JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2021
01
VOL.26

ISSN1006-8961
CN11-3758/TB

“自动驾驶技术与应用”专刊

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第26卷第1期（总第297期）
2021年1月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@radi.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@radi.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

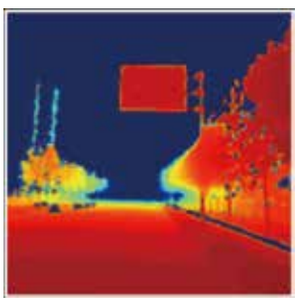
CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2021年1月 第26卷 第1期（总第297期）
中国图象图形学报 Zhongguo Tuxiang Tuxing Xuebao

目次

面向智能驾驶的平行视觉感知：基本概念、框架与应用(第0067页)



Edge-guided GAN：边界信息引导的深度图像修复(第0186页)

利用边缘计算的多车协同激光雷达SLAM(第0218页)

序言 …… 王飞跃，张云泉 I
编者按 …… II

综述

面向智能驾驶测试的仿真场景构建技术综述
任秉韬，邓伟文，白雪松，李江坤，纵瑞雪，朱冰，丁娟 …… 0001
自动驾驶软件测试技术研究综述
冯洋，夏志龙，郭安，陈振宇 …… 0013
强化学习的自动驾驶控制技术研究进展
潘峰，鲍泓 …… 0028
自动驾驶地图的数据标准比较研究
詹骄，郭迟，雷婷婷，屈宜琪，吴杭彬，刘经南 …… 0036
视觉感知的端到端自动驾驶运动规划综述
刘旖菲，胡学敏，陈国文，刘士豪，陈龙 …… 0049

平行驾驶

面向智能驾驶的平行视觉感知：基本概念、框架与应用
李轩，王飞跃 …… 0067
平行视觉的基本框架与关键算法
张慧，李轩，王飞跃 …… 0082

目标检测与跟踪

自动驾驶场景的尺度感知实时行人检测
徐歆恺，马岩，钱旭，张龑 …… 0093
实时视觉目标跟踪与视频对象分割多任务框架
李瀚，刘坤华，刘嘉杰，张晓晔 …… 0101
提升预测框定位稳定性的视频目标检测
郝腾龙，李熙莹 …… 0113
道路结构特征下的车道线智能检测
张翔，唐小林，黄岩军 …… 0123
适用全速域大曲率路径的自动驾驶跟踪算法
张龑，郑颖，鲍泓 …… 0135
伪3D卷积神经网络与注意力机制结合的疲劳驾驶检测
庄员，戚湧 …… 0143
基于眼部自商图—梯度图共生矩阵的疲劳驾驶检测
潘剑凯，柳政卿，王秋成 …… 0154

自动驾驶场景感知与仿真

结合局部平面参数预测的无监督单目图像深度估计
周大可，田径，杨欣 …… 0165
深度纯追随的拟人化无人驾驶转向控制模型
单云霄，黄润辉，何泽，龚志豪，景民，邹雪松 …… 0176
Edge-guided GAN：边界信息引导的深度图像修复
刘坤华，王雪辉，谢玉婷，胡坚耀 …… 0186
引入概率分布的深度神经网络贪婪剪枝
胡骏，黄启鹏，刘嘉昕，刘威，袁淮，赵宏 …… 0198

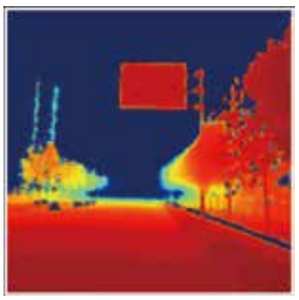
高精地图构建与SLAM

开放道路中匹配高精度地图的在线相机外参标定
廖文龙，赵华卿，严骏驰 …… 0208
利用边缘计算的多车协同激光雷达SLAM
崔明月，钟仕鹏，刘思瑶，李博洋，吴成昊，黄凯 …… 0218

Volume 26, Number 1
Published January 16, 2021

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Parallel visual perception for intelligent driving: basic concept, framework and application (P0067)



Edge-guided GAN: a depth image inpainting approach guided by edge information (P0186)

Cooperative LiDAR SLAM for multi-vehicles based on edge computing (P0218)

Review

Technologies of virtual scenario construction for intelligent driving testing
Ren Bingtao, Deng Weiwen, Bai Xuesong, Li Jiangkun, Zong Ruixue, Zhu Bing, Ding Juan ····· 0001

Survey of testing techniques of autonomous driving software
Feng Yang, Xia Zhilong, Guo An, Chen Zhenyu ····· 0013

Research progress of automatic driving control technology based on reinforcement learning
Pan Feng, Bao Hong ····· 0028

Comparative study on data standards of autonomous driving map
Zhan Jiao, Guo Chi, Lei Tingting, Qu Yiqi, Wu Hangbin, Liu Jingnan ····· 0036

Review of end-to-end motion planning for autonomous driving with visual perception
Liu Yifei, Hu Xuemin, Chen Guowen, Liu Shihao, Chen Long ····· 0049

Parallel Driving

Parallel visual perception for intelligent driving: basic concept, framework and application
Li Xuan, Wang Feiyue ····· 0067

The basic framework and key algorithms of parallel vision
Zhang Hui, Li Xuan, Wang Feiyue ····· 0082

Object Detection and Tracking

Scale-aware EfficientDet: real-time pedestrian detection algorithm for automated driving
Xu Xinkai, Ma Yan, Qian Xu, Zhang Yan ····· 0093

Multitask framework for video object tracking and segmentation combined with multi-scale inter-frame information
Li Han, Liu Kunhua, Liu Jiajie, Zhang Xiaoye ····· 0101

Video object detection method for improving the stability of bounding box
Hao Tenglong, Li Xiying ····· 0113

Intelligent detection of lane based on road structure characteristics
Zhang Xiang, Tang Xiaolin, Huang Yanjun ····· 0123

Autonomous vehicle tracking algorithm for high curvature path in full speed range
Zhang Yan, Zheng Ying, Bao Hong ····· 0135

Driving fatigue detection based on pseudo 3D convolutional neural network and attention mechanisms
Zhuang Yuan, Qi Yong ····· 0143

Fatigue driving detection based on ocular self-quotient image and gradient image co-occurrence matrix
Pan Jiankai, Liu Zhengqing, Wang Qiucheng ····· 0154

Scene Perception and Simulation of Autonomous Driving

Unsurpervised monocular image depth estimation based on the prediction of local plane parameters
Zhou Dake, Tian Jing, Yang Xin ····· 0165

Human-like steering model for autonomous driving based on deep pure pursuit method
Shan Yunxiao, Huang Runhui, He Ze, Gong Zhihao, Jing Min, Zou Xuesong ····· 0176

Edge-guided GAN: a depth image inpainting approach guided by edge information
Liu Kunhua, Wang Xuehui, Xie Yuting, Hu Jianyao ····· 0186

