基于深度学习的单目深度估计方法综述

江俊君 李震宇 刘贤明
(哈尔滨工业大学计算学部 哈尔滨 150001)

摘 要 深度估计是一种从单张或者多张图像预测场景深度信息的技术,是计算机视觉领域非常热门的研究方向,在三维重建、场景理解、环境感知等任务中起到了关键作用.当前深度估计技术可以分为多目深度估计和单目深度估计.因为单目摄像头具有成本低、设备较普及、图像获取方便等优点,与多目深度估计技术相比,从单目图像估计深度信息是当前更为热门和更具挑战的技术.近年来,随着深度学习的迅速发展,基于深度学习的单目深度估计方法被广泛研究.本文对基于深度估计的单目深度估计方法进行综述,首先给出单目深度估计问题的定义、介绍常用于训练的数据集与模型评价指标,然后根据不同的训练方式对国内外相关技术进行分析总结,将现有方法分为基于监督学习、无监督学习和半监督学习三大类,对每种类型方法的产生思路、优缺点进行详细分析,最后梳理、总结该技术的发展趋势与关键技术.

关键词 单目深度估计;深度学习;计算机视觉;场景理解
中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2022.01276

Deep Learning Based Monocular Depth Estimation: A Survey

JIANG Jun-Jun LI Zhen-Yu LIU Xian-Ming
(*Faculty of Computing, Harbin Institute of Technology, Harbin* 150001)

Abstract Depth estimation plays a crucial role in scene perception and understanding, which aims to predict distances between camera and real-world pixels from single or multiple images. It is a popular research field in computer vision, applying as an important step in many practical tasks such as 3D reconstruction. In recent years, depth estimation methods have drawn increasing attention and intensive research as a low-level vision task. Traditional methods use lidar to obtain high-precision depth information but cannot be widely used in practice due to the high cost of obtaining dense and accurate depth maps. In contrast, image-based depth estimation methods directly estimate depth based on input RGB images without expensive equipment, yielding more favor in applications. Specifically, image-based depth estimation methods can be divided into the multi-view and monocular estimation according to the number of required input pictures. Since the monocular camera has the advantages of low cost, more standard equipment, and convenient image acquisitions, compared with multi-view depth estimation methods, estimating depth information from monocular images is currently a more popular research field. With the rapid development of deep learning, monocular depth estimation based on deep neural networks has been widely studied, and many excellent methods have been proposed. This paper provides an overview of thestate-of-the-art on monocular depth estimation methods. First, the definition of monocular depth estimation, commonly used datasets, evaluation metrics, and applications are introduced. Then, we review representative methods according to different training manners:

收稿日期:2021-08-18;在线发布日期:2022-03-17.本课题得到国家自然科学基金面上项目(61971165)、国家自然科学基金优秀青年项目(61922027)资助.**江俊君**,博士,教授,主要研究领域为计算机视觉、图像处理、遥感影像处理与分析.E-mail:jiangjunjun@hit.edu.cn.**李震宇**,硕士研究生,主要研究方向为计算机视觉、图像处理.**刘贤明**,博士,教授,主要研究领域为人工智能、图像处理、计算成像.

supervised, unsupervised and semi-supervised. The existing methods based on different learning manners are divided into several types, respectively. We summarize supervised methods and classify them into enhancing framework, introducing auxiliary information, improving loss function, classification-based methods, applying conditional random field, applying generative adversarial network, and methods based on partial depth labels. As for unsupervised methods, we classify them into models trained by image pairs and monocular videos. Improving strategies are divided into mask-based methods, applying visual odometry, applying generative adversarial network, and methods towards fast and light runtime performance. In terms of semi-supervised methods, they are similarly trained on image pairs or monocular videos. The proposed methods are classified into methods that apply a generative adversarial network and introduce semantic information, respectively. The ideas, advantages, and disadvantages of each type of methods are analyzed in detail. Finally, we sort out trends of future development and key technologies of monocular depth estimation methods based on deep learning. We aim to inspire readers to make further breakthroughs based on existing research summarized in this paper. Compared with the previous overviews, our paper mainly focuses on deep learning methods. We elaborated methods more comprehensively, point out ideas and characteristics of each method and use a more fine-grained and more accurate classification criterion, which is more helpful for readers to understand the overall research progress of depth estimation methods. Moreover, we provide detailed comparison results among all kinds of methods for quick investigation in tables. We also discuss the technical difficulties and development trends in monocular depth estimation, hoping to further inspire readers.

Keywords　monocular depth estimation; deep learning; computer vision; scene understanding

1　引　言

深度估计是场景感知中重要的一环，测量与物体间的距离是所有生物赖以生存的技能. 在计算机视觉领域中，深度估计也扮演着类似的角色，是许多高层任务的基石，其结果可广泛应用于三维立体重建[1]、障碍物检测[2]、视觉导航[3]等方向. 近年来，深度估计方法作为底层视觉任务受到了广泛关注和深入研究. 传统方法使用激光雷达获取深度信息，但是由于获得稠密而准确的深度图的成本过高，无法得到广泛应用. 相比之下，基于图像的深度估计方法根据输入的 RGB 图像直接估计场景深度信息，无须价格高昂的设备，应用空间更为广阔. 基于图像的深度估计方法根据要求输入图片的多少可分为多目深度估计与单目深度估计. 其中，多目深度估计方法根据观测得到的多张图像，估计出深度信息，比较经典的有：从运动中恢复(Structure from Motion，SFM)、使用单目摄像机捕获的图像序列估计深度[4]、多视图重建(Multiview System，MVS)[5-6]、三角测量法(Triangulation)[7-8]等. 这些方法大多需要成对图像或图像序列作为输入、要求相机信息已知，对输入有较强的限制且预测结果受特征选择和匹配的影响较大. 相比之下，仅通过单张图像估计深度的单目深度估计方法更为灵活. 但由于缺乏深度线索，它是一个不适定性问题(Ill-posed Problem)[9]，早期的方法从不同角度提取深度线索进行深度估计，常见的手段有：从阴影中恢复(Shape from Shading)[10]、从对焦或离焦中获取深度(Depth from Focus or Depth from Defocus)[11-12]等. 表 1 总结了早期的深度估计方法.

随着深度学习的迅速发展，深度神经网络在图像分类[13]、自然语言处理[13]、语音识别[14]、医疗影像处理[15]等方向展现了极佳的性能. 鉴于深度神经网络强大的特征提取能力，学者们将深度学习技术引入深度估计领域，希望借助神经网络充分提取图像的深度特征并给出更准确的深度估计结果. 近年来，基于深度学习的单目深度估计方法发展迅速，根据训练方式的不同可以划分为有监督方法、无监督方法和半监督方法.

表 1 早期方法总结

方法	类型	主要思想
双目视觉法 (Binocular Stereo Vision,BSV)	多目	模拟人类利用双眼感知图像视差而获取深度信息,依赖于图片像素匹配,可以使用模板比对或者对极几何法等匹配方法
立体视觉法 (Multiple View Stereo,MVS)	多目	由双目视觉法发展而来,将多个相机设置于视点,或用单目相机在多个不同的视点拍摄图像,增加了算法的稳健性
从运动中恢复形状 (Structure from Motion,SFM)	多目	利用二维图像序列间的特征点匹配关系,估计相机的运动参数与场景点云信息
三角测量法 (Triangulation)	多目	借由测量目标点与固定基准线的已知端点的角度,测量目标距离
从阴影中恢复形状 (Shape from Shading,SFS)	单目	利用图像中物体表面的明暗变化来恢复表面形状
纹理恢复形状法 (Shape from Texture,SFT)	单目	通过纹理在经过透视等变形后在图像上的变化来逆向计算深度数据
从对焦中获取深度 (Depth from Focusing, DFF)	单目	利用图像中像素的聚焦信息结合摄像机的参数计算图像深度
从离焦中获取深度 (Depth from Defocusing,DFD)	单目	利用图像中像素的离焦信息结合摄像机的参数计算图像深度

本文对国内外近五年来基于深度学习的单目深度估计方法进行综述,如图 1,本文安排如下:第 2 节首先介绍深度估计问题的定义、常用数据集、模型评价指标和相关应用;第 3、4、5 节详细分析近年来基于深度学习的深度估计方法,将有监督改进方法分为改进网络结构的方法、引入辅助信息的方法、改进损失函数的方法、基于分类的方法、使用条件随机场的方法、使用生成式对抗网络的方法和基于部分深度信息的方法七种;将无监督方法分为使用图像对训练的方法、使用视频训练的方法两类,将改进方法分为基于可解释性掩膜的方法、基于直接视觉测距的方法、引入辅助信息的方法、使用生成式对抗网络的方法和面向实时与轻量级的方法五种;将半监督方法分为使用图像对训练的方法、使用视频训练的方法两类,将改进方法分为使用生成式对抗网络的方法、引入辅助信息的方法两种. 第 6 节讨论了该领域存在的难题,总结了未来可能的发展方向.

