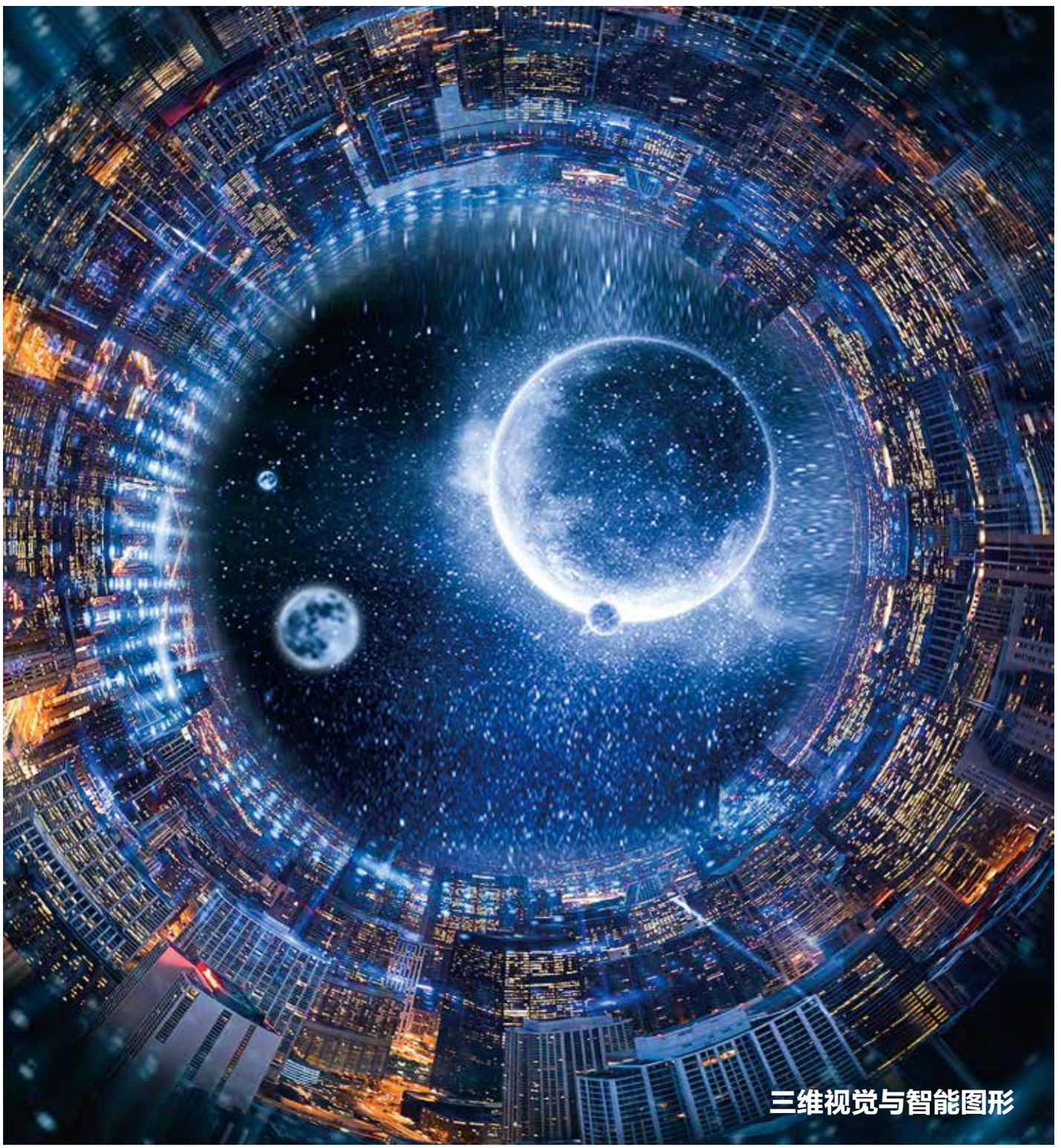
JOURNAL OF IMAGE AND GRAPHICS

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

2022
02
VOL.27

ISSN1006-8961
CN11-3758/TB



三维视觉与智能图形

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第27卷第2期（总第310期）
2022年2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@aircas.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@aircas.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

2022年2月 第27卷 第2期（总第310期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

面向本征图像分解的高质量渲染数据集与非局部卷积网络(第0404页)

高效鲁棒三维结构化重建(第0421页)

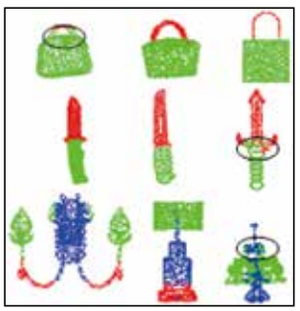
利用隐式解码器的三维模型簇协同分割(第0550页)

Volume 27, Number 2
Published February 16, 2022

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

High quality rendered dataset and non-local graph convolutional network for intrinsic image decomposition(P0404)

Efficient and robust 3D structure-aware reconstruction(P0421)



Co-segmentation of 3D shape clusters based on implicit decoder(P0550)

E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2022)02-0390-14

论文引用格式:Luo H L and Zhou Y F. 2022. Review of monocular depth estimation based on deep learning. Journal of Image and Graphics,27(02):0390-0403(罗会兰,周逸风. 2022. 深度学习单目深度估计研究进展. 中国图象图形学报,27(02):0390-0403)[DOI:10.11834/jig.200618]

深度学习单目深度估计研究进展

罗会兰*,周逸风

江西理工大学信息工程学院,赣州　341000

摘　要:单目深度估计是从单幅图像中获取场景深度信息的重要技术,在智能汽车和机器人定位等领域应用广泛,具有重要的研究价值。随着深度学习技术的发展,涌现出许多基于深度学习的单目深度估计研究,单目深度估计性能也取得了很大进展。本文按照单目深度估计模型采用的训练数据的类型,从3个方面综述了近年来基于深度学习的单目深度估计方法:基于单图像训练的模型、基于多图像训练的模型和基于辅助信息优化训练的单目深度估计模型。同时,本文在综述了单目深度估计研究常用数据集和性能指标基础上,对经典的单目深度估计模型进行了性能比较分析。以单幅图像作为训练数据的模型具有网络结构简单的特点,但泛化性能较差。采用多图像训练的深度估计网络有更强的泛化性,但网络的参数量大、网络收敛速度慢、训练耗时长。引入辅助信息的深度估计网络的深度估计精度得到了进一步提升,但辅助信息的引入会造成网络结构复杂、收敛速度慢等问题。单目深度估计研究还存在许多的难题和挑战。利用多图像输入中包含的潜在信息和特定领域的约束信息,来提高单目深度估计的性能,逐渐成为了单目深度估计研究的趋势。

关键词:单目视觉;场景感知;深度学习;3维重建;深度估计

Review of monocular depth estimation based on deep learning

Luo Huilan*, Zhou Yifeng

School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou　341000, China

Abstract: The development of computer technology promotes the development of computer vision. Nowadays, more researchers focus on the field of 3D vision while monocular depth estimation is one of the basic tasks of 3D vision. Depth estimation from a single image is a critical technology for obtaining scene depth information. This technology has important research value because it has potential applications in intelligent vehicles, robot positioning, and other fields. Compared with traditional depth acquisition methods, monocular depth estimation based on deep learning has the advantages of low cost and simple operation. With the development of deep learning technology, many studies on monocular depth estimation based on deep learning have emerged in recent years, and the performance of monocular depth estimation has made great progress. The monocular depth estimation model needs a large a large amount of data to train the model. The commonly used training data types include RGB and depth (RGB-D) image pairs, stereo image pairs, and image sequences. The depth estimation model training by RGB-D images first extracts the image features through convolutional neural network and then predicts the depth map by using the method of continuous depth value regression. After predicting the depth map, sev-

收稿日期:2020-11-13;**修回日期**:2021-03-09;**预印本日期**:2021-03-16

*通信作者:罗会兰　luohuilan@sina.com

基金项目:国家自然科学基金项目(61462035,61862031);江西省教育厅科技项目(GJJ180483,GJJ200859,GJJ200884)