Greedy pruning of deep neural networks fused with probability distribution
Hu Jun, Huang Qipeng, Liu Jiaxin, Liu Wei, Yuan Huai, Zhao Hong ····· 0198

High-Definition Map Construction and SLAM

Online extrinsic camera calibration based on high-definition map matching on public roadway
Liao Wenlong, Zhao Huaqing, Yan Junchi ····· 0208

Cooperative LiDAR SLAM for multi-vehicles based on edge computing
Cui Mingyue, Zhong Shipeng, Liu Siyao, Li Boyang, Wu Chenghao, Huang Kai ····· 0218

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2021)01-0165-11

论文引用格式:Zhou D K, Tian J and Yang X. 2021. Unsurpervised monocular image depth estimation based on the prediction of local plane parameters. Journal of Image and Graphics,26(01):0165-0175(周大可,田径,杨欣. 2021. 结合局部平面参数预测的无监督单目图像深度估计. 中国图象图形学报,26(01):0165-0175)[DOI:10.11834/jig.200364]

结合局部平面参数预测的无监督单目图像深度估计

周大可[1,2],田径[1],杨欣[1]

1. 南京航空航天大学自动化学院, 南京　211100; 2. 江苏省物联网与控制技术重点实验室, 南京　211100

摘　要:**目的**　无监督单目图像深度估计是3维重建领域的一个重要方向,在视觉导航和障碍物检测等领域具有广泛的应用价值。针对目前主流方法存在的局部可微性问题,提出了一种基于局部平面参数预测的方法。**方法**将深度估计问题转化为局部平面参数估计问题,使用局部平面参数预测模块代替多尺度估计中上采样及生成深度图的过程。在每个尺度的深度图预测中根据局部平面参数恢复至标准尺度,然后依据针孔相机模型得到标准尺度深度图,以避免使用双线性插值带来的局部可微性,从而有效规避陷入局部极小值,配合在网络跳层连接中引入的串联注意力机制,提升网络的特征提取能力。**结果**　在KITTI(Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago)自动驾驶数据集上进行了对比实验以及消融实验,与现存无监督方法和部分有监督方法进行对比,相比于最优数据,误差性指标降低了10%～20%,准确性指标提升了2%左右,同时,得到的稠密深度估计图具有清晰的边缘轮廓以及对反射区域更优的鲁棒性。**结论**　本文提出的基于局部平面参数预测的深度估计方法,充分利用卷积特征信息,避免了训练过程中陷入局部极小值,同时对网络添加几何约束,使测试指标及视觉效果更加优秀。

关键词:无监督学习;单目深度估计;注意力机制;局部平面参数估计;局部可微性

Unsurpervised monocular image depth estimation based on the prediction of local plane parameters

Zhou Dake[1,2], Tian Jing[1], Yang Xin[1]

1. College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing　211100, China;
2. Jiangsu Key Laboratory of Internet of Things and Control Technologies, Nanjing　211100, China

Abstract: **Objective** Scene depth information plays a vital role in many current research topics, such as 3D reconstruction, obstacle detection, and visual navigation. Obtaining dense and accurate depth image information often requires expensive equipment, resulting in high costs. The method of using color images for depth estimation does not require expensive equipment and has a wider range of applications. Stereo matching is a traditional method used for estimating the depth with RGB images. A large estimation error is found for weak texture regions because stereo matching relies heavily on feature matching. With the wide application of convolutional neural networks in image processing, the depth estimation of monocular images has been widely investigated. However, the monocular image is essentially a pathological problem because it lacks depth clues related to motion and stereo. Many methods are currently used to estimate the depth of monocular image. Without the use of real depth data, the method of using binocular images for unsupervised learning uses image reconstruction as

收稿日期:2020-07-10;**修回日期**:2020-10-15;**预印本日期**:2020-10-22

基金项目:国家自然科学基金项目(61573182)

Supported by:National Natural Science Foundation of China (61573182)

a supervised signal to train a depth estimation model. This task currently has achieved a large breakthrough although depth estimation depends on the geometric features. How to effectively use the information in the shallow features of the image and how to add geometric constraints to the prediction output while ensuring high convergence performance have been widely investigated to improve the accuracy of depth estimation. In the commonly used multi-scale estimation, the sampling method of bilinear interpolation has local differentiability, easily making the network fall into a local minimum and affecting the training effect. A method based on local plane parameter prediction is proposed to address these problems. This method is applied to multi-scale prediction by using a completely differentiable method with geometric constraints, thereby effectively limiting the convergence of multi-scale depth map prediction in the same direction. **Method** This study presents an unsupervised monocular depth estimation network based on local plane parameter prediction. The main structure is a coding-decoding network and is mainly composed of three parts: a ResNet50-based coding network, a decoding network that introduces a serial double attention mechanism in the skip layer connection, and multi-scale prediction using local plane parameter estimation module. During the training, the network estimates the depth of an image in stereo images, reconstructs another view, and uses the real image of the other view as a supervision for training. Our training set includes 22 600 images in the KITTI (Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago) dataset. The model is built on PyTorch framework, and the input image is 640 × 192 pixels for training. NVIDIA GTX 2080 equipment is used for training, and the training involves 20 epochs. In the multi-scale prediction module, we convert the depth estimation problem into a local plane parameter estimation problem. The local plane parameter prediction module is used to replace upsampling and depth map generation in multi-scale estimation. The depth map prediction of each scale is restored to the standard scale in accordance with the local plane parameters. The standard scale depth map is obtained in accordance with a pinhole camera model to avoid the local differentiability caused by bilinear interpolation, thereby effectively avoiding falling into the local minimum value. A serial attention mechanism is introduced in the network layer hopping connection to obtain clear edge contour information. **Result** We compared our model with multiple unsupervised and supervised methods on the KITTI test dataset. Quantitative evaluation indicators include absolute relative error (Abs Rel), squared relative error (Sq Rel), linear root mean square error (RMSE), logarithmic root mean square error (RMSElog), and threshold accuracy index δ. The dense depth map results for each method are compared. The experimental results show that the proposed method performs well in the depth estimation of various errors and accuracy indicators. In the comparative test, the error indicators are relatively reduced by 10% to 20%, and the accuracy indicators are increased by 1% to 2%. The generated depth map has a relatively clear outline and can separate the important depth values of pedestrians and vehicles from the complex background. It also has a certain robustness to the reflection area, thereby improving the quality of depth estimation. We conducted a series of ablation experiments in the test set to clearly show the effectiveness of the proposed algorithm. **Conclusion** In this study, we proposed a depth estimation method based on local plane parameter prediction. The proposed method utilizes convolution feature information, avoids the local minimum during training, and adds geometric constraints to the network to obtain excellent test indicators and visual effects.