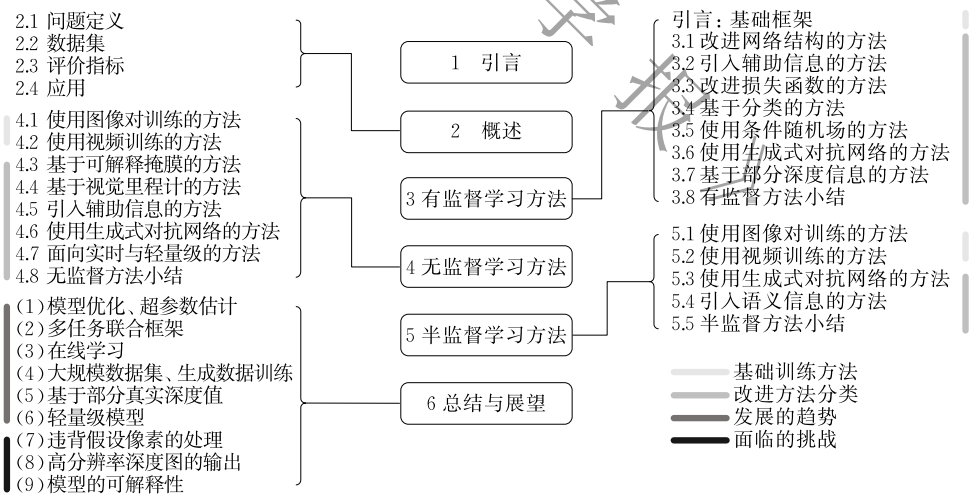


图 1 文章架构

本篇综述旨在归纳总结基于深度学习的单目深度估计方法,启发读者在现有研究基础上做出进一步的突破. 相比于 Bi 等人[16]的综述主要对参数学习、非参数学习和部分有监督深度估计方法进行介绍,我们的综述主要关注于深度学习方法,除有监督方法外,还对无监督、半监督方法进行了详细的总结. Bhoi[17]只总结了五个最有影响力的基于深度学习的单目深度估计方法,而我们更全面地对更多的方法进行了阐述,指出了每种方法的产生思路和特点. 此外,他们介绍的方法比较老旧,缺乏目前最新的算法介绍. 与综述[18-20]相比,我们使用了更细粒度、更准确的分类方法,更有助于读者了解深度估计方法的整体框架. 最后,我们还进一步讨论了单目深度估计领域存在的技术难题和发展趋势,希望进一步给读者启发.

2 概　述

2.1 问题定义

基于监督学习的单目深度估计任务可以描述为：利用一个包含大量二维 RGB 图像和深度图像对的数据集训练一个模型，该模型可以接受输入的 RGB 图像，输出其对应的深度图. 其中，深度图中每个像素的值表示输入 RGB 图像在对应位置处的深度. 此过程可以抽象为：设 I 为输入的二维 RGB 图像，D 为输出的灰度图. 给定一个训练集 $T=(I_i,D_i)_{i=1}^{M}$；$I_i\in I$，$D_i\in D$，使用该训练集训练模型，使其学习到一个映射 $\varphi: I\rightarrow D$，将输入 RGB 图像映射成预测的深度图.

有监督方法需要大量输入图像与深度图数据对来驱动模型的训练，但是由于真实的深度图获取困难，此类方法的成本高昂. 自然地，无监督方法因为在训练时不需要真实深度图，成为了此领域的研究热点. 但是因为此类方法缺乏强有力的监督信号，模型精度难以超越有监督方法. 半监督方法介于两者之间. 但是由于如今有监督和无监督方法仍有较大的发展空间，相比之下，半监督相关的研究内容较少. 我们仍然对其进行分析与分类，以保证文章的完整性.

2.2 数据集

在深度估计问题中，由于室内外深度范围差异较大，可根据不同的场景构建不同的数据集. 在对模型进行训练时，需依据问题的特性选择不同特点的数据集，表 2 总结了常见数据集的基本信息.

表 2　常见数据集汇总

名称	是否带有深度标签	图片分辨率	可用图片数量	适用场景	采集工具
NYU depth[21]	√	640×480	1449	室内	RGB-D 相机
Make3D[22]	√	2272×1704	534	室外	LASER
KITTI[23]	√	375×1242	93000+	自动驾驶/室外	LIDAR
Cityscapes[24]	√	1024×2048	22973	城市	相机+人工标注
MegaDepth[25]	√	1600×1600	130000+	地标/建筑	SFM+MVS
Scene Flow[26]	√	960×540	39000+	综合	人工生成
DIODE[27]	√	1024×768	27858	综合	faro
Holopix[28]	×	720p or 360p	50000+	综合	RED Hydrogen One 等
Mirror3D[29]	√	—	7011	室内	人工标注

对于室内场景，比较常用的数据集是由纽约大学发布的 NYU Depth 数据集[21]. 如图 2(a)所示，该数据集使用 RGB-D 摄像机获取室内场景的单目视频与对应的真实深度信息，含有 464 个室内场景照片，其中 249 个作为训练集，215 个作为测试集，是有监督单目深度估计方法常用的训练数据集. 由于 RGB 相机和深度相机之间具有不同的帧率，直接获得的 RGB 图像与深度图像的像素之间不具备一一对应的关系. 在数据处理时，通常通过关联最近像素点与引入几何投影关系对齐 RGB 图像与深度图像，由于投影的稀疏性，没有对应真实深度值的像素点将在实验时被筛除.

由斯坦福大学发布的 Make3D 数据集[22]包含成对 RGB 和深度图像，主要用于监督学习任务. 由于没有相邻帧信息，无监督学习任务往往无法直接使用该数据集进行训练，但该数据集常作为无监督学习方法的测试数据集以评估模型在其他场景下的泛化能力. Make3D 数据集主要通过激光扫描仪获取场景的深度信息，如图 2(b)所示，场景主要为白天城市与自然风光. 数据集包括 534 个 RGB 图像和深度图像对，其中 400 对样本用于训练，134 对样本用于测试.

对于室外与自动驾驶场景，由德国卡尔斯鲁厄理工学院和丰田美国技术研究院共同构建的 KITTI 数据集[23]较为常用. 除深度估计任务外，该数据集被广泛应用于光流、视觉测距、3D 物体检测等计算机视觉任务. 不同于 NYU depth 数据集，KITTI 数据集由装配有高分辨率彩色摄像机灰度摄像机、扫描摄像机、激光扫描仪和 GPS 定位系统的自动驾驶汽车采集，如图 2(c)所示，包含在市区、乡村和高速公路等场景下采集的真实图像数据. 在深度估计任务中，该数据集可以提供 93000 对 RGB 图像和深度图像.

此外，还有一些视频序列构成的数据集，可以用于无监督学习方法，比较常见的是 Cityscapes 数据集[24]. 该数据集主要关注语义分割问题，包含大约 5000 张带有精细注释的图像、20000 张带有粗略注释的图像，目前大部分方法只使用精细部分训练模型，对结果进行评估，如图 2(d)所示. 同时，该数据集还包括多个城市不同月份的双目视频序列，可以为无监督单目深度估计模型的训练提供 22973 对图像. 此数据集多用来做热身训练，用在此数据集上

(a) NYU depth数据集示例

(b) Make3D数据集示例

(c) KITTI数据集示例

(d) Cityscapes数据集示例

(e) DIODE数据集示例

图 2 数据集图像示例

训练完毕的模型在 KITTI 模型上进行训练，得到最终的模型结果.

为了进一步扩充数据集规模，由丰田日本研究院、芝加哥大学和北京航空航天大学共同构建的 DIODE[27] 数据集提供了包含室内和室外两种场景下的深度图和平面法线数据集，该数据集包括25 458张训练图片，771 张验证图片和 1629 张测试图片. 此数据集在采集时因为使用了更精准的 faro 传感器，得到的深度图和法线图更为准确. 如图 2(e)所示，采集的图像清晰地还原了真实场景的细节.