Supported by: National Natural Science Foundation of China (61462035, 61862031); Science and Technology Research Project of Jiangxi Provincial Department of Education (GJJ180483, GJJ200859, GJJ200884)

eral models use conditional random fields or other methods to optimize the depth map. Unsupervised learning is often used to train the monocular depth estimation model when the training data types are stereo image pairs and image sequences. The monocular estimation model training by stereo image pairs first predicts the disparity map and then estimates depth by using the disparity map. When an image sequence is used to train the model, the model first predicts the depth map of an image in the image sequence, and then the depth estimation model is optimized by images reconstructed by the depth map and other images in the sequence. To improve the accuracy of depth estimation, several researchers use semantic tags, depth range, and other auxiliary information to optimize depth maps. Several data sets can be used for multiple computer vision tasks such as depth estimation and semantic segmentation. Several researchers improve the accuracy of depth estimation by learning depth estimation and semantic segmentation model jointly because depth estimation has a strong correlation with semantic segmentation. When establishing the depth estimation data set, depth camera or light laser detection and ranging (LiDAR) is used to obtain the scene depth. Depth camera and LiDAR are based on the principle that light and other propagation media will reflect when they encounter objects. The depth range obtained by depth cameras and LiDAR is fixed because the propagation medium is dissipated in the transmission, and depth cameras and LiDAR cannot measure depth while the propagation medium energy is very small. Several models first divide the depth range into several depth intervals, take the median value of the depth interval as the depth value of the interval, and then use the method of multiple classifications to predict the depth map. Different training data types not only result in different network model structures but also affect the accuracy of depth estimation. In this review, the current monocular depth estimation methods based on deep learning are surveyed from the perspective of the training data type used by the monocular depth estimation model. Moreover, the single-image training model, the multi-image training model, and the monocular depth estimation model of auxiliary information optimization training are separately discussed. Furthermore, the latest research status of monocular depth estimation is systematically analyzed, and the advantages and disadvantages of various methods are discussed. Finally, the future research trends of monocular depth estimation are summarized.

Key words: monocular vision; scene perception; deep learning; 3D reconstruction; depth estimation

0 引 言

从场景中获取深度信息是计算机视觉的基础问题,在智能汽车(胡云峰 等,2019)和机器人定位(Cadena 等,2016)等领域应用广泛。获取深度信息的传统方法是激光雷达。激光雷达通过测量激光的折返时间精确地获取深度信息,但由于价格等原因很难大规模应用。与传统方法相比,基于图像的深度估计没有过高的硬件要求,只需要拍摄图像就能得到深度信息。因此,基于图像的深度估计具有更广泛的应用范围和使用人群。基于图像的深度估计分为基于单目视觉的深度估计方法和以立体视觉技术为代表的多目视觉深度估计方法。测试时,单目视觉从单幅图像中获取场景深度信息,而立体视觉技术类似于人眼的成像原理,需要使用两个摄像头获取同一场景的两幅图像,然后对两幅图像进行匹配,利用匹配信息得到深度信息。在立体视觉技术中,左右两图像的基线距离和相机的焦距已知,已知的相机参数信息有助于精度的提高,但左右匹配问题受到设备的限制,从而使得双目视觉应用范围有较大局限性。相较而言,单目深度估计只需单幅场景图像即可估计深度信息,不需要额外的信息,所以具有更广泛的应用价值和重要的研究意义,逐渐成为当前计算机视觉领域的研究热点之一。

从图像中获得深度信息的本质是构建一个关联图像信息和深度信息的模型(Saxena 等,2005),随着深度学习的迅速发展,深度卷积神经网络在单目视觉中的应用成为研究热点。深度学习在解决目标识别、图像分割等问题中性能优异,这些方法也大量迁移应用于深度估计。从 Eigen 等人(2014)首次将神经网络应用于深度估计至今,随着 VGG(Visual Geometry Group)(Simonyan 和 Zisserman, 2015)、ResNet(He 等,2016)等网络框架和空洞卷积(Yu 和 Koltun,2016)等卷积结构相继提出,出现了大量性能优异的深度估计网络。普通相机在拍摄时只记录场景的色彩信息,深度信息的丢失导致在不同场景下拍摄的图像可能是相同的,所以单目深度估计是一个病态问题。由于单目深度估计中图像与深度图的对应关系存在一对多的可能性,引入辅助信息提

升深度估计精度是众多科研者的思路。

已经训练好的单目深度估计模型在测试或实际使用时,只需要输入单幅图像就可以得到深度估计结果。但在训练学习时,有些模型只需要使用单幅图像训练(Eigen 等,2014;Laina 等,2016)。也有许多研究者为了提升模型性能,使用立体图像对(Godard 等,2017;孙蕴瀚 等,2020)、图像序列(Yin 和 Shi,2018;Zhou 等,2017;Zou 等,2018)和深度范围(Fu 等,2018;Lee 等,2019)来训练模型。目前,深度估计最常用的数据集是 NYU depth v2(Silberman 等,2012)数据集和 KITTI(Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago)(Geiger 等,2012)数据集。NYU depth v2 是美国纽约大学 Silberman 团队使用 Kinect 深度相机创建的室内深度估计数据集,包含 464 个室内场景的彩色图像和对应的深度图,有效深度范围为 0.5 ~ 10 m。KITTI 数据集由数据采集车沿途采集的彩色图像、深度图和 GPS(global positioning system)位置等信息组成,包括城市街道和乡间小道等道路场景,是常用的室外场景深度估计数据集。单图像训练模型采用特征提取主干网络和上采样网络获取深度图,网络结构简单,易于针对特定任务调整结构,并且对硬件要求低,但是需要像素级的深度真值图进行训练。多图像训练模型分为视差模型和运动结构重建模型。视差模型分别估计输入图像的左右视差图,并通过视差图学习图像的深度图;运动结构重建模型由深度估计模块和相机轨迹估计模块组成。多图像训练单目深度估计模型使用图像间的空间关系或时序关系构造监督信息,在训练时不需要图像的深度真值,训练数据很容易获得,所以在大数据量上训练得到的模型的泛化性能也较好。深度范围也可以用来训练模型,一些单目深度估计模型先将深度范围离散化,然后根据卷积神经网络提取的特征估计每一个像素点的深度范围。

与传统的深度获取方法和多目深度估计方法相比,单目深度估计只需一张照片就能获得场景的深度信息,应用范围更广泛,是计算机视觉的研究热点。2014 年以来,许多研究者从事单目深度估计的研究,提出了不同的单目深度估计模型。Khan 等人(2020)和 Zhao 等人(2020)以模型使用监督学习、无监督学习或半监督学习为依据,对单目深度估计研究进行了分类综述。与上述综述不同,本文从单目深度估计模型采用训练数据类型的角度,对近年来基于深度学习的单目深度估计研究进行综述,将单目深度估计方法分为单图像训练模型、多图像训练模型、语义及数据集深度范围优化深度估计,旨在为根据获得的训练数据类型和辅助信息选择和设计模型提供参考。

1 单图像输入训练模型

单图像输入的训练模型以单幅彩色图像为输入,参考深度信息为监督信息,训练网络模型实现深度估计。在使用卷积神经网络拟合得到深度图后,还可以通过一些优化方法对深度图进行优化以提高预测精度。

1.1 卷积神经网络拟合深度图

卷积神经网络拟合深度图的方法是使用卷积神经网络提取图像特征后,采用全连接或上采样的方法获取图像的深度图。此类方法性能的提升依赖于深度卷积神经网络的学习能力。随着深度学习技术的发展,出现了 AlexNet(Krizhevsky 等,2017)、VGG(Simonyan 和 Zisserman,2015)和 ResNet(He 等,2016)等网络结构。新型网络结构提高了网络的深度,更深的网络模型可以提取更丰富的特征信息,有利于提高拟合深度图的精度。Eigen 等人(2014)将 AlexNet 应用于深度估计,并提出由粗到细的两步策略。网络中使用粗细两个尺度的神经网络,其中粗尺度网络获取全局深度,细尺度网络在粗尺度网络输出的基础上丰富局部细节。粗尺度网络使用 AlexNet 网络作为主干网络,输入的彩色图像经过 AlexNet 网络提取特征,这些特征通过全连接层映射为分辨率是输入图像 1/4 的粗略深度图。在细尺度网络中,使用卷积提取原图像的局部特征并与全局深度图拼接,经过卷积获得最终的深度图。但 Eigen 等人(2014)提出的网络缺少上采样层,使得网络输出的深度图分辨率较小,细尺度网络的优化效果也不太理想,深度估计的精度有所欠缺。2014 年,牛津大学提出了 VGG 网络(Simonyan 和 Zisserman,2015),使用多个小卷积核的叠加替代 AlexNet 中的大卷积核,加深了网络,从而提高了网络性能。在 Eigen 等人(2014)的基础上,Eigen 和 Fergus(2015)采用 VGG 网络提取图像的特征,并在 Eigen 等人(2014)提出的网络结构的基础上加入第三尺度网