Key words: unsupervised learning; monocular depth estimation; attention mechanism; local plane parameters prediction; local differentiability

0 引 言

场景深度信息在当下许多研究课题中都起着至关重要的作用,如3维立体重建、障碍物检测、遮挡处理与光照估计等(李阳 等,2019;徐维鹏 等,2013)。获得稠密而精确的深度图像信息往往需要昂贵的设备,导致成本较高。与此相比,利用彩色图像进行深度估计的方法,无需昂贵的设备,有着更加广泛的应用范围。立体视觉方法是利用图像估计深度的传统手段之一,由于严重依赖于特征匹配,对于弱纹理区域会有较大的估算误差(Zhao 等,2019)。随着卷积神经网络(convolutional neural network, CNN)在图像领域的广泛应用,对于单目图像的深度估计成为了研究的热点。但由于单目图像缺少运动及立体相关的深度线索,所以本质上是一个病态问题,目前有多种方式来实现对单目图像的深度估计。根据学习方式的不同,基于卷积神经网络的单

目图像深度估计方法主要分为有监督和无监督方法(黄军 等,2019):有监督的单目图像深度估计是直接以设备获得的深度图作为监督信息,通过卷积神经网络产生密集的像素深度估计(Eigen 和 Fergus,2015);而基于无监督的方法则是利用双目图像或视频序列图像,通过预测视差图合成新视点,最小化新视点和目标视点的差距以完成网络的训练,该方法免除了使用设备获取真实深度作为训练数据的过程。

在不使用真实深度数据的情况下,利用双目图像进行无监督学习的方法本质上是使用图像重建作为监督信号来训练深度估计模型。Deep3D(Xie 等,2016)首次提出了预测离散深度的模型,使用合成新视图的方法来完成无监督深度估计,Garg 等人(2016)通过预测连续深度值来扩展这种方法。Godard 等人(2017)在此基础上引入了多尺度估计和左右视点一致性来产生优于监督方法的结果,同时这种利用双目相机信息的无监督方法已经应用到了弱监督数据(Kuznietsov 等,2017;Luo 等,2018)、生成对抗网络(Aleotti 等,2018;Pilzer 等,2018),以及实时使用(Poggi 等,2018a)中。经过长期的研究,利用估计得到的深度图或视差图来合成新视图的方法逐渐成为无监督单目深度估计的标准思路。

然而在合成新视图的过程中,视图之间的像素点不是标准的整齐对应,所以通常使用双线性插值的方法获得新视图,而双线性插值的梯度范围始终来自周围的 4 个坐标点,所以具有局部可微性(Jaderberg 等,2015)。针对此,Godard 等人(2017)引入了多尺度估计的方法,该方法使系统更具鲁棒性,多尺度深度估计实现了从粗到精的深度图预测,使梯度的范围来自于离当前位置更远的点,一定程度上避免了陷入局部极小值,但经实验证明该方法容易生成纹理复制伪影,所以 Godard 等人(2019)对多尺度方法进行了改进,将视差图像的分辨率与用于计算重投影误差的彩色图像分辨率解耦,提升了实验效果,但其中仍然使用了双线性插值上采样的方法,并未做到完全可微,所以如何使用一个完全可微的多尺度深度估计模块来避免陷入局部极小值成为提高单目深度估计精度的重要研究方向。

本文针对无监督单目深度估计任务中存在的局部可微性问题,引入了局部平面参数预测,使用可微的几何方法将低分辨率结果和高分辨率结果联系起来,设计了完全可微并附加几何约束的多尺度深度预测模块,有效限制了多尺度深度图预测结果向同一方向收敛。为避免陷入局部极小值,提升网络精度,除设计完全可微的网络模块等显式方法外,往往还要从网络的特征提取与融合入手。添加注意力机制也是一种常用的手段,注意力机制可以模拟隐性依赖关系,已广泛应用于图像视觉中,如 Fu 等人(2019)引入了空间注意力机制来对不同尺度的特征进行主动选择,获得更好的结果,不同于 Ye 等人(2019)在编码—解码网络中间引入了空间与通道注意力机制,将双注意力机制应用在跳层连接之中,以获取深度图局部特征上的非局部依赖关系。

综上所述,本文针对无监督单目深度估计中存在的易陷入局部极小值问题,提出了局部平面参数预测模块,在多尺度预测中使用完全可微的几何方法来代替常用的双线性插值,引导网络向重建高分辨率图像的方向工作,同时隐式地提升网络的特征提取能力,对网络结构的跳层连接支路采用了串联的空间和通道注意力机制,进一步提高性能,最终得到较为精确的稠密深度图。

1　基于局部平面参数预测的深度估计

1.1　网络整体结构

采用 U-Net(Ronneberger 等,2015)作为基础网络,主要由 3 个部分组成,分别是以 ResNet50(He 等,2016)为基础的编码网络、在跳层连接中引入了串联双注意力机制的解码网络、采用局部平面参数估计的多尺度预测模块,网络整体结构如图 1 所示。

编码部分采用预训练好的 ResNet50 网络,如图 1 所示分为 5 个部分,每部分输出不同尺度的特征图 $\boldsymbol{C}_i$;编码器交替使用卷积层和上采样层将特征图尺度放大,图中每个 Upconv 包括卷积和上采样操作,同时采用了跳跃连接的结构,将浅层特征经过串联的双注意力机制后与深层特征结合,更有效地对浅图像浅层信息进行选择利用;将编码器多个尺度特征输出经过局部平面参数预测模块,得到多个标准尺度深度图 $\boldsymbol{d}^i$。由于本文使用的是无监督方法,所以在得到深度图后需要合成双目对应视图来与真实视图计算重建误差,具体结构如图 2 所示。

图 2 中 $\boldsymbol{I}_1$,$\boldsymbol{I}_2$为双目相机拍摄的一对图像,根据针孔相机模型可知,在已知深度图和相机内参 $\boldsymbol{K}$ 的情况下,可利用式(1)将像素坐标(u_i,v_i)转化为