2.3 评价指标

深度估计问题中衡量模型性能的指标主要为：均方根误差（*RMSE*）、对数均方根误差（*RMSE Log*）、相对误差（*Abs Rel*）、平方相对误差（*Sq Rel*）、准确率（*Accuracy*）[9]. 以下是这些指标的计算公式：

$$RMSE=\sqrt{\frac{1}{|N|}\sum_{i\in N}\|d_i-d_i^*\|^2} \tag{1}$$

$$RMSE\ Log=\sqrt{\frac{1}{|N|}\sum_{i\in N}\|\log d_i-\log d_i^*\|^2} \tag{2}$$

$$Abs\ Rel=\sum_{i\in N}\frac{|d_i-d_i^*|}{d_i^*} \tag{3}$$

$$Sq\ Rel=\sum_{i\in N}\frac{\|d_i-d_i^*\|^2}{d_i^*} \tag{4}$$

$$Acc:\%\ \text{of}\ d_i\ \ \text{s.t.}\ \max\left(\frac{d_i}{d_i^*},\frac{d_i^*}{d_i}\right)=\delta<thr \tag{5}$$

其中，d_i 表示像素 i 点处的预测灰度值，d_i^* 表示真实灰度值，N 表示真实灰度值像素点数量，thr 表示阈值，常取 1.25、1.25^2、1.25^3.

具体地，*RMSE*、*RMSE Log*、*Abs Rel*、*Sq Rel* 均用来衡量预测结果与真实深度的误差，值越小，模型的深度估计效果越好. *RMSE* 的特点是对异常值敏感，其值的大小可反映离群点（预测结果误差较大的点）数量的多少. 由于 *RMSE* 受较大误差值主导，可以通过对数值取对数后再计算误差的方式来减小此偏差，例如 *RMSE Log* 指标. *Abs Rel* 体现了预测结果与深度值之间的相对误差，为了使这种相对差距更突出，可对其分子做平方运算，得到 *Sq Rel*. *Acc* 则是衡量模型准确率的指标，值越高，模型的估计效果越好. 通过比对预测深度在真实深度不同阈值范围内的结果百分比，可以更全面地了解模型的估计效果. 表 3 对比了上述五种评价指标.

表 3 模型评价指标对比

名称	优点	缺点
RMSE	能够反映预测结果与真实值之间的绝对误差	受异常值影响较大，无法衡量相对误差
RMSE Log	在 *RMSE* 的基础上使用对数运算，减小了异常值的影响	无法衡量相对误差
Abs Rel	能够反映预测结果与真实值之间的相对误差	差异体现得不明显，无法衡量绝对误差
Sq Rel	在 *Abs Rel* 的基础上使用平方运算，差异更显著	无法衡量绝对误差
Acc	直观地反映出预测结果的准确性	无法衡量预测结果偏离真实值的大小

2.4 应用

深度图的 3D 信息对于三维重建、导航定位、虚拟现实等应用都有比较大的帮助[30]. 在计算机视觉领域中，深度图信息可以简化很多任务的算法，它可以给出物体前后的位置信息，可用于人体姿态检测[31]、语义分割[32]、视频插帧[33]等任务. 在同时定位与地图构建中，基于图像-深度图对的方法在追踪和定位上有更稳定的效果. 对于单目重建问题来说，

从单张图(或者是静止的图序列)难以得到准确的深度信息.通过单目深度估计算法给出的一个粗略的预测,可以作为一种先验信息,对于算法的收敛性和鲁棒性会有很大的帮助[34].在工业生产方面,尽管有不少硬件能够直接得到深度图,但它们都有各自的缺陷.比如雷达设备非常昂贵、基于结构光的深度摄像头如 Kinect 等在室外无法使用,且得到的深度图噪声比较大、双目摄像头需要利用立体匹配算法,计算量比较大,且对于低纹理场景效果不好.单目摄像头相比来说成本最低,设备也较普及,所以从单目摄像头估计出深度仍然是一个可选的方案,其应用场景比其它方案更加广泛.

3　有监督学习方法

由于具有输入 RGB 图像对应的真实深度图,有监督单目深度估计可以视作单像素回归问题[9],即对于输入的 RGB 图像中的每个像素点,模型需要预测出图像对应像素位置处的深度值,因而模型输出的深度图像中的每个像素点都可以为模型的训练提供监督信息.最直接的方法是使用真实深度图与预测深度图之间的均方损失来指导模型的训练:

$$L_2=\frac{1}{|N|}\sum_i^N\|d_i-d_i^*\|^2 \tag{6}$$

从而获得具有深度估计能力的神经网络模型,整体架构如图 3 所示,其中箭头代表像素级别的对应关系和系统流程.

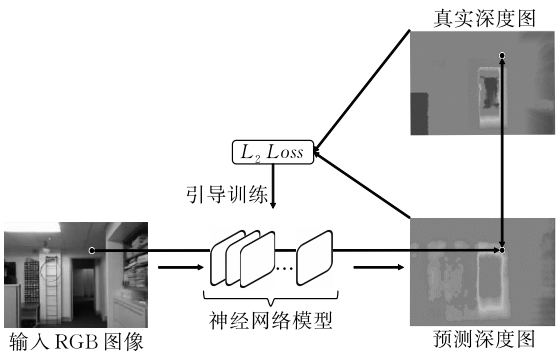


图 3　有监督学习方法基本框架[9]

据我们所知,Eigen 等人[9]最早提出了使用卷积神经网络(CNN)[35]解决深度估计问题.他们提出的网络架构由全局粗略尺度网络(Global Coarse-Scale Network)和局部精细尺度网络(Local Fine-Scale Network)组成.在深度估计时,输入的 RGB 图像经过全局粗尺度网络得到全局尺度下场景深度的粗略估计结果,然后局部精细尺度网络对全局粗略估计结果进行优化,重建更多的局部细节信息.

单目深度估计问题本质上具有不适定性,在先验约束不足的情况下,可能得到多个可行的解.由于许多解描述的场景是不可能在真实世界存在的,经过训练的模型可以在一定程度上避免预测出违背常理的深度估计结果.但是全局尺度问题仍然会导致估计结果不准确.譬如,对于一个使用普通房间图片训练得到的深度估计模型,当利用它来估计一个房屋玩具模型图片的深度时,往往会得到比真实值更大的深度估计结果,这是由于图片缺少全局尺度信息,模型并不清楚待预测深度的具体范围所导致的.全局尺度模糊将影响模型的泛化能力.

Eigen 等人[9]指出虽然全局尺度发生变化,但场景内的物体之间的深度关系不变.在此基础之上,他们提出了尺度不变误差:

$$L_{scale}=\frac{1}{|N|}\sum_i^N y_i^2-\frac{\lambda}{|N|}\left(\sum_i^N y_i\right)^2 \tag{7}$$

其中 $y_i=\log(d)-\log(d^*)$,λ 代表权重因子,当 $\lambda=0$ 时,式(7)退化成均方损失式(6),当 $\lambda=1$ 时,式(7)为完全的尺度不变损失,作者使用 $\lambda=0.5$ 均衡两种误差的权重影响,提升模型的尺度感知能力.

他们的实验结果证明了使用神经网络进行深度估计的可行性,启发了更多学者进一步研究基于深度学习的单目深度估计方法.他们的研究对本领域的发展有着深远的影响.对比图 3 中描述的基本框架与文献[9]提出的方法,我们可以初步得到以下几个基本改进方向:构建更复杂精妙的神经网络模型(3.1 节)、引入与深度相关的信息辅助深度估计(3.2 节)、使用更有效的损失函数监督网络训练(3.3 节).

3.1　改进网络结构的方法

深度学习领域的发展很大程度上体现在神经网络结构的优化上,同样地,深度估计的发展与网络结构的改进密切相关,本节将整理一些成功应用于深度估计领域的网络结构,希望给予读者启发.

深度神经网络可以被视作黑盒方法,模型在监督信息的指导下可以学习到提取特征信息、预测深度的能力.更深、更宽的网络模型往往被认为可以获取更多的信息、拥有更强的表征能力.经典的神经网络模型主要关注对网络在宽度与深度方面进行不同程度的扩增.借助于大规模数据的训练,AlexNet[36]、VGG-16[37]、VGG-19[37]等经典网络模型通过增加宽度或深度有效地提升了特征提取能力,被应用于早期的深度学习深度估计方法[9,38-39]中.

但是网络的加深将导致梯度爆炸和梯度消失问题，模型的训练难度也大大增加. 同时模型也产生了退化现象(Degradation of Deep Network)，性能反而下降. 为了克服这些困难，He 等人[40]提出了深度残差网络(ResNet)，通过恒等映射(Identity Mapping)和相关变体模型将网络的层数进一步加深. Laina 等人[41]首次将残差网络应用到深度估计领域，他们的模型整体为一个编码器-解码器结构，在编码端使用的 ResNet 使模型获得了更强的特征提取能力，获得信息更丰富的特征图再经过解码器得到最终的深度估计结果. 受此启发，以 ResNet 为基础的 DenseNet[42]也被应用于深度估计任务[43-44].