络,进一步优化细节,同时网络中增加上采样层,使输出深度图的尺寸提高,为输入图像的 1/2。

Eigen 等人(2014)使用全连接层生成深度图,全连接层巨大的参数量限制了深度图分辨率的提高。Shelhamer 等人(2015)在为解决语义分割问题提出的全卷积网络(fully convolutional networks,FCN)中使用特征图取代全连接的结果,经过上采样将特征图还原到输入图像的尺寸,在更少参数量的情况下获得了原图像同样分辨率的输出。在此基础上,Laina 等人(2016)将 ResNet 和 FCN 网络应于单目深度估计。网络在特征提取端使用 ResNet50 提取彩色图像的深层特征,特征图通过 4 次上采样输出分辨率为原图像 1/2 的深度图,并使用 reverse Huber(Owen,2007;Zwald 和 Lambert-Lacroix,2012)函数替代 L_2 范数为损失函数提高深度估计精度。由于 ResNet 有效缓解了梯度消失问题,相比 Eigen 等人(2014)提出的模型,Laina 等人(2016)的模型的全局误差显著降低,深度估计的精度明显提高。

Laina 等人(2016)提出的网络模型存在物体形状扭曲、小物体缺失和物体边界模糊等问题,一些研究者对此提出新的网络模型。Hu 等人(2019)认为不同尺度的特征图包含不同的深度信息,基于这个观点,Hu 等人(2019)采用与 Laina 等人(2016)类似的主干网络,不同尺度的特征图上采样后与主干网络的输出融合生成最终的深度图,尝试使用 ResNet、DenseNet(Huang 等,2017)和 SENet(Hu 等,2020)提取特征,其中 SENet 的深度估计效果最为优异。网络充分融合多尺度特征,有效增强了深度图局部细节,提高了深度估计的精度。与 Hu 等人(2019)方法不同,Zhao 等人(2019)使用 Shi 等人(2016)提出的亚像素卷积融合多尺度特征,提高模型对细节的深度预测能力。为了提高物体边界处的深度估计精度,谢昭等人(2020)基于编解码结构提出了一种采样汇集网络模型,在上采样和下采样层中加入局部汇集模块,通过采样汇集跨层的上采样网络提高特征图分辨率,局部汇集模块的使用减少了物体边界定位误差对深度估计的影响,提升了估计的精确度。

1.2 利用图像特征优化深度图

在卷积网络模型估计得到深度图后,可以利用输入图像的区域相似性和梯度等图像自身的特征进一步优化深度图,其中利用图像相似性优化深度图的方法主要有条件随机场(conditional random field,CRF)(Lafferty 等,2001)。Liu 等人(2015)认为图像中相似区域的深度相近,根据图像的区域相似性使用连续条件随机场优化深度图。首先,将图像划分为多个区域块,即超像素,然后通过神经网络估计超像素的深度值,并根据相邻超像素生成相似矩阵优化深度估计。在 CRF 模型中,超像素作为图模型的节点,每个超像素的深度映射函数用 U 表示,节点间的连线用二元项 V 表示,二元项反映节点间的相似度,卷积神经网络通过优化 U 和 V 这两个函数,使得相似度高的超像素有相近的深度值。输入图像分为 N 个超像素,以超像素的质心为中心在原图中截取图像块作为一元项 U 的训练图像,经过卷积层和全连接操作后输出超像素质心对应的深度值。N 个超像素有 K 组相邻超像素对,K 组超像素对的相似度组成相似性矩阵。一元项部分的网络以 L_2 范数为损失函数,二元项以 CRF 损失进行优化。但网络模型需要计算超像素的深度和超像素间的相似性,计算量大,耗时严重,而且深度估计的结果以超像素为单位降低了深度估计的精度。在此基础上,Liu 等人(2016)提出使用掩膜层加快计算,将图像分割成多个超像素作为超像素掩膜层。彩色图像经过卷积和最近邻插值得到与输入图像分辨率相同的深度图,深度图与超像素掩膜层结合,将同一超像素中的深度值平均后作为超像素的深度值,再进行条件随机场优化。由于神经网络不需要重复估计超像素的深度值,网络的运行速度得到提升。

Xu 等人(2017)将条件随机场作为独立的网络分支,设计了一个由卷积神经网络和条件随机场组成的深度估计网络,卷积神经网络使用全卷积网络产生深度图,在上采样网络中将不同尺度特征图生成的深度图逐层输入条件随机场,得到精度逐步提高的深度图。与使用条件随机场优化深度图的方法不同,Li 等人(2017)利用图像梯度信息优化深度估计,设计了一个独立分支充分提取输入图像的梯度特征,并与深度估计融合,达到优化深度图的目的。Lee 等人(2018)提出一种在频域中分析候选深度图以提高深度图精度的深度学习网络,使用裁剪率不同的裁剪图像得到多个深度预选图,在频域中学习各深度预选图的分量,并通过 2 维傅里叶逆变换产生最后的深度图。由于不同裁剪率生成的深度

预选图在全局和局部的表现不同,深度学习模型充分学习了全局分量和局部细节,使得深度估计的精度得到提高。

1.3 小结

典型的单图像输入训练模型的网络结构都比较简单。FCN 网络提出前都使用卷积层加全连接层来产生深度值,FCN 网络提出后深度估计模型大多采用了编解码结构。与卷积神经网络拟合深度图的方法相比,虽然深度图优化的方法预测深度更精确,但在训练时多采用分步训练,网络需要更多迭代次数,时间复杂度更高。随着性能更好的新型骨干网络框架的提出,单图像训练模型更倾向于采用卷积神经网络拟合的方法。典型单图像训练模型的概要如表 1 所示。

表 1 典型单图像训练模型概要总结

Table 1 Summary of typical single image training models

文献	模型结构	主要贡献
Eigen 等人(2014)	卷积神经网络	首次使用深度学习模型估计深度图
Laina 等人(2016)	全卷积网络	采用全卷积网络和提出新型上采样模块
Hu 等人(2019)	全卷积网络	使用多尺度特征信息提高深度估计性能
Liu 等人(2015)	卷积神经网络	采用连续条件随机场优化深度图
Xu 等人(2017)	全卷积网络	级联的连续条件随机场逐级优化深度图
Li 等人(2017)	卷积神经网络	首次根据图像的梯度信息优化深度估计
Lee 等人(2018)	卷积神经网络	首次在频域中优化深度估计

2 多图像输入训练模型

多图像输入训练模型分为使用立体图像对训练的方法和使用图像序列训练的方法。以立体图像对作为训练数据的单目深度估计模型通过视差图来学习场景的深度,而以图像序列为训练数据的单目深度估计模型,需要结合学习深度估计和相机姿态,通过场景重建来优化深度估计。多图像输入单目深度估计与立体视觉有显著不同,在测试时,单目深度估计只需单幅图像,而立体视觉还需要立体图像对。

2.1 立体图像对训练模型

Mayer 等人(2016)在 FlowNet(Dosovitskiy 等,2015)的基础上,提出了用于视差估计的 DispNet 模型,在 FlowNet 的解码端增加了卷积层和上采样层,获得了更平滑、分辨率更高的视差图。Godard 等人(2017)将视差法应用单目深度估计,网络结构如图 1 所示。