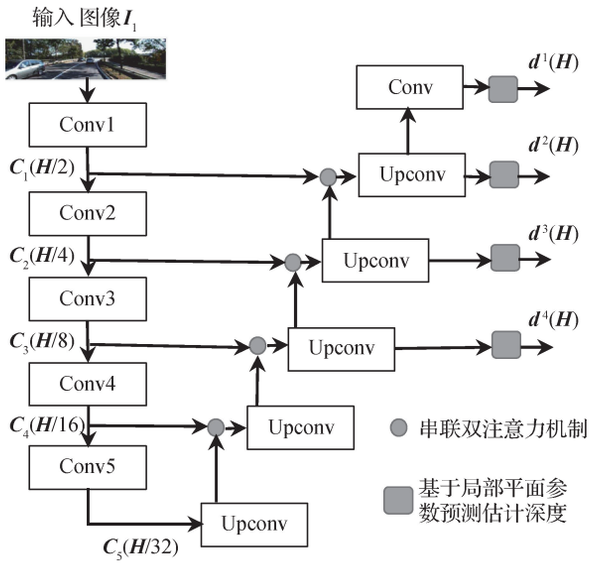


图 1　本文算法整体网络结构示意图

Fig. 1　Overall network structure of our algorithm

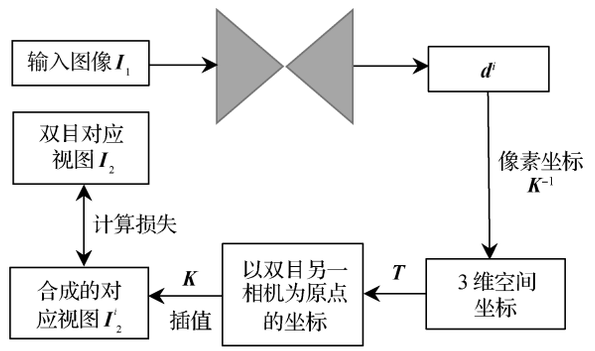


图 2　无监督训练

Fig. 2　Unsupervised training method

3 维空间坐标(x_i, y_i, y_i)，然后将坐标系更换为以双目另一相机光心为原点的坐标系，投影到像素平面中获得坐标对应关系，利用插值的方法可以合成输入图像的双目对应视图$\boldsymbol{I}_2^i$，与真实视图$\boldsymbol{I}_2$计算重建误差即可实现无监督训练

$$z_i\begin{bmatrix}u_i\\v_i\\1\end{bmatrix}=\boldsymbol{K}\begin{bmatrix}x_i\\y_i\\z_i\end{bmatrix} \tag{1}$$

1.2　局部平面参数预测

在深度估计任务中，为了避免训练陷入局部极小值，常常使用多尺度预测的方法来对网络进行约束。Godard 等人(2017)采用多尺度预测来进行图像重建，输出多个尺度的深度图从而合成多个尺度的目标视图，将真实目标视图降采样与多个尺度合成视图进行比较，这倾向于在中等较低分辨率深度图的大型低纹理区域形成纹理复制伪影(即深度图的细节从彩色图像错误地转移)，使得深度网络的任务变得复杂。Godard 等人(2019)对多尺度公式进行了改进，将视差图像的分辨率与用于计算重投影误差的彩色图像解耦，首先将较低分辨率的深度图使用双线性插值的方式上采样到输入图像分辨率，再重新投影计算误差。然而双线性插值具有局部可微性(Jaderberg 等，2015)，双线性插值过程可以表示为

$$\boldsymbol{V}_i=\sum_n^H\sum_m^W\boldsymbol{U}_{nm}\max(0,1-|x_i-m|)\times\max(0,1-|y_i-n|) \tag{2}$$

式中，$\boldsymbol{V}$为双线性插值的输出特征图，$\boldsymbol{U}$为输入特征图，$\boldsymbol{V}_i$为输出特征图的第i个像素值，(x_i, y_i)为$\boldsymbol{V}_i$对应$\boldsymbol{U}$中的坐标，$\boldsymbol{U}_{nm}$为特征图上坐标为(m, n)的值。以对x_i求偏导为例

$$\frac{\partial\boldsymbol{V}_i}{\partial x_i}=\sum_n^H\sum_m^W\boldsymbol{U}_{nm}\max(0,1-|y_i-n|)\times\begin{cases}0 & |m-x_i|\geqslant 1\\1 & m\geqslant x_i\\-1 & m<x_i\end{cases} \tag{3}$$

如式(3)所示，以局部可微的双线性插值作为上采样方法可能会影响网络训练过程，同时双线性插值作为线性插值过程没有考虑每个被预测像素之间的局部相关性，这种弱数据依赖的上采样过程无法产生相对较高质量的深度图。本文提出基于局部平面参数预测的方法来同时替代多尺度预测中的上采样以及深度估计的过程。

本文尝试引入局部平面假设，将特征图引导至全分辨率。具体来说，将深度估计任务中直接预测深度图转化为预测 3 维空间点所在平面参数的问题，使用由尺度为$\boldsymbol{H}/k$特征图中预测的每一组平面参数来代表高分辨率图像中$k\times k$平面的参数，即假设图像中像素点对应的 3 维空间点均在局部$k\times k$平面上。具体网络结构如图 3 所示。即在多尺度估计中将$\boldsymbol{H}/k$特征图利用多个 1×1 卷积核逐步将特征图压缩至 3 通道，前 2 个通道特征值为平面法线的极坐标角度表示$\theta\in[0,\pi]$，$\varphi\in[0,2\pi]$，由极坐标表示方法可得

$$n_1=\sin\theta\cos\varphi \tag{4}$$

$$n_2=\sin\theta\sin\varphi \tag{5}$$

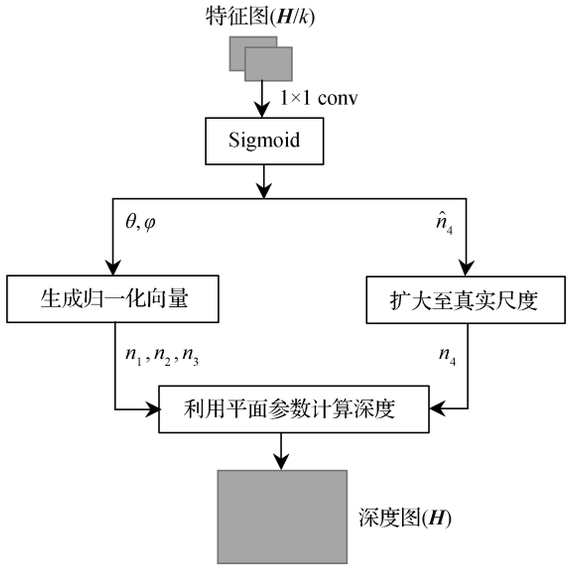


图 3　局部平面参数预测模块结构图

Fig. 3　Local plane parameters prediction model structure

$$n_3 = \cos\theta \tag{6}$$

将第 3 通道估计的参数 $\hat{n}_4$ 进行适当放大作为平面参数中的 n_4，由于与真实世界尺度以及最大深度范围有关，所以设定最大深度值 $d_{\max}$，以及初始值为 1 的可训练参数 γ，得到 n_4 为

$$n_4 = \gamma d_{\max}\hat{n}_4 \tag{7}$$

根据平面参数 (n_1,n_2,n_3,n_4) 即可得到 3 维点所在平面方程，即

$$n_1X + n_2Y + n_3Z + n_4 = 0 \tag{8}$$

式中，X、Y、Z 指 3 维点在空间中的坐标。

而针孔相机模型为

$$z_i\begin{bmatrix}u_i\\v_i\\1\end{bmatrix} = \begin{bmatrix}f_x & 0 & u_0\\0 & f_y & v_0\\0 & 0 & 1\end{bmatrix}\begin{bmatrix}x_i\\y_i\\z_i\end{bmatrix} \tag{9}$$