为了使网络可以接受任意分辨率的输入图片、减少网络中的参数数量，Dai 等人[45]使用 1×1 卷积替代全连接层，提出了全卷积网络架构(Fully Convolutional Networks，FCN). 文献[46]提出的方法最早将全卷积网络架构引入深度估计领域，在此基础之上，Mancini 等人[47]提出了使用光流信息引导的更加鲁棒的深度估计方法.

深度图像中单个像素的深度值与其所在的场景有着密切关联，全局信息对于深度估计同样关键. 因此，可以利用空洞卷积(Dilated Convolution)[48]增大感受野，扩大模型的感知区域以提取更丰富的全局信息[49]. 更进一步地，多尺度空洞卷积的空洞空间卷积池化金字塔(ASPP)[50]模块也被用于提取多尺度的全局信息[51].

在一些使用视频进行训练的方法中，也可使用循环神经网络(RNN)[52]、LSTM 模块[53]、元学习模块[54]进一步提取相邻帧之间丰富的深度信息、增强模型的抗遗忘性[55-57]. 为了计算深度的先验概率密度，Xia 等人[58]使用条件变分编码器结构[59]，结合来自额外信息的似然值进行深度估计，提出了基于额外真实信息的通用深度估计框架. 在一些任务中，深度估计被视作图像生成问题，研究者使用生成式对抗网络完成深度图的预测. 此外，为了获得轻量级和具有更优秀架构的模型，蒸馏网络[60]、NASNET[61]网络也被引入深度估计领域[62-63]. 由于对抗式生成网络设计思路巧妙，我们将在 3.6 节中对其进行详细的总结.

2020 年至 2021 年，自然语言处理领域的 Transformer[64]大火，研究者们将这种基于注意力机制的网络结构引入计算机视觉领域，取得了极佳的效果. 在深度估计领域，Ranftl 等人[65]首次尝试使用 Transformer 替换卷积神经网络作为特征提取器，将其应用于密集预测问题，通过交融全局与局部注意力特征，取得了极佳的深度估计结果. 作为第一篇将这种结构引入深度估计的文章，他们的结果进一步引发了学者们对 Transformer 结构的思考，如何设计更加配适于深度估计的 Transformer 结构是需要进一步探索的地方.

从不同角度出发，Bhat 等人[66]将 Transformer 作为一个空间注意力模块引入他们的框架中，他们的方法进一步挖掘了 Transformer 这种网络结构的潜力，同时证明了空间注意力机制对单目深度估计起到了重要的作用.

一些较为新颖的方法如网络结构搜索[67]、孪生网络[68]、胶囊网络[69]、自监督对比学习[70]模型等还没有被引入深度估计任务中，他们在深度估计领域的效果还等待着学者们进一步发掘.

单目深度估计领域的网络模型优化主要着重于提升模型的特征提取能力、全局信息的感知能力. 可以预见，随着深度学习的进一步发展，更多的优秀模型架构将被提出，基于网络结构改进的方法仍然有很大的发展空间.

3.2 引入辅助信息的方法

深度神经网络的灵感来源于人脑神经元的级联，类似地，基于辅助信息的改进方法主要受人类感知模式的启发. 当人类进行单目深度估计任务时，人脑会结合图片中不同物体的类别、运动情况、表面情况等多方面的场景信息进行深度估计. 类比于深度神经网络，研究者提出为深度估计引入一些场景相关的感知信息作为辅助，引导模型进行深度估计以获得更好的效果.

场景感知包含多方面内容：深度估计体现空间上的几何关系、语义信息表征不同实体的含义、表面法向量体现空间平面关系等，这些子任务在一定程度上共享相似的上下文信息，从这个角度出发，学者们提出了很多有效方法[32,38,71-84]，尝试将不同的场景感知信息融入深度估计中.

Eigen 等人[38]在文献[9]的基础之上，设计了三个尺度的网络架构，三层网络逐层补充深度图的细节，获得了更精确的估计结果. 此外，他们首次将深度估计、表面法线预测、语义分割三个子问题结合，通过他们之间的约束，进一步提升模型的效果. 在实验中，他们提出的方法在三个问题上都取得了当时最佳的性能，这印证了这些辅助任务之间的确存在联系，促进了人们对融合辅助信息进行深度估计方法的进一步深入研究.

在结合语义信息的方法[32,71-72,74-75,78-83]中，比较有代表性的是 Jiao 等人[32]提出的工作，他们指出由于深度信息服从长尾分布特性，即对于大多数实际场景，前景物体往往只占整个场景的一小部分，大部分区域为深度较大的远景，深度图主要受具有低深度值的像素主导，模型更倾向于作出低深度预测. 同样地，语义分割任务也有类似的情况. 针对这个问题，作者设计了一个协同网络，将深度估计与语义分割两个任务结合起来，两方面信息可以在模型之间进行动态传递. 此外作者使用注意力驱动机制指导模型训练，明显改善了模型对远距离物体的估计效果.

从平面法向量角度出发，文献[76]提出了同时用于深度估计与平面法向量估计的几何神经网络，通过几何约束循环迭代更新深度信息与平面法向量，可以精确预测出几何一致的深度与平面法向量结果. 类似地，引入平面法向量信息的方法还有文献[73,77,84].

为了使引入的辅助信息能够更有效地引导模型进行深度估计，学者们提出了一些很有代表性的方法[73-74]. 文献[73]利用多任务引导来设计蒸馏预测网络结构，通过产生的中间辅助任务，为学习目标提供多模态信息. 直接学习适用于多个模型的网络是很困难的，受迁移学习的启发，文献[74]使用学徒网络结构将已训练好的深度估计网络与语义分割网络结合，使学生网络同时具备两种能力，将语义信息融入到了深度估计中.

更一般地，除场景感知信息之外，一些其他相关的信息的引入也有助于深度估计. 文献[85-87]提出的方法分别将相机摄像头孔径信息、深度光学与物体边缘信息引入模型，使用不同的辅助信息引导深度估计，都获得了更准确的深度估计结果.

通过总结可以发现，大多数基于辅助信息的优化方法主要关注引入相关的场景感知信息，从单独使用某方面的场景感知信息[32,74,76]到综合使用多种场景感知信息[71,73,75]、再到引入场景感知以外的其他相关信息作为引导信号[85-87]的发展趋势. 同时，基于辅助信息的方法依赖于网络模型的改进，增强模型融合信息的能力[73-74]也是此方向进一步突破的关键.

3.3　改进损失函数的方法

损失函数用来训练模型，描述了预测值和真实值之间的差距. 在训练模型时，我们不断迭代计算，使用梯度下降的优化算法，缩小损失函数值，逐渐增强模型的深度估计能力. 损失函数体现了模型的优化目标，是深度学习领域模型优化的关键. 许多的深度估计优化方法都包含了损失函数优化部分，本小节将对其中具有里程碑意义的损失函数形式进行总结.

L_2 损失是常用于深度学习的回归损失. 在深度估计领域，因为损失中平方项的设计，着重于惩罚误差较大的估计结果. 由于深度值服从长尾分布，多数像素点集中于较低深度处，多数近距离的估计结果的误差较小，仅仅使用 L_2 损失并不合适. Laina 等人[41]在他们的工作中提出使用逆 Huber 损失(berHu loss)代替简单的 L_2 损失：

$$L_{huber}=\begin{cases}|d_i-d_i^*|, & |d_i-d_i^*|\le c\\ \dfrac{(d_i-d_i^*)^2+c^2}{2c}, & |d_i-d_i^*|>c\end{cases} \tag{8}$$

其中 c 被设置为 $1/5\max_i(d-d^*)$，当预测 $|d_i-d_i^*|\le c$ 时，该损失等价于 L_1 损失，当 $|d_i-d_i^*|>c$ 时，此项等价于 L_2 损失. 逆 Huber 损失平衡了 L_1、L_2 两种损失函数，提升了高误差像素的权重，而对低误差像素使用了更合适的 L_1 损失，被广泛应用于深度估计领域[26,88-90].

由于深度图的结构信息同样可以用于表示高层语义，Wang 等人[91]提出了一种分层次嵌入损失(Hierarchical Embedding Loss)，进一步衡量深度估计结果与真实值之间的语义空间距离. 此外，因为深度估计领域存在多种不同损失函数，不同损失之间的权重选择是一个关键难题. Lee 等人[92]提出了一种自动均衡各种损失权重的方法，充分的实验证明了该方法的有效性，为深度估计损失函数设计提供了一种新的思考角度.

从另一个角度出发，人类视觉可以很好地感知视野内的物体之间的相对深度，而不是直接估计准确的量化数据，换句话说，人们更善于估计场景中各种物体之间的相对关系. 由此便产生了一种新的深度估计思路，使用带有相对深度标记的数据集训练网络从而使其具有估计相对深度的能力，通过对相对深度信息进行整合，此类方法可以得图像的绝对深度图.