网络训练时,将左视图 $\boldsymbol{I}^{\mathrm{l}}$ 输入卷积网络提取特征,获得从左到右和从右到左的两幅视差图 $\boldsymbol{d}^{\mathrm{r}}$ 和 $\boldsymbol{d}^{\mathrm{l}}$。右视差图 $\boldsymbol{d}^{\mathrm{r}}$ 经过双线性上采样后与左视图 $\boldsymbol{I}^{\mathrm{l}}$ 结合重建右视图 $\tilde{\boldsymbol{I}}^{\mathrm{r}}$,左视差图 $\boldsymbol{d}^{\mathrm{l}}$ 与右视图 $\boldsymbol{I}^{\mathrm{r}}$ 结合重建左视图 $\tilde{\boldsymbol{I}}^{\mathrm{l}}$,同时视差图 $\boldsymbol{d}^{\mathrm{r}}$ 和 $\boldsymbol{d}^{\mathrm{l}}$ 生成相对应的反向

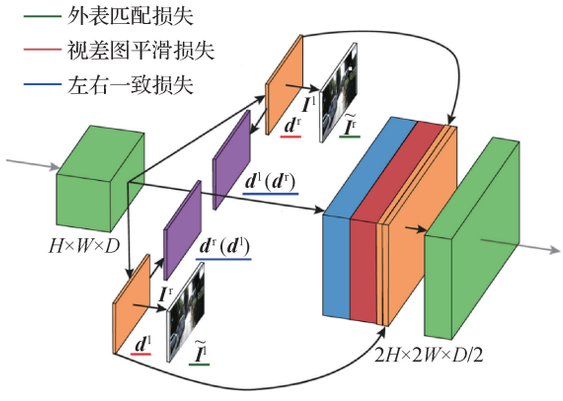


图 1 MonoDepth 网络结构示意图(Godard 等,2017)

Fig. 1 The framework of MonoDepth(Godard et al. ,2017)

视差图 $\boldsymbol{d}^{\mathrm{l}}(\boldsymbol{d}^{\mathrm{r}})$ 和 $\boldsymbol{d}^{\mathrm{r}}(\boldsymbol{d}^{\mathrm{l}})$。上采样后的左右视差图和左视图特征拼接后生成左视图的深度图。损失函数由外表匹配损失、视差图平滑损失和左右视差一致损失组成。其中,外表匹配损失衡量输入视图 $\boldsymbol{I}^{\mathrm{l}}$ 和 $\boldsymbol{I}^{\mathrm{r}}$ 与对应重建视图 $\tilde{\boldsymbol{I}}^{\mathrm{l}}$ 和 $\tilde{\boldsymbol{I}}^{\mathrm{r}}$ 的一致性,视差图平滑损失使视差图 $\boldsymbol{d}^{\mathrm{r}}$ 和 $\boldsymbol{d}^{\mathrm{l}}$ 平滑,左右视差一致损失使视差图 $\boldsymbol{d}^{\mathrm{r}}$ 和 $\boldsymbol{d}^{\mathrm{l}}$ 与对应的视差图 $\boldsymbol{d}^{\mathrm{r}}(\boldsymbol{d}^{\mathrm{l}})$ 和 $\boldsymbol{d}^{\mathrm{l}}(\boldsymbol{d}^{\mathrm{r}})$ 一致。网络训练时以左右视图作为输入计算损失和优化网络参数,测试时仅需输入左视图估计左视图的深度图。

基于视差估计深度的网络模型,其效果严重依

赖于视差估计结果。所以一些研究者致力于视差估计模型的研究，GA-Net（guided aggregation net）（Zhang 等，2019）和 Bi3DNet（Badki 等，2020）等视差预测模型有效提高了深度估计的效果。基于 Godard 等人（2017）的网络模型和损失函数，孙蕴瀚等人（2020）通过 DenseNet（Huang 等，2017）提取特征后使用特征金字塔结构（Lin 等，2017）估计多尺度深度图，并在损失函数中使用多尺度损失，深度估计精度有所提高。在 Godard 等人（2017）的方法中，右视差图 $\boldsymbol{d}^{\mathrm{r}}$ 是左视图 $\boldsymbol{I}^{\mathrm{l}}$ 生成的，造成右视图 $\boldsymbol{I}^{\mathrm{r}}$ 与右视差图 $\boldsymbol{d}^{\mathrm{r}}$ 不匹配，从而影响了深度图的精度。Repala 和 Dubey（2019）使用双流网络并行处理左右视图，估计左右视差图，使左右视差图与对应视图匹配，得到了更精确的深度图。

2.2　图像序列训练模型

Kendall 等人（2015）提出了 PoseNet 网络模型，通过分析图像序列获得相机的运动轨迹，实现相机姿态估计。在此基础上，Zhou 等人（2017）提出了运动结构重建模型 SfMLearner（structure from motion learner），在假定没有运动物体和场景物体是理想反射以及相机参数已知的条件下，将深度估计与相机姿态估计结合，使用运动结构重建来学习图像中的深度信息。网络结构如图 2 所示，以图像序列作为输入，$\boldsymbol{I}_t$ 为目标图像，$\boldsymbol{I}_{t-1}$ 和 $\boldsymbol{I}_{t+1}$ 表示目标图像的前后帧图像，将它们输入相机姿态估计网络中，分别估计从 $\boldsymbol{I}_t$ 到 $\boldsymbol{I}_{t-1}$ 和从 $\boldsymbol{I}_t$ 到 $\boldsymbol{I}_{t+1}$ 的相机运动轨迹。深度估计网络估计输入图像 $\boldsymbol{I}_t$ 的深度图，然后图像 $\boldsymbol{I}_t$ 的深度图结合相机运动轨迹和相机参数将图像 $\boldsymbol{I}_t$ 中所有点的位置映射到 $\boldsymbol{I}_{t+1}$ 和 $\boldsymbol{I}_{t-1}$ 中，最后根据映射关系使用 $\boldsymbol{I}_{t+1}$ 和 $\boldsymbol{I}_{t-1}$ 中像素点的亮度值重建图像 $\boldsymbol{I}_t$。网络以重建后的图像和原图像间的误差作为损失函数，同时优化深度估计网络和姿态估计网络。训练

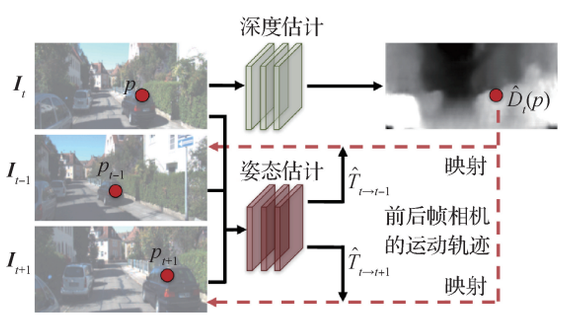


图 2　SfMLearner 网络结构示意图（Zhou 等，2017）

Fig. 2　The framework of SfMLearner（Zhou et al. ，2017）

时，网络以图像序列作为训练数据，预测时只需输入单幅图像估计图像的深度图。

考虑到图像序列间的亮度差异会影响深度估计的精确度，Yang 等人（2020）在 Zhou 等人（2017）的基础上，使用亮度转换函数和亮度不确定参数来减少匹配图像序列间的亮度误差，提高了深度估计的整体精度。一些研究者在 SfMLearner 基础上，通过优化姿态估计网络和深度估计网络提高深度估计模型的性能。梁欣凯等人（2019）在姿态估计网络中加入长短期记忆网络，并在深度估计网络中使用空洞特征金字塔网络（Chen 等，2018）提取特征，提升了深度估计的精度。同样的，Johnston 和 Carneiro（2020）使用自注意力（Wang 等，2018）和离散差异体积卷积（Kendall 等，2017）增强深度估计模块的性能。SfMLearner 模型隐含地假定了目标图像帧与其相邻帧间不存在遮挡或分离现象，为了克服此局限性，Godard 等人（2019）采用 U-Net（Ronneberger 等，2015）网络作为特征提取主干网络来估计图像的深度，重建损失项中使用最小重投影损失取代平均重投影损失降低遮挡物对深度估计的影响，从而提高了深度图的精确性。