式中，$(u_i,v_i,1)$ 为像素平面归一化坐标，(x_i,y_i,z_i) 为以光心为原点的 3 维空间点坐标，f_x 和 f_y 均为与相机焦距和尺度相关的固定参数，将式(9)经过变换可得

$$x_i = (u_i - u_0)z_i/f_x \tag{10}$$

$$y_i = (v_i - v_0)z_i/f_y \tag{11}$$

在已知相机内参的情况下，2 维图像像素点坐标可以转化为 3 维空间点坐标，代入预测平面即可获得该像素点对应的深度

$$d_i = |z_i| = \left|\frac{n_4}{n_1(u_i-u_0)/f_x + n_2(v_i-v_0)/f_y + n_3}\right| \tag{12}$$

使用局部平面参数估计，对于 $k \times k$ 区域，只需要 4 个参数即可实现完全可微并具有几何联系的上采样以及深度估计，进而将多个尺度的特征图输入转换成多个标准尺度的深度图，方便进行图像重建及训练。

1.3　串联双注意力机制

通过对深度图像的观察，图像中具有相似外观的邻域像素具有相近的深度，所以可以引入空间和通道注意力机制来捕获深度图局部特征上的非局部依赖关系(Ye 等，2019)。在跳层连接中，将浅层特征与深层特征合并，更好地结合了浅层几何特征以及深层语义特征。考虑到特征图在空间以及维度之间的特征依赖性，引入注意力机制来获得与图像深度估计几何信息关联度较大的特征，增加相关特征的权重。所以本文在网络中每一级跳层连接中加入了注意力机制，将空间注意力机制和通道注意力机制串联使用。

如图 4 所示，从编码网络得到的浅层特征 $\boldsymbol{A} \in \mathbf{R}^{H\times W\times C1}$，以及编码器得到的上采样特征 $\boldsymbol{B} \in \mathbf{R}^{H\times W\times C2}$，其中 H 为特征图的高，W 为特征图的宽，$C1$ 和 $C2$ 为特征图的通道数。首先将特征经过空间注意力机制，空间注意力机制能将更广泛的上下文信息编码为局部特征，从而增强其特征表示能力(Peng 等，2017；Zhao 等，2017)。空间注意力机制本质上为与输入特征图相同大小的权重图，通过网络训练更新权重，指示网络应该关注特征图的部分，本文添加的空间注意力子网络如图 4 所示，将特征图 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 经过 1×1 卷积核压缩至相同通道数 $C1/2$(通道数 $C1$ 的一半)叠加，然后依次经过 ReLU 激活

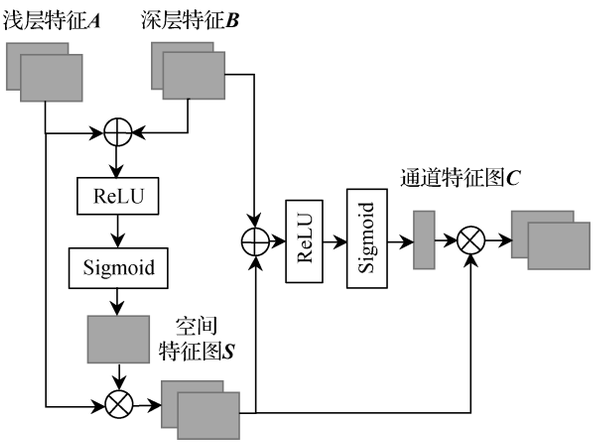


图 4　串联的双注意力机制

Fig. 4　Tandem dual attention mechanism

函数、1×1 卷积核压缩通道,Sigmoid 激活函数,最终得到与特征图相同大小的权重图 $\boldsymbol{S}\in\mathbf{R}^{H\times W\times 1}$,与浅层特征 $\boldsymbol{A}$ 相乘后即为经过了空间注意力机制处理过的特征图 $\boldsymbol{A}'$。跳层连接的浅层特征的每个通道可以被视为若干几何特征的响应,并且彼此相互关联,通过利用通道图之间的相互依赖性,构建了一个通道注意力模块,以明确地模拟通道之间的相互依赖关系。与空间注意力机制同理,将经过空间注意力机制的浅层特征 $\boldsymbol{A}'\in\mathbf{R}^{H\times W\times C1}$ 和 $\boldsymbol{B}$ 经过 1×1 卷积核压缩至相同通道数 $C1$ 后叠加,依次经过最大池化、ReLU 激活函数、1×1 卷积核通道映射,Sigmoid 激活函数最终得到通道方向的权重向量 $\boldsymbol{C}\in\mathbf{R}^{1\times 1\times C1}$,与浅层特征 $\boldsymbol{A}'$ 相乘后即可得到与高级特征拼接的特征图。通过在编码器与解码器之间的跳层连接中添加空间和通道注意力机制可以获得可训练的空间和通道权重,指导网络关注特征图中更有意义的部分,提取到更丰富的特征,同时也能够起到加速收敛的作用。

1.4 损失函数

无监督单目图像深度估计本质上是将学习问题转化为视图合成问题,通过限制网络输出(深度图)执行目标视图的合成,来从模型中提取可解释的深度。与 Zhou 等人(2017)的方法相似,将损失函数 L_{total} 表达为

$$L_{\text{total}}=\sum_{i=1}^{4}\omega_i(L_p^i+L_s^i) \tag{13}$$

使用 4 个尺度的合成视图来计算损失,其中 L_p^i 为合成视图与真实视图之间的匹配损失,L_s^i 为生成深度图的平滑损失,在 Zhou 等人(2017)的基础上,训练过程中使用了权重规划来代替直接对不同尺度的损失求平均,随着网络的训练,标准尺度的损失逐渐占较大的比重,约束网络实现由粗到精的训练过程,权重规划如表 1 所示。

表 1 多尺度损失权重规划

Table 1 Multiscale loss weights schedule

epoch	ω_1	ω_2	ω_3	ω_4
5	0.25	0.25	0.25	0.25
10	0.32	0.16	0.08	0.04
15	0.64	0.32	0.04	0.02
20	1	0	0	0

匹配损失 L_p^i 用于衡量合成的视图与真实视图之间的差距,使用 L1 距离和结构相似性(structural similarity,SSIM)的加权和构建匹配损失函数,即

$$L_p^i=\frac{\alpha}{2}(1-SSIM(\boldsymbol{I}^i,\boldsymbol{I}))+\frac{1}{N}\sum_{x,y}(1-\alpha)\|\boldsymbol{I}_{x,y}^i-\boldsymbol{I}_{x,y}\| \tag{14}$$

式中,$\boldsymbol{I}^i$ 为合成视图,$\boldsymbol{I}$ 为真实视图,$\boldsymbol{I}_{x,y}^i$ 为合成视图的像素点,$\boldsymbol{I}_{x,y}$ 为真实视图的像素点,N 为像素总数,α 为权重系数,经过实验,将系数 α 的值定为 0.85。L1 表达了图像像素值之间的差异性,SSIM 反映合成视图与真实视图的结构相似性,根据 Wang 等人(2004)关于图像结构相似性的论述,其计算为

$$SSIM(\boldsymbol{X},\boldsymbol{Y})=\frac{(2\mu_X\mu_Y+c_1)(2\sigma_{XY}+c_2)}{(\mu_X^2+\mu_Y^2+c_1)(\sigma_X^2+\sigma_Y^2+c_2)} \tag{15}$$