Zoran 等人[93]提出了一个框架，对像素之间的相对关系进行估计，之后将这些相对关系汇总，输出连续而密集的深度图像. 与估计每一点的绝对深度相比，估计成对点的相对关系有以下几个优点：估计点对的相对关系更加简单，相当于将复杂问题简单化、人类在相对关系问题上做得很好，故通过人工即可完成数据集的标注、消除全局尺度增加了系统的

鲁棒性.

更进一步地,Chen 等人[39]构造了一个相对深度的数据集:Depth in the Wild,每张图片仅仅标注两个随机点的相对深度,该数据集主要关注室外环境的单目深度估计,因此场景中存在深度值较大的景物(如天空).文章同时提出了一个对应的算法:首先使用排序损失(Rank Loss)训练模型预测图像上两点深度顺序关系的能力,之后给定输入图像,反复使用分类器预测稀疏点对之间的顺序关系,再通过求解二次约束优化从预测的关系中重构深度图,通过添加平滑正则项,可协调点对之间可能存在的冲突关系,使深度图更加平滑.但是由于采样的随机性,他们的方法可能产生伪纹.在此基础之上,文献[94]提出的方法将查询相对深度值点对集中于图像边缘、语义分割边缘,明显减少了伪纹的产生.

这类方法相当于引入了一种问答机制,即给出图像两点,询问哪个点距离相机更近,这种思路提供了一种全新的视角.文献[94]中提出的改进方法则关注于能否更有效地利用问答机制,将注意力集中于物体的边缘,进一步贴近了人类的感知方式.但是,此类方法仍存在缺陷:得到的深度图只包含相对深度信息,通过算法整合得到的绝对信息并不准确.将绝对信息融合进此类方法是一个可行的发展方向[95].与优化模型架构和引入辅助信息的方法相比,改进损失函数的优化方法在形式上更为简单,只需调整损失函数.但是,提出可以大幅度提升深度估计效果的损失函数,需要很精妙的设计手段,在难度上大于前者.

3.4 基于分类的方法

基于深度学习的单目深度估计方法最早作为一个单像素回归问题被提出[9],随着进一步的优化,此类方法面临着网络结构日益复杂、计算速度下降、计算成本高昂、容易陷入局部最优解等问题.为了克服这些困难,一些学者尝试将深度值离散化,将回归问题转化为分类问题.

文献[96]最早提出使用深度残差网络解决深度分类问题的方法.更进一步地,由于深度值具有有序的特点[51],深度分类问题可以归结为一个有序回归问题(Ordinal Regression).因为预测深度的不确定性随着深度的增长而变大,预测较大深度值时可以允许相对较大的误差,因此常规的均匀分段策略(UD)并不适用,作者提出使用间断距离增加(对数空间均分)的分段策略(SID).他们提出的整体网络结构可分为特征提取、场景理解、有序回归三大模块,其中场景理解模块又可以分解为全局编码器、ASPP 模块、跨通道注意力学习三个部分.

在场景模块中,利用以空洞卷积为基础的 ASPP 模块增大感受野、避免多次下采样导致的分辨率过低问题.这种架构无须复杂的信息融合手段,只需将各个部分的结果连接起来即可获得最终的特征图.为进一步降低全局编码器训练难度、提高训练效率,作者提出了一个新的编码器结构,将经过池化的特征图送入全连接层获得编码结果.

不同于传统的分类问题经过 softmax 层归一化后使用交叉熵函数作为损失,有序回归需要对损失函数进行修改,作者将损失函数定义为图像每个像素的平均有序损失:

$$L(x,\theta)=-\frac{1}{N}\sum_{w=0}^{W-1}\sum_{h=0}^{H-1}\Psi(w,h,x,\theta) \tag{9}$$

其中 θ 为权重参数,W 是图像宽度,H 是图像高度,每个图像点的有序损失为 Ψ:

$$\Psi(w,h,x,\theta)=\sum_{k=0}^{l(w,h)-1}\ln P_{(w,h)}^{(k)}+\sum_{k=l(w,h)}^{K-1}(1-\ln P_{(w,h)}^{(k)}) \tag{10}$$

其中 $x_l(w,h)$ 是像素点(w,h)在离散化深度序列中所处的位置,K 为区间总数,$P_{(w,h)}^{(k)}$ 是预测的像素点深度值大于离散化深度序列中第 k 个阈值的概率:

$$P_{(w,h)}^{(k)}=\frac{e^{y(w,h,2k+1)}}{e^{y(w,h,2k)}+e^{y(w,h,2k+1)}} \tag{11}$$

其中 $y(w,h,2k)$ 是像素点(w,h)在有序回归中的预测值,由于深度的有序性,某点的深度值要么大于或等于第 k 个阈值,要么小于第 k 个阈值,因此这个损失函数类似于二分类问题的损失函数.

该方法启发了许多相关的工作,目前深度估计最佳方法(the State of the Art, Sota),Adabins[66]就是改进方法之一,他们将文献[51]提出的分段策略进一步泛化,提出自适应区间分段,得到了极佳的效果.Huynh 等人[97]将这种有序性转化成注意力形式,用于监督网络的训练.Saha 等人[98]则是将有序性与语义信息建立起关系,用于风格迁移任务.Lee 等人[99]则着力于提出更好的离散区间表示形式.

文献[51]直至本文完成时,仍是 KITTI 数据集下公开的深度估计方法排行榜中的第五名.鉴于方法的优秀性能,本节花费了较大篇幅总结其网络结构特点、算法思想、相关改进,希望给读者以启发.

从分类问题角度思考深度估计是一个很巧妙的想法,但这种方法会存在准确率降低的问题.因为分类时模型会将一个范围内的所有预测深度归于一个

相同的区间中值，这难免会产生误差、损失精度. 但是文献[51]中提出的方法仍然取得了最好的估计结果，这体现了有序回归方法在深度估计领域的优越性，如何进一步挖掘有序回归在深度估计领域上的潜力是一个可行的发展方向[100]. 与前几种广泛应用于深度学习领域的方法相比，这个优化方式抓住了深度值有序的关系，更加契合深度估计问题的特性，但发展空间不如前者广阔，要求研究者具有更有创造性的思路.

3.5　使用条件随机场的方法

使用一个或多个网络优化深度估计结果显得繁琐复杂，一些学者希望能够使用更精简的方法完成深度图改进(Refine). 条件随机场(Conditional Random Field，CRF)是在马尔可夫随机场(Markov Random Field，MRF)的基础上发展起来的方法，它可以利用相邻节点的联系与整个观测场的信息对局部判断加以指导，从而更加合理地提取目标. 鉴于条件随机场在语义分割领域表现优秀，深度估计问题中单点深度值又与其周边深度有着密切的联系，人们自然而然地想到将条件随机场应用于改善深度估计的结果. 本节将总结此方向几个比较具有代表性的方法.

Liu 等人[101]提出了一个深度卷积神经场模型，在统一深度卷积框架中学习连续条件随机场的一元势和成对势，进行单幅图像的深度的预测. 由于给定深度值的连续性，概率密度函数中的分区函数可以进行解析计算，因此，无须任何近似即可解决对数似然优化问题. 为了提高计算效率、提高预测的准确性，文献[102]在文献[101]的基础上，提出将超像素信息编码进网络中的方法. 同年，Li 等人[103]提出先用深度神经网络对超像素尺度的深度进行回归，然后用条件随机场进行处理，通过结合超像素、像素尺度的深度，这些方法可获得更优的结果. 这三篇文章是最早将条件随机场应用于深度估计领域的工作，奠定了这个方向的基础，随后产生了许多基于他们的优化方法.

与 3.2 节中总结的方法类似，研究者们尝试将更多辅助信息引入条件随机场模型中. 文献[104]提出了在条件随机场的基础上联合训练语义分割和深度估计两个任务的方法. 基于深度图与语义标签之间的结构一致性，通过语义分割和深度信息之间相互作用，提高深度估计效果. 深度估计和语义分割任务被下式所示监督信号共同训练：

$$L=\frac{1}{N}\Big(\sum_{i}^{N}(\log d_i-\log d_i^*)^2\Big)-\frac{\lambda}{N}\sum_{i}^{N}(\log P(l_i^*))^2 \tag{12}$$

$$P(l_i^*)=\frac{\exp(z_{i,l_i^*})}{\sum_{l_i}\exp(z_{i,l_i})} \tag{13}$$

其中 l_i^* 表示真实的语义标签，l_i 表示预测的语义标签，$\exp(z_{i,l_i})$ 表示语义节点的输出. 此外他们还提出了双层条件随机场模型以更新深度细节. 文献[104]提出的双层条件随机场模型在语义分割类别数量增加时估计效果变差，为了解决这个问题，Mousavian 等人[105]使语义分割和深度估计网络共享高层特征，用端到端的训练方式得到了估计速度更快、估计效果更好的模型. 为了使辅助信息能够更好地引导模型训练，Xu 等人[106]在端到端条件随机场模型的基础上引入了结构注意力引导模块，增强了对应特征之间的信息转换.