为了处理存在运动物体的场景，Yin 和 Shi（2018）结合 FlowNet（Dosovitskiy 等，2015）提出了 GeoNet 模型，采用分治策略，根据场景中物体是否存在相对移动，将场景中的物体分为静态和动态物体，深度估计和姿态估计网络处理静态物体，光流网络计算场景中运动物体的光流。首先使用运动结构重建的思想实现场景重建，然后将重建后的场景图输入后续的光流网络中，通过前后向光流一致性优化深度估计网络、相机姿态估计网络和光流网络，很好地解决了运动结构重建无法处理运动物体的问题。同样地，为了解决运动物体问题，Zou 等人（2018）提出 DF-Net 模型，使用光流网络，通过前后向一致性生成掩膜层来明确静态与动态物体的区域。而 Ranjan 等人（2019）使用运动分割的方法明确静态与动态区域。这两种方法使得动态物体的区域更为明确，进一步提升了网络应对运动物体的能力。

相机将 3 维场景映射为 2 维图像时，2 维图像丢失了图像中物体尺寸与实际物体尺寸的比例关系，因此无法准确估计单幅图像中物体的长度。立体图像对中，由于两视图间的基线距离已知，根据基

线距离能够准确计算图像与实际物体尺寸的比例，所以使用立体图像训练的深度模型能够准确估计图像的深度值。与立体图像对不同，图像序列间缺少相机移动的实际度量尺度，度量尺度的缺失引起的尺度模糊问题会影响深度估计的效果。Zhan 等人(2018)考虑到图像序列中存在的尺度模糊问题，提出使用立体图像序列训练单目深度估计的网络模型。在立体图像序列中，左视图到右视图的相机运动轨迹已知，网络分别估计左视图序列和右视图序列中的相机运动轨迹，然后结合相机运动轨迹与深度图重建视图。考虑到重建图与原图间存在亮度差，而这种亮度差会影响深度估计的效果，因此重建图像的同时还重建了图像的特征，在图像重建损失的基础上增加了原图与重建图的特征匹配损失项，降低了原图与重建图像间亮度不同对网络性能的影响。使用立体图像序列降低了图像序列中尺度模糊的影响，同时立体图像序列在图像序列中增加了左右视图间的空间约束，提升了深度估计的精确性。

2.3 小结

视差法深度估计模型和运动结构重建模型作为多图像训练的典型模型，与单图像训练模型相比，网络结构复杂度明显提高。但是多图像训练模型在训练时不需要深度真值作为监督，极大降低了训练数据的获取难度，增加的训练数据提高了网络的泛化性能。相对于视差法深度估计模型，运动结构重建模型的结构更为复杂，网络包括深度估计和相机轨迹估计两个前端模块和光流网络等后端处理模块，训练过程更复杂，收敛速度更慢。典型多图像训练模型的概要如表 2 所示。

表 2 典型多图像训练模型概要总结

Table 2 Summary of typical multi-image training models

文献	模型结构	主要贡献
Godard 等人(2017)	全卷积网络	采用立体图像对训练单目深度估计模型
Repala 和 Dubey(2019)	全卷积网络	双流网络分别估计视差图缓解匹配
孙蕴瀚 等人(2020)	全卷积网络	采用多尺度特征信息和多尺度损失函数
Zhou 等人(2017)	全卷积网络	提出运动结构重建模型，使用图像序列训练单目深度估计模型
Johnston 和 Carneiro(2020)	全卷积网络	使用自注意力和离散差异体积卷积
Yin 和 Shi(2018)	全卷积网络	使用光流网络分别估计场景中的动态和静态物体的深度
Zhan 等人(2018)	全卷积网络	使用立体图像序列降低尺度模糊问题的影响

3 结合语义及数据集深度范围优化深度估计

单幅图像单目深度估计方法根据神经网络提取的图像特征拟合深度图，多图像单目深度估计通过视差图和运动结构重建计算图像的深度。为了进一步提高深度估计的精度，许多研究者结合语义标签和数据集深度范围等信息来获得更好的深度估计效果。

3.1 结合语义分割与深度估计

语义分割与深度估计是计算机视觉中的两项基本任务，都是对图像场景的理解，语义分割中物体边缘普遍存在深度跳跃而分割物体内部的深度大多连续，两项任务具有非常高的相关性，结合进行训练可以学习到额外的信息，从而提高深度估计的性能。而且两项任务可以共享一个主干网络，减少训练两项任务的参数量。因此，许多研究者通过结合语义分割与深度估计来提高深度估计效果。

Wang 等人(2015)首次提出在网络框架中联合利用语义和深度信息，旨在提升深度估计和语义分割的效果。网络结构如图 3 所示，全局卷积神经网络以整个图像作为输入，同时估计整幅图像的深度图和语义概率图。而区域卷积神经网络以 over-segmentation(Liu 等，2011)算法分割的图像块为输入，估计归一化相对深度模板和语义标签。然后利用像素点的深度与语义标签、图像块的深度与语义标签以及图像块与像素间的包含关系组成分层条件随机场。最后联合利用像素级深度信息和区域级深度信息，使用循环置信传播(loopy belief propagation，

LBP)算法估计每个图像块的语义标签,再使用最大后验概率算法估计图像块中每个像素点的语义标签。在获得语义标签后,通过 LBP 算法估计图像块中心深度和深度变化尺度,结合区域卷积网络估计的归一化模板,获得图像块的绝对深度。模型通过最小化双层 CRF 联合估计语义标签和深度图,充分利用了图像中语义与深度信息的关联性,精度得到了一定提升。

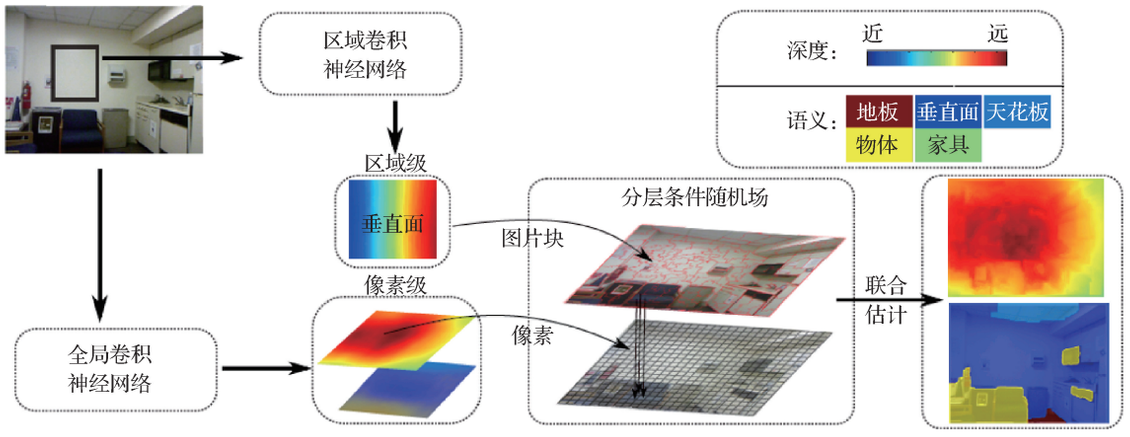


图 3　分层条件随机场网络结构示意图(Wang 等,2015)

Fig. 3　The framework of hierarchical conditional random field(Wang et al. ,2015)

Jiao 等人(2018)通过分析数据集中深度值的分布,提出使用语义信息和注意力驱动损失来优化深度图。图像在提取特征后分别输入深度估计和语义分割两个分支网络,两个分支网络间采用横向共享单元将同尺度下的特征相互拼接实现特征共享,同时每个分支中使用跳跃连接将各上采样层融合,得到最后的深度图和语义分割图。网络同时进行深度估计和语义分割,横向共享单元使两项任务的效果协同提高,有效提高了深度估计的精度。

Chen 等人(2019)提出采用立体视图同时预测图像的深度信息和语义分割,将左视图生成的语义分割图通过视差图转化为右视图的语义分割图,右视图生成的语义分割图通过同样的操作获得左视图的语义分割图,然后使用相对应语义分割图差值之和优化网络,相比 Godard 等人(2017)单纯使用立体图像对进行深度估计,精度获得了显著提升。