式中,$\boldsymbol{X},\boldsymbol{Y}$ 为一对需要计算 SSIM 的图像,μ_X,μ_Y 分别为 $\boldsymbol{X},\boldsymbol{Y}$ 的像素均值,σ_X,σ_Y 分别表示图像 $\boldsymbol{X},\boldsymbol{Y}$ 的标准差,σ_{XY} 为图像 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的协方差,c_1 和 c_2 为常数,避免分母为 0 而维持稳定,通常 c_1 和 c_2 分别取 2.55^2 和 7.65^2。

平滑损失 L_s^i 用于衡量生成深度图的连续与平滑程度,由于图像中具有相似外观的区域具有相近的深度,则可根据原视图的梯度来指导衡量深度图的平滑程度,用于抑制原视图较为平滑而深度图梯度较大的区域,损失函数为

$$L_s^i=\frac{1}{N}\sum_{x,y}|\partial_x d_{x,y}^i|\mathrm{e}^{-\|\partial_x I_{x,y}\|}+|\partial_y d_{x,y}^i|\mathrm{e}^{-\|\partial_y I_{x,y}\|} \tag{16}$$

式中,$d_{x,y}^i$ 为多尺度估计模块中得到的深度图像素点,$I_{x,y}$ 为真实视图的像素点,∂_x 为求该点 x 方向上的梯度,∂_y 为求该点 y 方向上的梯度,N 为像素总数。

2 实验结果与分析

2.1 实验设置

2.1.1 数据集

使用自动驾驶数据集 KITTI(Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago)(Geiger 等,2012)作为训练数据,以便与现有的工作进行比较。该数据集包含从 61 个场景拍

摄的 42 382 对经过矫正的双目图像数据。原始的图像大小为 1 242 × 375 像素，为了方便在 4 个尺度进行预测计算，将图像的输入大小设为 640 × 192 像素。在实验过程中，使用 Eigen 等人(2014)提出的拆分数据集的方法，使用 697 幅图像作为测试集，其中包含 29 个场景的图像。其余的 32 个场景包含 23 488 幅图像，保留其中的 22 600 幅图像作为训练集，其余图像作为验证集，测试集中图像的真实深度是由激光雷达探测到的 3 维点云投影到左视图中得来的。

2.1.2　训练细节

使用的训练平台为英伟达 RTX 2080，整个算法使用 pytorch 框架实现。输入图像大小为 640 × 192 像素，采用随机水平翻转、亮度及对比度调整来实现数据增强，利用 Adam 作为权值优化算法，初始学习率为 0.000 1，采用多尺度权重规划损失函数，15 个 epoch 后学习率缩小为原来的 1/10，共训练 20 个 epoch，batch 大小为 10，编码器 ResNet50 采用在 ImageNet 预训练好的模型。

2.1.3　评估指标

本文使用单目深度估计最常用的评估指标，其中包括误差指标(值越小结果越好)：平均相对误差(absolute relative error，Abs Rel)、平方相对误差(squared relative error，Sq Rel)、线性均方根误差(linear root mean square error，RMSE)、对数均方根误差(logarithmic root mean square error，RMSElog)，以及阈值准确性指标(值越大结果越好)δ。

2.2　实验结果

2.2.1　消融实验

使用 U-Net 为基础网络，在跳层连接中添加串联的注意力机制，同时在多尺度估计中使用基于局部平面参数预测模块。为了进一步探究增加的模块各部分对实验结果的影响，分别去除串联注意力机制和局部平面参数预测模块，将得到的测试指标结果进行比较，如表 2 所示，其中，去除局部平面参数预测模块后，在多尺度预测中的深度图结果为由粗到精不同尺度。采用两种常用方法实现与真实视图的比较，方法 a 为将真实视图降采样至与合成视图相同尺度进行比较，方法 b 为将预测得到的不同深度图上采样至标准尺度后进行图像重建，得到标准尺度的新视图再与真实视图比较。根据表 2 结果可得，增加注意力机制和增加局部平面参数预测模块均会带来一定的性能提升。

2.2.2　对比实验

本文使用 Eigen 等人(2014)提出的数据集拆分方法，将 KITTI 数据集中激光雷达的数据作为真实数据，据此来评价提出算法的性能，直接利用原始论文的结果，将本文实验结果与目前几种代表性方法进行了定量与定性的对比(各对比算法的结果数据来自于其对应的论文，部分论文中只给出了 80 m 的结果)。表 3 显示了使用深度估计评价指标和其他代表性方法之间的定量比较。其中包括使用真实深度数据做有监督学习，以及使用双目数据或视频序列数据做无监督训练的方法。为了验证不同景深情

表 2　在 KITTI 数据集上消融实验结果

Table 2　Results of ablation study on the KITTI dataset

方法	Abs Rel	Sq Rel	RMSE	RMSElog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
	最大深度值:80 m						
本文	**0.104**	**0.820**	**4.785**	**0.200**	**0.874**	**0.953**	**0.978**
去除注意力机制	0.108	0.856	4.852	0.205	0.865	0.950	0.976
去除局部平面参数预测 a	0.140	1.340	5.816	0.242	0.813	0.930	0.967
去除局部平面参数预测 b	0.120	1.117	5.206	0.207	0.866	0.949	0.975
	最大深度值:50 m						
本文	**0.098**	**0.606**	**3.626**	**0.189**	**0.886**	**0.958**	**0.979**
去除注意力机制	0.104	0.762	3.789	0.193	0.883	0.953	0.977
去除局部平面参数预测 a	0.135	0.965	4.460	0.222	0.820	0.935	0.970
去除局部平面参数预测 b	0.109	0.889	3.832	0.196	0.879	0.955	0.978