受条件随机场模型启发，一些研究者尝试使用与条件随机场类似的模型完成深度优化任务，文献[107-108]分别提出了使用放置场、随机森林提升深度估计结果的准确度的方法.

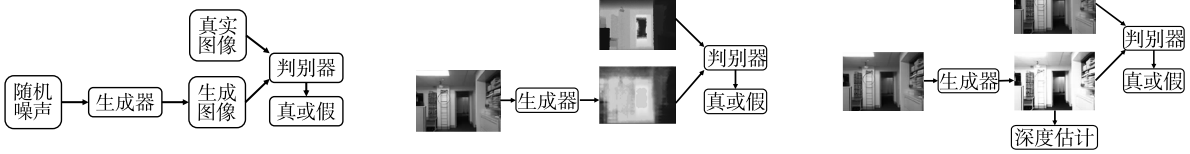
条件随机场在其他任务中表现良好，被引入到深度估计领域并取得了不错的效果，类似的工作还有文献[109-113]. 通过总结，可以得到基于条件随机场的方法的发展轨迹：首先文献[101]提出了基本方法，接下来文献[102]提出使用更复杂的结构，文献[104-106]为深度估计引入更多辅助信息，文献[107-108]尝试使用其他类似条件随机场的方法对深度进行优化.

此类优化方法比较特殊，是一种几乎纯优化的数学方法，近几年新方法产出较少.

3.6　使用生成式对抗网络的方法

生成式对抗网络(Generative Adversarial Networks，GAN)是蒙特利尔大学的 Goodfellow 等人于 2014 年提出的一种生成模型[114]，在之后引起了研究学者们的广泛关注与研究.

如图 4(a)生成式对抗网络由生成器(Generator)与判别器(Discriminator)组成，在训练过程中通过相互竞争同时让这两个模型得到增强. 由于判别器的存在，生成器在没有大量先验知识的前提下也能很好地生成逼近真实数据的图像，使判别器无法区分生成器生成的图片与真实图片. 设 x 是真实深度图，$\hat{x}$ 表示生成器生成的深度值，训练过程中的优化目标可以写作如下形式：



(a) GAN 的基本框架[114] (b) 使用GAN的单目深度估计框架[115] (c) 使用生成数据训练的框架[116]

图 4 使用生成式对抗网络方法的基本框架

$$\arg\min_G \arg\max_D E_{x\sim P_{gt}}[\log D(x)]+E_{\hat{x}\sim P_{gt}}[\log D(\hat{x})] \tag{14}$$

深度估计某种程度上也是图像生成任务,因此 Jung 等人[115]将生成式对抗网络结构引入深度估计领域,生成器接收 RGB 图像生成对应的深度图,判别器判别该深度图是否为真实深度,整体架构如图 4(b).

以文献[114]提出的基本架构为基础,产生了多种类型的生成式对抗网络,它们被广泛地使用在深度估计领域,比较有代表性的有使用堆叠的生成式对抗网络(Stacked GAN)的文献[116](无监督,但描述相同思想)、使用条件生成式对抗网络(Conditional GAN)的文献[117-118]等,他们整体框架与图 4(b)类似,都是将深度估计视作深度图像生成问题.

基于生成式对抗网络的方法的主要思路是将基本的单像素回归问题整合起来,转化为一个图像生成问题,进而依靠生成式对抗网络的强大生成能力完成深度图估计. 基于生成式对抗网络的方法一个巨大的优势是有着非常不错的主观效果. 但是,生成式对抗网络训练困难,这导致一些本来合理的模型难以产生更优的结果.

此外,生成式对抗网络在估计时存在不确定性,客观指标较差,故此类方法不适用于一些需要高确定性估计结果的场合.

使用生成式对抗网络的另一个目的是实现域迁移. 基于监督学习的单目深度估计方法在训练时需要大量真实数据,这往往是难以获得的,一个解决的方案是使用人工生成的数据(Synthetic Data).

但是,生成数据与真实数据之间存在域偏差,直接由生成数据训练出的模型在真实世界的泛化效果并不理想,学者们通过引入生成式对抗网络实现从生成数据域到真实数据域的迁移,从而保证由生成数据训练出的模型的泛化能力. 基础模型如图 4(c)所示,通过生成器将真实世界数据分布迁移至生成数据分布,这样使用生成数据训练的模型就可以给出较准确的真实世界深度估计结果.

根据 Li 等人[119]提出的风格迁移和域自适应之间的关系,Atapour-Abarghouei 等人[120]提出了使用循环生成对调网络(Cycle GAN)[121]完成生成数据到真实世界数据的风格迁移过程. 该模型由两个对称的生成式对抗网络组成,在训练时,前向判别损失、后向判别损失与一致性损失共同组成监督信号,引导模型进行训练. 训练结束后,他们使用真实数据到生成数据的生成式对抗网络对 RGB 图像进行域迁移,再进行深度估计. 在此基础上,Chen 等人[122]将语义分割信息作为引导,进一步生成了更加准确的深度图像. 文献[123-124]也是类似的使用生成数据进行训练的方法.

基于上述基本框架,目前一些新型的方法着力于探索新的约束条件对域适应进行限制,试图提升域迁移效果,获得具有更高性能的模型. Lopez-Rodriguez 等人[125]引入语义分割知识,使用高层的语义信息作为引导,增强模型的深度估计能力. 在 Cycle GAN 的基础框架之上,Zhao 等人[126]结合深度估计任务本身的几何约束性进一步监督模型训练,提出了几何对称的域自适应深度估计框架. 通过进一步结合语义分割、光流估计等任务,Chen 等人[127]提出了基于多任务的跨域约束,从而进一步提升了模型的泛化能力. Akada 等人[128]则是在文献[129]的基础上进一步探究如何学习到具有域不变性的几何特征.

领域迁移可以减少模型的训练数据量,且此类方法的目的更贴近于学习人类举一反三的能力,这些特点使基于风格迁移的方法成为了目前比较火热的研究方向. 但是通过总结我们可以观察到,相关方法多拘泥于基本的 GAN 的框架,能否做出进一步的突破,将深度估计的几何特性作为核心,创造出更具任务适配性的风格迁移方法,是此领域进一步发展的关键.

3.7 基于部分深度信息的方法

在实际应用中,一些设备可以获取部分场景的深度信息. 如图 5 所示,图 5(a)代表真实深度图,其他子图分别代表了不同系统可以获得到的部分深度信息. 更具体地,图 5(b)中的系统获取到了场景中正方形框部分的深度值,图 5(c)代表了雷达投影深

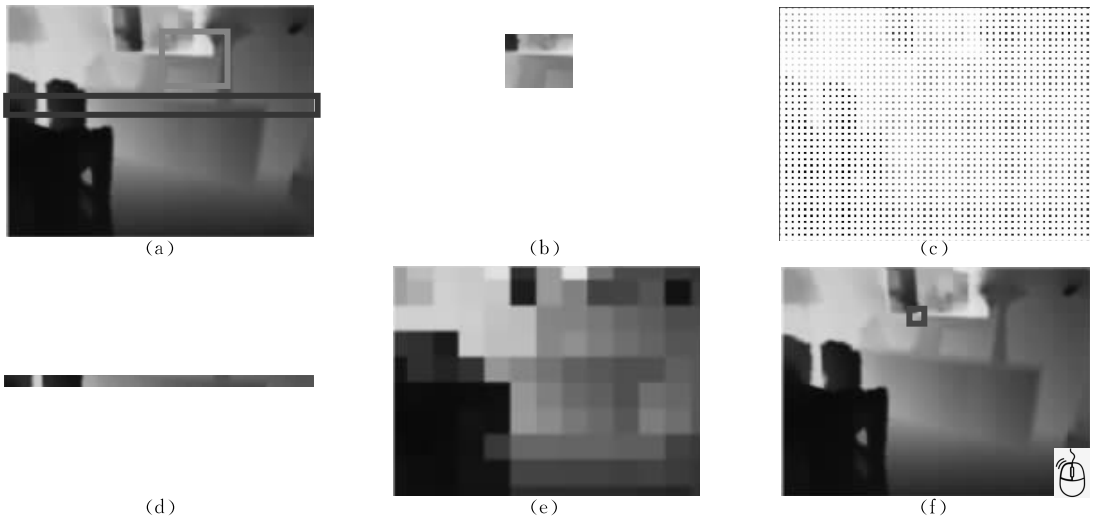


图 5　几种常见的部分深度图示例

度图(并非与原始深度图场景对应),图 5(d)是激光获取的场景图 5(a)中长条形框部分的深度图,图 5(e)是真实深度图是经过下采样处理后的深度图,图 5(f)代指用户可以标注场景某一区域的真实深度.利用获取到的部分深度信息,神经网络可以对 RGB 图像的深度进行更准确地估计,本小节将对此类任务中比较具有代表性的几种方法进行总结.