有些研究者将实例分割和全景分割等新型的图像分割任务与深度估计结合。Wang 等人(2020)将深度估计、语义分割和实例分割三者结合,提出了 SDC-DepthNet(semantic divide-and-conquer network)模型,采用分治策略先对图像进行语义分割和实例分割,再结合分割结果估计图像深度。语义分割图促使图像中相同语义类别的物体有相近的深度值;实例分割图明确界定了同种物体中不同个体的边界,使图像中同类物体的不同个体在深度图中存在明显的深度差。SDC-DepthNet 模型利用语义分割和实例分割的特点,提升了深度估计的精确度。

3.2　深度范围作为辅助信息

常用的深度估计数据集有 NYU depth 和 KITTI 数据集,NYU depth 数据集的深度范围是 0 ～10 m,KITTI 数据集的最大深度是 120 m。在数据集的有效深度范围已知条件下,一些研究者采用像素级分类的方法实现深度估计。

Cao 等人(2018)将数据集深度范围离散化后对每个像素点进行深度分类,实现深度估计。网络通过 ResNet152 提取图像特征,并控制输出特征图的通道数与深度范围离散个数相同。然后,特征经过 softmax 层后生成概率图,获得像素点的深度范围,以深度范围的中值作为像素点的深度值。最后,将每个像素点深度值的对数损失作为一元项,使用高斯核函数衡量输入图像中像素点对间的颜色和位置相似度,高斯核函数与对应像素点对的深度绝对差之积作为二元项来构造条件随机场,分类器估计深度范围后使用条件随机场优化深度图,最终得到的精确度比 Laina 等人(2016)的方法更高。

Fu 等人(2018)在数据集深度范围已知的条件下,将深度估计看做是一个分类问题,采用顺序回归的方法实现单目深度估计。网络以整幅图像作为输入,输入图像经过密集特征提取后输入场景理解模块,最后通过有序回归实现深度估计。场景理解模

块由空洞特征金字塔网络(Chen 等,2018)和全图编码器组成,空洞特征金字塔网络充分融合多尺度特征,特征经过编码进行有序回归。有序回归时需要多个间隔递增离散化标签,考虑到深度估计误差随着图像深度的增大而增大,在对数域中将数据集深度范围平均离散化作为有序回归的类别。网络通过有序回归获得每个像素点的深度区间后,将深度区间的中点作为像素点的深度值,取得了 2018 年 KITTI 深度估计的最优效果。

Su 和 Zhang(2020)同样将深度估计作为分类任务,提出了 SSE-Net(scale-sernantic exchange network)模型,使用与 Fu 等人(2018)相同的方法对数据集深度范围离散化。SSE-Net 网络以 ResNet101 为主干网络提取多尺度特征,多尺度特征输入信息交换网络充分融合,然后经过卷积产生类别概率图,最后使用软回归网络获得深度图。在软回归网络中,考虑到数据集深度范围严格离散化会对深度的连续性和模型处理歧义的性能产生影响,引入了尺度项和偏移项对生成的类别概率图进行修正。信息交换网络和软回归的使用有效提升了深度估计的效果。

与上述像素级分类方法不同,Lee 等人(2019)利用数据集深度范围提出新型的场景重建网络,以数据集深度范围作为辅助信息,提出使用局部平面指导层解决下采样过程中的信息丢失问题。输入图像通过 DenseNet(Huang 等,2017)提取特征,接着使用空洞特征金字塔网络(Chen 等,2018)充分融合特征图的上下文信息,然后通过局部平面指导层重建场景,最后结合多尺度的场景重建图生成深度图。局部平面指导层进行场景重建时,数据集深度范围与特征图结合,将特征图上每个位置的特征值转换为包含方位信息的特征向量。然后,根据射线平面交叉的原理得到场景重建图。通过数据集深度范围,将每个尺度下的特征图映射为反映场景信息的场景重建图,充分利用了各尺度的特征信息,深度估计的精度有很大提升。

3.3 小结

辅助信息的加入提高了深度估计的精确度,但是使用辅助信息会使网络更复杂,训练需要的时间更长。语义信息不仅使网络更复杂,而且增加了训练难度,一些结合语义分割的深度估计模型不得不采用分步训练的方法。与语义信息相比,在深度估计网络中加入深度范围更灵活,深度范围不仅应用于像素级分类深度估计模型,还可以加强网络对输入图片场景的理解能力,并且使用深度范围作为辅助信息的深度估计模型采用统一训练的方法,训练简单。典型辅助信息训练模型的概要如表 3 所示。

表 3 典型辅助信息训练模型概要总结

Table 3 Summary of typical auxiliary information training models

文献	模型结构	主要贡献
Wang 等人(2015)	卷积神经网络	语义分割与深度估计结合,并使用分层条件随机场
Jiao 等人(2018)	全卷积网络	语义分割与深度估计特征共享和新型注意力损失
Chen 等人(2019)	全卷积网络	在立体图像对中联合训练语义分割和深度估计
Wang 等人(2020)	全卷积网络	利用语义分割和实例分割信息提升深度估计精度
Cao 等人(2018)	卷积神经网络	平均离散深度范围,使用像素级分类生成深度图
Fu 等人(2018)	卷积神经网络	在对数域中平均离散深度范围并有序回归深度值
Lee 等人(2019)	全卷积网络	使用深度范围构建新型场景重建模型

4 评价指标及代表性方法对比

单目深度估计模型常用的性能评价指标有绝对值相对误差(*AR*)、平方相对误差(*SQ*)、对数平均误差(*ML*)、均方根误差(*RMS*)、对数均方根误差(*LRMS*)以及阈值准确率(f_{correct})(Zhuo 等,2015),计算方法为

$$AR = \frac{1}{N}\sum_{i=1}^{N}\frac{|D_i - D_i^*|}{D_i^*} \tag{1}$$

$$SQ = \frac{1}{N}\sum_{i=1}^{N}\frac{|D_i - D_i^*|^2}{D_i^*} \tag{2}$$

$$ML = \frac{1}{N}\sum_{i=1}^{N}|\lg D_i - \lg D_i^*| \tag{3}$$

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|D_i - D_i^*|^2} \tag{4}$$

$$LRMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|\lg D_i - \lg D_i^*|^2} \tag{5}$$

$$f_{\text{correct}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\max\left(\frac{D_i}{D_i^*},\frac{D_i^*}{D_i}\right) = \delta < T\right\| \tag{6}$$

$$\delta = \max\left(\frac{D_i}{D_i^*},\frac{D_i^*}{D_i}\right) \tag{7}$$

式中，N 表示图像的像素点个数，D_i 表示图中第 i 个像素点的预测深度值，D_i^* 表示对应像素点的真值，T 表示阈值，当比值 δ 小于阈值时认为是正确预测($\|\cdot\|$)，然后计算正确预测的像素点比例作为准确率，阈值的设定通常有 3 种。

表 4、表 5 和表 6 分别比较了单图像输入单目深度估计模型、多图像输入单目深度估计模型和利用辅助信息优化深度模型在 NYU 数据集和 KITTI 数据集上的测试性能。表中数据取自各方法的测试结果，KITTI 数据集的深度范围采用 0 ～80 m，表 5 中训练网络使用的数据类型 M 和 S 分别表示图像序列和立体图像对。

从表 4 可以看出，Eigen 等人(2014)方法与 Hu 等人(2019)方法的结果相比，单图像单目深度估计

表 4　部分单图像训练模型的评估函数值

Table 4　Quantitative evaluation of some single image training models

方法	数据集	误差					不同阈值下的准确率		
		AR	*SQ*	*ML*	*LRMS*	*RMS*	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen 等人(2014)	NYU	0.215	0.212	–	0.285	0.907	0.611	0.887	0.971
Liu 等人(2015)	NYU	0.23	–	0.095	–	0.824	0.614	0.883	0.971
Laina 等人(2016)	NYU	0.127	–	0.055	0.195	0.573	0.811	0.953	0.988
Xu 等人(2017)	NYU	0.121	–	0.052	–	0.586	0.811	0.954	0.987
Li 等人(2017)	NYU	0.143	–	0.063	–	0.635	0.788	0.958	0.991
Lee 等人(2018)	NYU	0.139	0.096	–	0.193	0.572	0.815	0.963	0.991
Hu 等人(2019)	NYU	0.115	–	0.050	–	0.530	0.866	0.975	0.993