注：加粗字体表示各列最优结果。

表 3　在 KITTI 数据集上结果比较

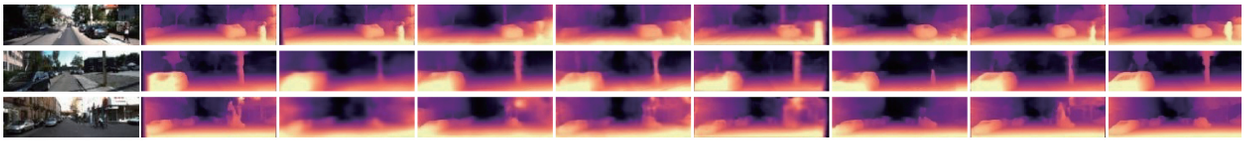
Table 3　Results on the KITTI set

方法	是否监督	Abs Rel	Sq Rel	RMSE	RMSElog	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
				最大深度值:80 m				
Eigen 等人(2014)	是	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu 等人(2015)	是	0.201	1.584	6.471	0.273	0.680	0.898	0.958
Zhou 等人(2017)	否	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Garg 等人(2016)	否	0.152	1.226	5.849	0.246	0.784	0.921	0.967
Godard 等人(2017)	否	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Zhan 等人(2018)	否	0.135	1.132	5.585	0.229	0.820	0.933	0.971
GeoNet(Yin 和 Shi,2018)	否	0.155	1.296	5.857	0.233	0.793	0.931	0.973
DDVO(Wang 等,2018)	否	0.151	1.257	5.583	0.228	0.810	0.936	0.974
Ranjan 等人(2019)	否	0.148	1.149	5.464	0.226	0.815	0.935	0.973
3Net(Poggi 等,2018b)	否	0.119	1.201	5.888	0.208	0.844	0.941	0.978
AsiANet(Yusiong 和 Naval,2019)	否	0.145	1.349	5.909	0.230	0.824	0.936	0.970
本文	否	**0.104**	**0.820**	**4.785**	**0.200**	**0.874**	**0.953**	**0.978**
				最大深度值:50 m				
Zhou 等人(2017)	否	0.201	1.391	5.181	0.264	0.696	0.900	0.966
Garg 等人(2016)	否	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Godard 等人(2017)	否	0.140	0.976	4.471	0.232	0.818	0.931	0.969
Zhan 等人(2018)	否	0.128	0.815	4.204	0.216	0.835	0.941	0.975
GeoNet(Yin 和 Shi,2018)	否	0.147	0.936	4.348	0.218	0.810	0.941	0.977
AsiANet(Yusiong 和 Naval,2019)	否	0.122	0.786	4.014	0.198	0.864	0.953	0.978
本文	否	**0.098**	**0.606**	**3.626**	**0.189**	**0.886**	**0.958**	**0.979**

注:加粗字体表示各列最优结果,DDVO 为 differentiable implementation of direct visual odometry。

况下的算法性能,表 3 中上半部分为将最大深度设为 80 m 时各个方法预测得到的指标,下半部分为将最大深度设为 50 m 时各个方法预测得到的指标。可以看出,在不同景深的情况下,本文在误差指标和准确率指标上均优于目前无监督方法以及部分有监督方法,实现了所有指标的最佳性能。证明了将深度估计问题转化为局部平面参数估计的可行性以及性能的优越性。

由于测试集的真实数据是由激光雷达获得的点云数据投影至图像平面中,导致真实的深度数据是非常稀疏的,所以测试得到的指标仅为验证算法性能的一部分依据。图 5 为测试数据集中几幅具有代表性的测试图像在不同算法上得到的深度图结果,测试图像中包括各种类型的车辆,不同姿态的行人以及树木、电线杆等,这类物体的真实深度在深度估计的应用中相比街道背景占有更重要的地位,所以深度图中各个物体的边缘细致情况成为评估深度估计效果的重要参考。如在图 5 中,第 1 幅图右下角的行人,第 2 幅图中的电线杆等,使用本文方法得到了更加细致的轮廓,而在 Ranjan 等人(2019)的方法中几乎没有体现,在其他方法中也较为模糊。由于道路中情况非常复杂,如图 5 中第 3 幅图复杂交通背景下的骑自行车的人,很难将自行车的深度信息从背景中较好地隔离出来。本文方法能够将自行车的大致轮廓较好地从杂乱的背景中分离出来,而 Godard 等人(2017)方法无法将自行车很好地分离出来,DDVO(Wang 等,2018)等算法得到的深度图边界相对模糊。与目前具有代表性的算法估计的深



(a) 原图　(b) Godard 等人(2017)　(c) Zhou 等人(2017)　(d) DDVO　(e) GeoNet　(f) Zhan 等人(2018)　(g) Ranjan 等人(2019)　(h) 3Net　(i) 本文

图 5　测试集算法对比结果

Fig. 5　Comparison of experimental results on test set

((a) input image; (b) Godard et al. (2017); (c) Zhou et al. (2017); (d) DDVO; (e) GeoNet; (f) Zhan et al. (2018); (g) Ranjan et al. (2019); (h) 3Net; (i) ours)

度图相比,本文算法可以获得边缘更清晰的深度图,同时可以较好地将行人等重要物体从复杂背景中提取出来。

目前使用无监督的单目深度估计均是利用图像重建来指导网络估计图像深度,当场景中包含违反朗伯假设的物体时(如场景中某些汽车的表面),得到的图像深度图就会出现中断,如图 6 所示,输入图像的右下角车辆表面有部分反射和颜色饱和区域,导致 Godard 等人(2019)的深度估计结果会在车表面出现深度图的中断,本文方法由于使用局部平面参数预测,在多尺度预测中使用局部平面假设生成标准尺度深度图,所以一定程度上能够对深度图进行平滑,对于图像中有反射和颜色饱和的区域具有更好的鲁棒性,如图 6 中本文算法得到的深度图汽车表面深度较为平滑,没有中断的区域。

(a) 原图　(b) Godard等人(2019)　(c) 本文

图 6　反射区域结果对比

Fig. 6　Comparison of reflection area results

((a) input image; (b) Godard et al. (2019); (c) ours)

综上所述,本文算法在深度估计各项误差和准确率指标上表现较好,同时生成的深度图具有较为清晰的轮廓,能够将行人车辆等较为重要的深度值从复杂背景中分离出来,同时对反射区域也有一定的鲁棒性,提高了深度估计的质量。

3　结　论

为了避免在多尺度预测中使用双线性采样带来的局部可微问题,本文提出了一种基于局部平面参数预测的无监督单目深度估计方法,将深度预测问题转换为各尺度局部平面参数预测问题,使用完全可微的方式代替双线性采样来将多尺度预测恢复到标准尺度,在恢复过程中添加了邻域的几何关联,同时为网络添加了一定的几何约束。结合在跳层连接中添加串联的双注意力机制,可以更好地利用图像特征信息。本文使用双目图像数据进行训练,实现了无监督学习,在 KITTI 自动驾驶数据集上各项指标均优于目前先进的无监督方法以及部分有监督方法,在视觉效果上有更好的鲁棒性以及更清晰的深度边界。

本文方法的不足之处是:在生成的深度图中,部分纤细物体如路灯、树木等深度图轮廓存在不完整的情况,原因是目前提取到的特征不足以产生更加精细化的深度图轮廓。在今后的进一步研究中,将考虑通过结合语义信息来指导生成更精细的深度图。

参考文献(References)

Aleotti F, Tosi F, Poggi M and Mattoccia S. 2018. Generative adversarial networks for unsupervised monocular depth prediction//Proceedings of 2018 European Conference on Computer Vision. Munich, Germany: Springer: 337-354 [DOI: 10.1007/978-3-030-11009-3_20]

Eigen D and Fergus R. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2650-2658 [DOI: 10.1109/ICCV.2015.304]

Eigen D, Puhrsch C and Fergus R. 2014. Depth map prediction from a single image using a multi-scale deep network//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: NIPS: 2366-2374 [DOI: 10.5555/

