图像部分区域缺乏纹理时,会出现物体包含空洞、物体边缘模糊、细长条物体无法显示等问题。当融合 160 个稀疏激光点后,如图 8(e)所示,第 2 列石墩的轮廓已清晰可辨,且空洞被填充,第 3 列的灯柱也可以较好的显示。当融合 1440 个稀疏激光点时,如图 8(g)所示,石墩轮廓已非常清晰,灯柱的形状更贴近现实。

3.3 融合不同数量稀疏激光点性能分析

本文选用绝对相对误差,均方根误差,以及绝对相对误差的像素比率作为网络的评价指标:

绝对相对误差(AbsRel)

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\|d_{\text{pred}}-d_{\text{gt}}\|}{d_{gt}} \tag{9}$$

均方根误差(RMSE)

$$\sqrt{(d_{\text{pred}}-d_{\text{gt}})^2} \tag{10}$$

绝对相对误差在 1.25^k 以内的像素比率,其中 $k\in\{1,2,3\}$

$$\left[\max\left(\frac{d_{\text{gt}}}{d_{\text{pred}}},\frac{d_{\text{pred}}}{d_{\text{gt}}}\right)=\delta_k<1.25^k\right] \tag{11}$$

其中,d_{pred} 为估计出的深度;d_{gt} 为激光雷达扫出来的深度真值。表 1 给出了融合不同数量稀疏激光点时,本文算法针对上述评价指标的对比结果。从表中可以看出,当融合 160 个激光点时,绝对相对误差从 10.1 % 降低至 7.6 %,已可显著提高深度估计的精度;随着融合激光点个数的增加,错误率指标与准确率指标均逐渐优化;当融合的激光点数量增长至 1280 时,各指标趋于平稳,绝对相对误差可达到 4.1 %。

表 1 融合不同数量稀疏激光点性能指标

Tab. 1 The performance for fusing different numbers of sparse laser points

稀疏激光点个数	错误率指标/% ↓		准确率指标/% ↑		
	AbsRel	RMSE	δ_1	δ_2	δ_3
0	10.1	5.160	86.5	94.5	97.6
160	7.6	5.103	91.7	96.6	98.3
480	5.6	4.597	94.0	97.4	98.7
800	4.9	4.412	94.8	97.7	98.8
1120	4.4	4.081	95.6	98.1	99.1
1280	4.1	3.961	95.9	98.2	99.1
1440	4.1	3.928	95.9	98.2	99.1

4 总 结

针对单目深度估计由于缺少视差等几何约束,限制了深度预测精度提升的问题,本文提出二维图像与稀疏激光点融合的深度估计网络。同时,为解决自采集数据激光雷达分布不均匀问题,加入相对位姿估计网络与深度估计网络联合训练,并增加光度一致性损失函数和深度重投影损失函数;最后,通过 KITTI 数据预训练、自采集数据调优的策略训练模型,从融合不同数量稀疏激光点的对比实验中可以看出,融合 160 个稀疏激光点后,即可显著提升模型精度,融合 1280 稀疏个激光点时,精度提升趋于平稳,这为后续的硬件设计提供一定思路。目前,稀疏激光点是从激光雷达捕获的数据中随机选取的,未来工作将对稀疏激光点在图像中的位置选取进行分析,并进一步压缩使用的稀疏激光点数量,使其可在现实应用场景中使用。

参考文献:

[1] Tsai Y M, Chang Y L, Chen L G. Block-based vanishing line and vanishing point detection for 3D scene reconstruction[C]//Intelligent Signal Processing and Communications, 2006: 586 – 589.

[2] Subbarao M, Surya G. Depth from defocus: a spatial domain approach[J]. International Journal of Computer vision, 1994, 13(3): 2181 – 2184.

[3] R Zhang, P Tsai, J Cryer, et al. Shape from shading: a survey[J]. IEEE Trans Pattern Analysis & Machine Intelligence, 1999, 21: 690 – 706.

[4] Bao S Y, Savarese S. Semantic structure from motion

[C]//CVPR,2011:2025 – 2032.

[5] C Godard, O M Aodha, M Firman, et al. Digging into self-supervised monocular depth estimation[C]//ICCV, 2019: 3827 – 3837.

[6] Fu H, Gong M, Wang C, et al. Deep ordinal regression network for monocular depth estimation[C]//CVPR, 2018: 2002 – 2011.

[7] Laina I, Rupprecht C, Belagiannis V, et al. Deeper depth prediction with fully convolutional residual networks[C]// 3DV, 2016: 239 – 248.

[8] David Eigen, Christian Puhrsch, Rob Fergus. Depth map prediction from a single image using a multi-scale deep network[C]//NIPS, 2014: 2366 – 2374.

[9] Y Kuznietsov, J Stuckler, B Leibe. Semi-supervised deep learning for monocular depth map prediction [C]// CVPR, 2017: 6647 – 6655.

[10] J Li, R Klein, A Yao. A two-streamed network for estimating fine-scaled depth maps from single RGB images [C]//ICCV, 2017: 3372 – 3380.

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

[12] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]// CVPR, 2012: 3354 – 3361.

[13] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//CVPR, 2016: 770 – 778.

[14] TZhou, M Brown, N Snavely, et al. Unsupervised learning of depth and ego-motion from video[C]//CVPR, 2017: 1851 – 1858.