注：“–”表示原文献未提供数据。

表 5　部分多图像训练模型的评估函数值

Table 5　Quantitative evaluation of some multi-image training models

方法	数据集	数据类型	误差					不同阈值下的准确率		
			AR	*SQ*	*ML*	*LRMS*	*RMS*	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Godard 等人(2017)	KITTI	S	0.148	1.344	–	0.247	5.927	0.803	0.922	0.964
Repala 和 Dubey(2019)	KITTI	S	0.122	1.306	–	0.213	6.069	0.841	0.939	0.976
孙蕴瀚 等人(2020)	KITTI	S	0.099	0.965	–	0.184	5.333	0.875	0.957	0.985
Zhou 等人(2017)	KITTI	M	0.208	1.768	–	0.283	6.856	0.678	0.885	0.957
Yin 和 Shi(2018)	KITTI	M	0.155	1.296	–	0.233	5.857	0.793	0.931	0.973
Zou 等人(2018)	KITTI	M	0.15	1.124	–	0.223	5.507	0.806	0.933	0.973
Ranjan 等人(2019)	KITTI	M	0.14	1.07	–	0.217	5.326	0.826	0.941	0.975
Godard 等人(2019)	KITTI	M	0.115	0.903	–	0.193	4.863	0.877	0.959	0.981
Johnston 和 Carneiro(2020)	KITTI	M	0.106	0.861	–	0.185	4.699	0.889	0.962	0.982
Zhan 等人(2018)	KITTI	MS	0.135	1.132	–	0.229	5.585	0.82	0.933	0.971

注：“–”表示原文献未提供数据。

表 6　部分辅助信息优化深度模型的评估函数值

Table 6　Quantitative evaluation of some auxiliary information training models

方法	数据集	误差					不同阈值下的准确率		
		AR	*SQ*	*ML*	*LRMS*	*RMS*	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Wang 等人(2015)	NYU	0.220	–	0.094	0.262	0.745	0.605	0.890	0.970
Jiao 等人(2018)	NYU	0.098	–	0.040	0.125	0.329	0.917	0.983	0.996
Lee 等人(2019)	NYU	0.110	0.066	0.047	0.142	0.392	0.885	0.978	0.994
Wang 等人(2020)	NYU	0.128	–	–	0.174	0.497	0.845	0.966	0.990
Cao 等人(2018)	KITTI	0.115	–	–	0.198	4.712	0.887	0.963	0.982
Fu 等人(2018)	KITTI	0.072	0.307	–	0.120	2.727	0.932	0.984	0.994
Chen 等人(2019)	KITTI	0.112	0.673	–	0.198	3.871	0.852	0.951	0.980
Su 和 Zhang(2020)	KITTI	0.102	–	–	0.183	4.051	0.901	0.973	0.994

注：“–”表示原文献未提供数据。

模型的性能在 5 年时间内提升了近一倍。多图像单目深度估计模型是 2017 年出现的，虽然模型出现较晚，但发展势头强劲。从表 5 可以看出，Johnston 和 Carneiro(2020)方法与 Zhou 等人(2017)方法的结果相比，预测误差明显降低。从表 6 的数据可以发现，通过利用辅助信息优化深度估计后，性能相比表 4 和表 5 中模型有较大提升。表 6 中 Chen 等人(2019)的 *RMS* 与表 5 中 Repala 和 Dubey(2019)模型相比，降低了 2.198。Wang 等人(2015)模型与表 4 中 Eigen 等人(2014)模型相比，*RMS* 降低了 0.162。

5　结　语

深度估计的应用领域对单目深度估计的精确度、估计速度和泛化性等方面有很高的要求。本文根据网络输入数据的类型对单目深度估计模型进行了分类综述，分别介绍了单图像输入、多图像输入和结合辅助信息 3 类方法中的典型网络结构。以单幅图像作为训练数据的模型具有网络结构简单的特点，但泛化性能较差。采用多图像训练的深度估计网络有更强的泛化性，但网络的参数量大、网络收敛速度慢、训练耗时长。引入辅助信息的深度估计网络的深度估计精度得到了进一步提升，但辅助信息的引入会造成网络结构复杂、收敛速度慢等问题。单目深度估计研究还存在许多的难题和挑战，未来的研究趋势包含如下两个方面：

1)多图像训练模型成为研究热点。虽然采用单幅图像作为训练数据的网络模型具有模型简单的特点，但是训练所需的精确密集深度信息难于获得，训练数据的场景种类和样本数量非常有限，导致单图像训练单目深度估计模型的泛化性较差。而多图像训练单目深度估计网络模型在训练时不需要精确的深度信息，训练数据的成本更低，所以可以获得更大量的训练数据，从而使得模型的泛化性能得到提高，所以多图像输入单目深度估计逐渐成为单目深度估计的研究热点。

2)单目深度估计中加入约束信息。由于单目深度估计的病态性，仅从彩色图像中获取的场景信息很难提高深度估计的精确度。在这种情况下，在单目深度估计网络中加入辅助信息是提升单目深度估计的有效方法。语义标签与深度估计联系紧密，常用来提升深度估计的效果。特定领域约束信息的加入在解决特定问题中也能取得不错效果，如 Chen 等人(2016)在野外深度估计任务中加入相对深度信息提高深度估计的精度。

参考文献(References)

Badki A, Troccoli A, Kim K, Kautz J, Sen P and Gallo O. 2020. Bi3D: stereo depth estimation via binary classifications//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1597-1605 [DOI: 10.1109/CVPR42600.2020.00167]

Cadena C, Carlone L, Carrillo H, Latif Y, Scaramuzza D, Neira J, Reid

I and Leonard J J. 2016. Past, present, and future of simultaneous localization and mapping: toward the robust-perception age. IEEE Transactions on Robotics, 32(6): 1309-1332 [DOI: 10.1109/TRO.2016.2624-754]

Cao Y Z H, Wu Z F and Shen C H. 2018. Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE Transactions on Circuits and Systems for Video Technology, 28(11): 3174-3182 [DOI: 10.1109/TCSVT.2017.2740321]

Chen L C, Zhu Y K, Papandreou G, Schroff F and Adam H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 833-851 [DOI: 10.1007/978-3-030-01234-2_49]

Chen P Y, Liu A H, Liu Y C and Wang Y C F. 2019. Towards scene understanding: unsupervised monocular depth estimation with semantic-aware representation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2619-2627 [DOI: 10.1109/CVPR.2019.00273]

Chen W F, Fu Z, Yang D W and Deng J. 2016. Single-image depth perception in the wild//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc.: 730-738

Dosovitskiy A, Fischer P, Ilg E, Häusser P, Hazirbas C, Golkov V, Smagt P V D, Cremers D and Brox T. 2015. FlowNet: learning optical flow with convolutional networks//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2758-2766 [DOI: 10.1109/ICCV.2015.316]

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: MIT Press: 2366-2374

Fu H, Gong M M, Wang C H, Batmanghelich K and Tao D C. 2018. Deep ordinal regression network for monocular depth estimation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2002-2011 [DOI: 10.1109/CVPR.2018.00214]

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 J. 2019. Digging into self-supervised monocular depth estimation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): 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]

Hu J, Shen L, Albanie S, Sun G and Wu E H. 2020. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(8): 2011-2023 [DOI: 10.1109/TPAMI.2019.2913372]

Hu J J, Ozay M, Zhang Y and Okatani T. 2019. Revisiting single image depth estimation: toward higher resolution maps with accurate object boundaries//Proceedings of 2019 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 1043-1051 [DOI: 10.1109/WACV.2019.00116]

Hu Y F, Qu T, Liu J, Shi Z Q, Zhu B, Cao D P and Chen H. 2019. Human-machine cooperative control of intelligent vehicle: recent developments and future perspectives. Acta Automatica Sinica, 45(7): 1261-1280 (胡云峰, 曲婷, 刘俊, 施竹清, 朱冰, 曹东璞, 陈虹. 2019. 智能汽车人机协同控制的研究现状与展望. 自动化学报, 45(7): 1261-1280) [DOI: 10.16383/j.aas.c180136]

Huang G, Liu Z, van der Maaten L and Weinberger K Q. 2017. Densely connected convolutional networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2261-2269 [DOI: 10.1109/CVPR.2017.243]