2969033.2969091]

Fu J, Liu J, Tian H J, Li Y, Bao Y J, Fang Z W and Lu H Q. 2019. Dual attention network for scene segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3141-3149 [DOI: 10.1109/cvpr.2019.00326]

Garg R, Kumar B G V, Carneiro G and Reid I. 2016. Unsupervised CNN for single view depth estimation: geometry to the rescue//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 740-756 [DOI: 10.1007/978-3-319-46484-8_45]

Geiger A, Lenz P and Urtasun R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 3354-3361 [DOI: 10.1109/CVPR.2012.6248074]

Godard C, Aodha O M and Brostow G J. 2017. Unsupervised monocular depth estimation with left-right consistency//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6602-6611 [DOI: 10.1109/CVPR.2017.699]

Godard C, Aodha O M, Firman M and Brostow G. 2019. Digging into self-supervised monocular depth estimation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE: 3827-3837 [DOI: 10.1109/iccv.2019.00393]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Huang J, Wang C, Liu Y and Bi T T. 2019. The progress of monocular depth estimation technology. Journal of Image and Graphics, 24(12): 2081-2097 (黄军, 王聪, 刘越, 毕天腾. 2019. 单目深度估计技术进展综述. 中国图象图形学报, 24(12): 2081-2097) [DOI: 10.11834/jig.190455]

Jaderberg M, Simonyan K, Zisserman A and Kavukcuoglu K. 2015. Spatial transformer networks//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: IEEE: 2017-2025

Kuznietsov Y, Stückler J and Leibe B. 2017. Semi-supervised deep learning for monocular depth map prediction//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2215-2223 [DOI: 10.1109/CVPR.2017.238]

Li Y, Chen X W, Wang Y and Liu M L. 2019. Progress in deep learning based monocular image depth estimation. Laser and Optoelectronics Progress, 56(19): #190001 (李阳, 陈秀万, 王媛, 刘茂林. 2019. 基于深度学习的单目图像深度估计的研究进展. 激光与光电子学进展, 56(19): #190001) [DOI: 10.3788/LOP56.190001]

Liu F Y, Shen C H and Lin G S. 2015. Deep convolutional neural fields for depth estimation from a single image//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 5162-5170 [DOI: 10.1109/CVPR.2015.7299152]

Luo Y, Ren J, Lin M D, Pang J H, Sun W X, Li H S and Lin L. 2018. Single view stereo matching//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 155-163 [DOI: 10.1109/cvpr.2018.00024]

Peng C, Zhang X Y, Yu G, Luo G M and Sun J. 2017. Large kernel matters-improve semantic segmentation by global convolutional network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1743-1751 [DOI: 10.1109/cvpr.2017.189]

Pilzer A, Xu D, Puscas M, Ricci E and Sebe N. 2018. Unsupervised adversarial depth estimation using cycled generative networks//Proceedings of 2018 International Conference on 3D Vision. Verona, Italy: IEEE: 587-595 [DOI: 10.1109/3dv.2018.00073]

Poggi M, Aleotti F, Tosi F and Mattoccia S. 2018a. Towards real-time unsupervised monocular depth estimation on CPU//Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid, Spain: IEEE: 5848-5854 [DOI: 10.1109/IROS.2018.8593814]

Poggi M, Tosi F and Mattoccia S. 2018b. Learning monocular depth estimation with unsupervised trinocular assumptions//Proceedings of 2018 International Conference on 3D Vision. Verona, Italy: IEEE: 324-333 [DOI: 10.1109/3dv.2018.00045]

Ranjan A, Jampani V, Balles L, Kim K, Sun D Q, Wulff J and Black M J. 2019. Competitive collaboration: joint unsupervised learning of depth, camera motion, optical flow and motion segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 12232-12241 [DOI: 10.1109/cvpr.2019.01252]

Ronneberger O, Fischer P and Brox T. 2015. U-net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Wang C Y, Buenaposada J M, Zhu R and Lucey S. 2018. Learning depth from monocular videos using direct methods//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2022-2030 [DOI: 10.1109/CVPR.2018.00216]

Wang Z, Bovik A C, Sheikh H R and Simoncelli E P. 2004. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4): 600-612 [DOI: 10.1109/TIP.2003.819861]

Xie J Y, Girshick R and Farhadi A. 2016. Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural net-

works//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 842-857 [DOI: 10.1007/978-3-319-46493-0_51]

Xu W P, Wang Y T, Liu Y and Weng D D. 2013. Survey on occlusion handling in augmented reality. Journal of Computer-Aided Design and Computer Graphics, 25(11): 1635-1642 (徐维鹏, 王涌天, 刘 越, 翁冬冬. 2013. 增强现实中的虚实遮挡处理综述. 计算机辅助设计与图形学学报, 25(11): 1635-1642)

Ye X C, Zhang M L, Xu R, Zhong W, Fan X, Liu Z and Zhang J A. 2019. Unsupervised monocular depth estimation based on dual attention mechanism and depth-aware loss//Proceedings of 2019 IEEE International Conference on Multimedia and Expo. Shanghai, China: IEEE: 169-174 [DOI: 10.1109/ICME.2019.00037]

Yin Z C and Shi J P. 2018. Geonet: unsupervised learning of dense depth, optical flow and camera pose//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1983-1992 [DOI: 10.1109/CVPR.2018.00212]

Yusiong J P T and Naval P C. 2019. AsiANet: Autoencoders in autoencoder for unsupervised monocular depth estimation//Proceedings of 2019 IEEE Winter Conference on Applications of Computer Vision. Waikoloa Village, USA: IEEE: 443-451 [DOI: 10.1109/wacv.2019.00053]

Zhan H Y, Garg R, Weerasekera C S, Li K J, Agarwal H and Reid I M. 2018. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 340-349 [DOI: 10.1109/CVPR.2018.00043]

Zhao H S, Shi J P, Qi X J, Wang X G and Jia J Y. 2017. Pyramid scene parsing network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6230-6239 [DOI: 10.1109/CVPR.2017.660]

Zhao S Y, Zhang L, Shen Y, Zhao S J and Zhang H J. 2019. Super-resolution for monocular depth estimation with multi-scale sub-pixel convolutions and a smoothness constraint. IEEE Access, 7: 16323-16335 [DOI: 10.1109/ACCESS.2019.2894651]

Zhou T H, Brown M, Snavely N and Lowe D G. 2017. Unsupervised learning of depth and ego-motion from video//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6612-6619 [DOI: 10.110 s9/CVPR.2017.700]

作者简介

周大可,1974 年生,男,副教授,主要研究方向为计算机视觉。
E-mail:dkzhou@ nuaa. edu. cn

田径,男,硕士研究生,主要研究方向为计算机视觉。
E-mial:jingtian96@ nuaa. edu. cn

杨欣,男,副教授,主要研究方向为计算机视觉。
E-mail:yangxin@ nuaa. edu. cn