类比图 3,我们总结出此类方法的基本框架,如图 6.在将图片送入神经网络进行深度估计之前,把 RGB 图像与获得的部分真实深度图连接(Concatenate)起来,这样就可以将已知的部分真实深度信息引入至深度估计中.学者们在这个框架的基础之上进行改进,提出了许多有效的方法.

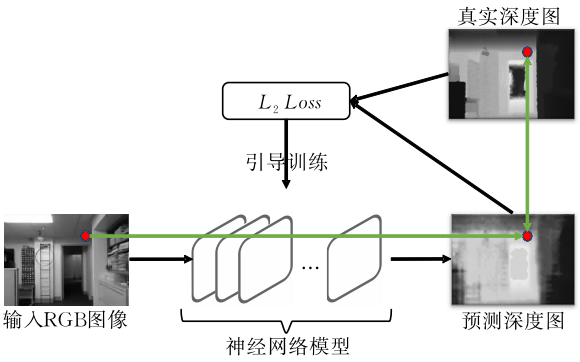


图 6　基于部分真实深度图方法的基本框架[129]

文献[129]是较早使用深度学习框架完成此类任务的方法,他们的模型主要关注于结合线性激光获取的深度信息进行深度估计,即图 5(d)所描述的情况.他们的实验结果表明,由于有真实深度的区域面积太小,深度信息无法被充分提取,基本框架中直接将 RGB 图像和线性激光获取的深度图连接起来的方法并不能得到更为准确的深度估计结果.

为了克服这个困难,他们提出构造参考深度图(Conference Depth Map)的方法,通过结合激光仪器和 RGB 相机,将真实深度纵向扩展,保证了整张图片的每个像素都可以直接接受引入的真实深度信息的引导.之后他们将参考深度图与 RGB 图像连接,送入整体为一个编码器-解码器结构的神经网络进行深度估计.在训练时,他们将回归损失与分类损失结合,获得了更准确的深度估计模型.

在基于如图 5(c)离散真实深度值的深度估计任务中,比较有代表性的是 Jaritz 等人[59]与 Chen 等人[130]提出的方法.文献[63]提出的方法思路与文献[129]类似,他们没有直接将 RGB 图像与稀疏深度图连接,而是都各自经过一个特征提取网络,并将提取的特征连接在一起进行深度估计.为了获取更轻量级的网络模型,他们使用了神经网络搜索方法降低参数数量.同时,他们的实验结果表明使用变化的稠密度的真实深度图对模型进行训练可以得到更好的结果.为了进一步提取稀疏深度图中的真实值的空间位置信息,文献[130]提出了稀疏深度输入参数化的方法,他们将离散的真实深度图映射成 S_1 与 S_2 两个分别包含深度信息与空间信息的图像,并将 S_1、S_2、RGB 图像连接起来进行深度估计.一些最新的方法,如文献[131]则着重探究如何联合基本深度估计框架与基于部分深度信息的框架,提出一种插入模块,巧妙地将部分深度信息引入到基本深度估计框架中.文献[132]则进一步优化离散深度估计方法的性能,探索准确估计物体边缘位置处深度的方法.

为了将多种任务结合起来,Wang 等人[133]提出了一个即插即用(Plug and Play,PnP)模块,可以改

善以任意稀疏深度模式为输入的深度预测结果，将一个深度估计任务转化成深度优化任务. 给定任何预先训练的深度预测模型，他们提出的即插即用模块迭代更新模型中间的特征映射，使模型输出与给定稀疏深度一致的新深度. 更进一步地，Xia 等人[58]使用条件变分自编码器对深度的概率密度进行估计，将深度的概率密度作为先验信息，额外的真实深度值作为似然，对深度进行最大先验估计(Maximum A Posteriori Estimate，MAP)，提出了可以同时完成多项深度估计任务的统一模型.

此类方法有着较为可观的应用空间. 在实际应用中，虽然获取稠密的深度图代价昂贵，但是通过传感器等设备获取部分真实深度值的成本是可以接受的，在部分真实深度值的基础上对深度图进行补全是一个可行的方向. 通过总结，可以发现此方向的趋势是融合多种模式，使用统一的框架完成深度估计.

3.8 有监督方法小结

基于有监督学习的深度估计问题最早作为单像素回归问题被提出，产生两大优化方向：(1) 更准确的估计结果；(2) 更快速的估计方法.

从准确性出发，研究者们使用了更复杂的网络结构(3.1 节)、尝试将更多辅助信息引入深度估计模型中(3.2 节)、使用更有效的损失函数(3.3 节)、使用条件随机场对深度估计结果进行优化(3.5 节)、使用生成式对抗网络进行深度图像生成(3.6 节)、基于部分深度信息的方法(3.7 节). 随着深度学习的迅速发展，提高准确性方向还将会涌现出很多优秀的方法.

从速度角度出发，研究者们提出将回归问题转化为分类问题的方法(3.4 节)和一些使用轻量级模型的方法如文献[134]. 相比较而言，这个方向的研究数量较少，但是又很关键，因为很多应用场景在准确性得到保证的前提下，要求使用更轻量级的模型(如嵌入式系统)、更快速的估计方法(如智能驾驶系统)，这限制了单纯从准确性出发的优化方法的应用范围，因而研究优秀的简化模型、加速算法是深度估计领域一个很有潜力的发展方向. 目前有一些无监督方法关注于实时性和轻量级，我们将在 4.7 节中一同给予阐述.

从应用角度出发，一些方法在已有部分深度值的基础上，对深度进行补全(3.7 节)，得到了更准确的深度估计结果.

表 4 对现有的有监督单目深度估计方法进行了分类. 值得注意的是，许多方法采用了多种优化手段，融合多种方法的优势是本领域发展的趋势. 更具体地，表 5 对每个有监督方法的主要贡献按时间顺序进行总结. 由于各种方法基于的数据集不同，我们总结了深度估计领域最常用的两个数据集：NYU 和 KITTI (截断到 80 m)数据集上各种方法的 *RMSE* 指标. 此外，因为 NYU 数据集主要包含室内场景，KITTI 数据集主要是室外场景，所以此表也体现了不同方法在不同数据集上的适应性. 一些方法由于评价指标的不同、使用的数据集不同，并没有对应的客观指标结果，这些地方我们使用“—”表示数值缺失. 从此表中我们还可以发现，有监督方法多关注于 NYU 室内数据集，大多数方法都提供了 NYU 数据集上的客观指标结果.

表 4 有监督方法分类

改进方式	优点
3.1 改进网络结构的方法	[38,39,41,47,51,102,135] [49,58,72,90,116,133] [60,63,65,66]
3.2 引入辅助信息的方法	[26,38,46,86,135] [43,72-75,103,122]
3.3 改进损失函数的方法	[39,41,51,91,92,94]
3.4 基于分类的方法	[51,66,96-99]
3.5 使用条件随机场的方法	[49,101,107,108,116]
3.6 使用生成式对抗网络的方法	[103,115,116,117,118] [122-128]
3.7 基于部分深度信息的方法	[58,63,129-133]

表 5 有监督方法总结

方法	提出年份	主要贡献	来源	NYU	KITTI
Eigen et al.[9]	2014	卷积神经网络	NIPS	—	7.156
Eigen et al.[38]	2015	多任务框架	ICCV	0.641	—
Shelhamer et al.[46]	2015	多任务框架	ICCV	—	—
Liu et al.[101]	2015	CRF	CVPR	—	6.986
Li et al.[103]	2015	CRF\多任务框架	CVPR	0.635	—
Mancini et al.[47]	2016	全卷积网络	IROS	—	6.863
Laina et al.[41]	2016	残差网络\berhu 损失	3DV	0.510	—
Chen et al.[39]	2016	rank 损失	NIPS	0.17	—
Mayer et al.[26]	2016	多任务框架	CVPR	—	—
Roy et al.[108]	2016	随机森林	CVPR	0.744	—
Cao et al.[96]	2017	分类任务	T CSV	0.688	6.209