Jiao J B, Cao Y, Song Y B and Lau R. 2018. Look deeper into depth: monocular depth estimation with semantic booster and attention-driven loss//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 55-71 [DOI: 10.1007/978-3-030-01267-0_4]

Johnston A and Carneiro G. 2020. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4755-4764 [DOI: 10.1109/CVPR42-600.2020.00481]

Kendall A, Grimes M and Cipolla R. 2015. PoseNet: a convolutional network for real-time 6-DOF camera relocalization//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2938-2946 [DOI: 10.1109/ICCV.2015.336]

Kendall A, Martirosyan H, Dasgupta S, Henry P, Kennedy R, Bachrach A and Bry A. 2017. End-to-end learning of geometry and context for

deep stereo regression//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 66-75 [DOI: 10.1109/ICCV.2017.17]

Khan F, Salahuddin S and Javidnia H. 2020. Deep learning-based monocular depth estimation methods — a state-of-the-art review. Sensors, 20(8): 2272 [DOI: 10.3390/s20-082272]

Krizhevsky A, Sutskever I and Hinton G E. 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90 [DOI: 10.1145/3065386]

Lafferty J, McCallum A and Pereira F C N. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data//Proceedings of the 18th International Conference on Machine Learning. Williamstown, USA: Morgan Kaufmann: 282-289

Laina I, Rupprecht C, Belagiannis V, Tombari F and Navab N. 2016. Deeper depth prediction with fully convolutional residual networks//Proceedings of the 4th International Conference on 3D Vision. Stanford, USA: IEEE: 239-248 [DOI: 10.1109/3DV.2016.32]

Lee J H, Han M K, Ko D W and Suh I H. 2019. From big to small: multi-scale local planar guidance for monocular depth estimation [EB/OL]. [2021-02-21]. https://arxiv.org/pdf/1907.10326.pdf

Lee J H, Heo M, Kim K R and Kim C S. 2018. Single-image depth estimation based on fourier domain analysis//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 330-339 [DOI: 10.1109/CVPR.2018.00042]

Li J, Klein R and Yao A. 2017. A two-streamed network for estimating fine-scaled depth maps from single RGB images//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 3392-3400 [DOI: 10.1109/ICCV.2017.365]

Liang X K, Song C and Zhao J J. 2019. Depth estimation technique of sequence image based on deep learning. Infrared and Laser Engineering, 48(S2): #S226002 (梁欣凯, 宋闯, 赵佳佳. 2019. 基于深度学习的序列图像深度估计技术. 红外与激光工程, 48(S2): #S226002)

Lin T Y, Dollár P, Girshick R, He K M, Hariharan B and Belongie S. 2017. Feature pyramid networks for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 936-944 [DOI: 10.1109/CVPR.2017.106]

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]

Liu F Y, Shen C H, Lin G S and Reid I. 2016. Learning depth from single monocular images using deep convolutional neural fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(10): 2024-2039 [DOI: 10.110-9/TPAMI.2015.2505283]

Liu M Y, Tuzel O, Ramalingam S and Chellappa R. 2011. Entropy rate superpixel segmentation//Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA: IEEE: 2097-2104 [DOI: 10.1109/CVPR.2011.5995323]

Mayer N, Ilg E, Häusser P, Fischer P, Cremers D, Dosovitskiy A and Brox T. 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4040-4048 [DOI: 10.1109/CVPR.2016.438]

Owen A B. 2007. A robust hybrid of lasso and ridge regression. Prediction and Discovery, 443: 59-71

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]

Repala V K and Dubey S R. 2019. Dual CNN models for unsupervised monocular depth estimation//Proceedings of the 8th International Conference on Pattern Recognition and Machine Intelligence. Tezpur, India: Springer: 209-217 [DOI: 10.1007/978-3-030-34869-4_23]

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]

Saxena A, Chung S H and Ng A Y. 2005. Learning depth from single monocular images//Proceedings of the 18th International Conference on Neural Information Proce-ssing Systems. Vancouver, British Columbia, Canada: MIT Press: 1161-1168

Shelhamer E, Long J and Darrell T. 2015. Fully convolutional networks for semantic segmentation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]

Shi W Z, Caballero J, Huszár F, Totz J, Aitken A P, Bishop R, Rueckert D and Wang Z H. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 1874-1883 [DOI: 10.1109/CVPR.2016.207]

Silberman N, Hoiem D, Kohli P and Fergus R. 2012. Indoor segmentation and support inference from RGBD images//Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer: 746-760 [DOI: 10.1007/978-3-642-33715-4_54]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2021-02-21]. https://arxiv.org/pdf/1409.1556.pdf

Su W and Zhang H F. 2020. Soft regression of monocular depth using

scale-semantic exchange network. IEEE Access, 8: 114930-114939 [DOI: 10.1109/ACCESS.2020.3003466]

Sun Y H, Shi J L and Sun Z X. 2020. Estimating depth from single image using unsupervised convolutional network. Journal of Computer-Aided Design and Computer Graphics, 32(4): 643-651 (孙蕴瀚, 史金龙, 孙正兴. 2020. 利用自监督卷积网络估计单图像深度信息. 计算机辅助设计与图形学学报, 32(4): 643-651) [DOI: 10.3724/SP.J.1089.2020.1782]

Wang L J, Zhang J M, Wang O, Lin Z and Lu H C. 2020. SDC-Depth: semantic divide-and-conquer network for monocular depth estimation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 538-547 [DOI: 10.1109/CVPR42600.2020.00062]

Wang P, Shen X H, Lin Z, Cohen S, Price B and Yuille A. 2015. Towards unified depth and semantic prediction from a single image//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 2800-2809 [DOI: 10.1109/CVPR.2015.7298897]

Wang X L, Girshick R, Gupta A and He K M. 2018. Non-local neural networks [EB/OL]. [2021-02-21]. https://arxiv.org/pdf/1711.07971.pdf

Xie Z, Ma H L, Wu K W, Gao Y and Sun Y X. 2020. Sampling aggregate network for scene depth estimation. Acta Automatica Sinica, 46(3): 600-612 (谢昭, 马海龙, 吴克伟, 高扬, 孙永宣. 2020. 基于采样汇集网络的场景深度估计. 自动化学报, 46(3): 600-612) [DOI: 10.16383/j.aas.c180430]

Xu D, Ricci E, Ouyang W L, Wang X G and Sebe N. 2017. Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 161-169 [DOI: 10.1109/CVPR.2017.25]

Yang N, von Stumberg L, Wang R and Cremers D. 2020. D3VO: deep depth, deep pose and deep uncertainty for monocular visual odometry//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1278-1289 [DOI: 10.1109/CVPR42600.2020.00136]

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]

Yu F and Koltun V. 2016. Multi-scale context aggregation by dilated convolutions [EB/OL]. [2021-02-21]. https://arxiv.org/pdf/1511.07122.pdf

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]

Zhang F H, Prisacariu V, Yang R G and Torr P H S. 2019. GA-Net: guided aggregation net for end-to-end stereo matching//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 185-194 [DOI: 10.1109/CVPR.2019.00027]

Zhao C Q, Sun Q Y, Zhang C Z, Tang Y and Qian F. 2020. Monocular depth estimation based on deep learning: an overview. Science China Technological Sciences, 63(9): 1612-1627 [DOI: 10.1007/s11431-020-1582-8]

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: 10.1109/CVPR.2017.700]

Zhuo W, Salzmann M, He X M and Liu M M. 2015. Indoor scene structure analysis for single image depth estimation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 614-622 [DOI: 10.1109/CVPR.2015.7298660]

Zou Y L, Luo Z L and Huang J B. 2018. DF-Net: unsupervised joint learning of depth and flow using cross-task consistency//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 38-55 [DOI: 10.1007/978-3-030-01228-1_3]

Zwald L and Lambert-Lacroix S. 2012. The BerHu penalty and the grouped effect [EB/OL]. [2021-02-21]. https://arxiv.org/pdf/1207.6868.pdf

作者简介

罗会兰，1974 年生，女，教授，硕士生导师，主要研究方向为机器学习、图像处理与模式识别。
E-mail: luohuilan@sina.com

周逸风，男，硕士研究生，主要研究方向为计算机视觉。
E-mail: colinzhou2010@foxmail.com