（续　表）

方法	提出年份	主要贡献	来源	NYU	KITTI
Jung et al.[115]	2017	Conditional GAN	ICIP	0.527	—
Liao et al.[129]	2017	线性雷达深度补全模型	ICRA	—	—
Fu et al.[51]	2018	ASPP\有序回归	CVPR	0.509	2.727
He et al.[43]	2018	DenseNet\焦距编码	TIP	0.572	4.014
Li et al.[49]	2018	空洞卷积\CRF	PR	0.505	4.687
Liu et al.[72]	2018	解卷积网络\多任务框架	T NNL	0.628	—
Xu et al.[73]	2018	多任务框架	ECCV	0.792	—
Chen et al.[130]	2018	稀疏深度图补全模型	ECCV	—	—
Atapour et al.[120]	2018	生成数据域迁移	CVPR	0.318	4.726
Jaritz et al.[63]	2018	稀疏深度图补全\多任务	3DV	—	0.917
Ye et al.[74]	2019	多任务框架\学徒网络	CVPR	0.687	—
Chang et al.[86]	2019	深度光学	ICCV	0.433	1.928
Chen et al.[90]	2019	LSTM	机器人	—	4.933
Chen et al.[122]	2019	生成数据训练\多任务框架	ICCV	—	—
Wang et al.[133]	2019	即插即用深度补全模块	ICRA	0.159	2.975
Eldesokey et al.[135]	2020	NCNN\置信度信息	CVPR	0.135	1.237
Xian et al.[94]	2020	结构引导的 rank 损失	CVPR	—	—
Ramamonjisoa et al.[107]	2020	放置场	CVPR	0.352	—
Xia et al.[58]	2020	统一建模深度补全任务	CVPR	—	—
Wang et al.[91]	2020	多模态深度补全框架	ECCV	0.493	—
Lee et al.[92]	2020	多层嵌入损失	ECCV	0.430	—
Huynh et al.[97]	2020	有序深度注意力机制	ECCV	0.412	—
Wang et al.[113]	2021	离散卷积\CRF	MLC	—	—
Guizilini et al.[131]	2021	插入式深度补全模块	CVPR	—	0.909
Imran et al.[132]	2021	关注边缘的深度补全方法	CVPR	0.092	0.840
Bhat et al. (SoTA)[66]	2021	动态区间\transformer	CVPR	0.364	2.360
Saha et al.[98]	2021	借助语义和深度完成域适应	CVPR	—	—
Lee et al.[99]	2021	改良的有序深度表示方式	CVPR	0.104	0.867
Ranftl et al.[65]	2021	Transformer 应用	Arxiv	0.357	2.573
Akada et al.[128]	2021	引入对比学习进行域适应	Arxiv	—	5.498

由于监督学习需要大量的带标注训练数据（真实深度值），而带标注的数据往往有限且难以获取，研究者们开始将目光转向不需要标注数据的无监督学习方法. 这种方法可以充分利用数据中的几何信息对模型进行训练，近年来成果颇丰，我们将在下一节对此类方法进行综述.

4　无监督学习方法

不同于代价高昂的有监督方法，无监督方法无须自带标签的数据即可完成模型的训练. 在训练时，算法根据数据本身的结构和特性，通过几何关系，将标签构造出来，进而就可以像有监督学习一样对模型进行训练. 早期的无监督方法使用图像对进行训练（4.1 节）. 为了进一步降低使用训练数据的限制，一些学者提出了使用视频训练的方法（4.2 节）. 此后，以这两种训练方法为基础，产生了许多相关的优化方法（4.3～4.7 节）.

4.1　使用图像对训练的方法

早期的无监督方法主要通过图像对之间的几何关系作为约束训练模型.

如图 7，图中左右相机 LR 拍摄的物体 $P(x,z)$ 分别在相机成像平面（Image Planes）上投影. 其中 x_l 代表左相机成像坐标，x_r 代表右相机成像坐标，f 表示相机焦距，dis 表示左右相机成像的视差，b 表示左右两相机之间的基线距离（Baseline），z 表示成像物体的深度. 根据几何知识，可以推导出左右相机之间成像的关系：

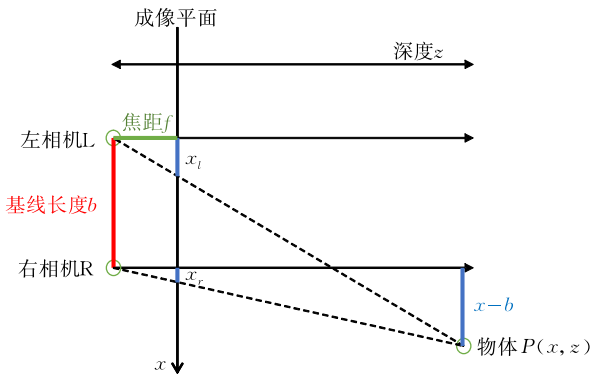


图 7　图像对之间的投影几何约束

$$x_l - x_r = \frac{f \times b}{z} = dis \tag{15}$$

在已知焦距 f、基线距离 b 的条件下，只需估计出视差 dis，就可以实现左相机至右相机或右相机至左相机的图像重建并计算出物体深度 z.

Garg 等人[136]最早提出使用无监督学习方法训练模型的基础框架. 如图 8，该框架由估计深度的卷积神经网络与图像重建模块构成. 由于深度服从长尾分布，深度值普遍较小，为了更快地训练，他们对逆深度即视差进行估计，根据式(15)在已知焦距 f、基线距离 b 的条件下，由视差 dis 可以进一步计算出深度. 图像重建模块根据左视图深度与右视图重建出新的左视图，这样左视图与重建图像之间的误差就可以作为监督信号，引导深度估计网络的训练. 他们使用的损失函数为

$$\begin{aligned} L_{rec} &= \sum_x \| I_l(x) - I_w(x) \| \\ &= \sum_x \| I_l(x) - I_r(x + dis(x)) \| \end{aligned} \tag{16}$$

其中 $I_l(x)$ 表示左相机成像，$I_w(x)$ 表示从右相机图像重建的左视图. 除此之外，为了获得轮廓更为清晰的深度图像，他们还设计了梯度平滑损失与重建损失一起引导模型进行训练. 为了计算反向传播的梯度，他们在变形重建时使用泰勒展开进行线性优化处理，但是此方法的结果并不完全可微，模型可能陷入局部最优解. 为克服这个困难，Godard 等人[137]引入可微的插值函数. 此外，他们还优化了损失函数，提出了后来被广泛应用于无监督深度估计领域的表面匹配损失 L_{ap} 和视差平滑损失 L_{ds}：

$$\begin{aligned} L_{ap} = \frac{1}{N} \sum_{i,j} \alpha \frac{1 - SSIM(I^l_{i,j} - \hat{I}^l_{i,j})}{2} + \\ (1-\alpha) | I^l_{i,j} - \hat{l}^l_{i,j} | \end{aligned} \tag{17}$$

$$\begin{aligned} L_{ds} = \frac{1}{N} \sum_{i,j} | \partial_x d^l_{i,j} | e^{-\| \partial_x I^l_{i,j} \|} + \\ \frac{1}{N} \sum_{i,j} | \partial_y d^l_{i,j} | e^{-\| \partial_y I^l_{i,j} \|} \end{aligned} \tag{18}$$

其中表面匹配损失 L_{ap} 结合了 L_1 损失项与 $SSIM$ 损失项，α 是一个权重因子，I 与 $\hat{I}$ 分别表示原图像与重建图像. 视差平滑损失 L_{ds} 中 ∂_x 表示图像横向梯度，∂_y 表示图像纵向梯度，d 代表预测的深度图，因为深度图和 RGB 图在物体边缘处都有较大的梯度，故此项损失约束了梯度的一致性.

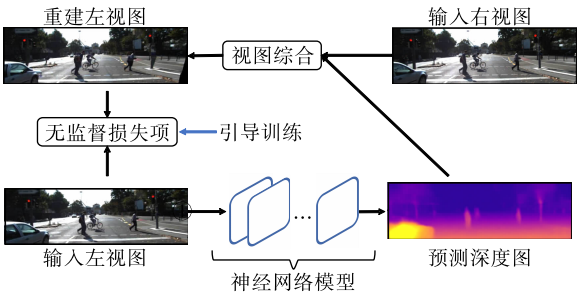


图 8 使用图像对训练的无监督学习方法[136]

此外，他们还在文献[136]的基础上引入了左右深度一致性约束，模型如图 9 所示，左右对称性更好地帮助了模型的训练. 更进一步地，Poggi 等人[138]提出使用三张图片之间的几何约束进行模型训练，有效减少了遮挡像素对模型训练的影响.

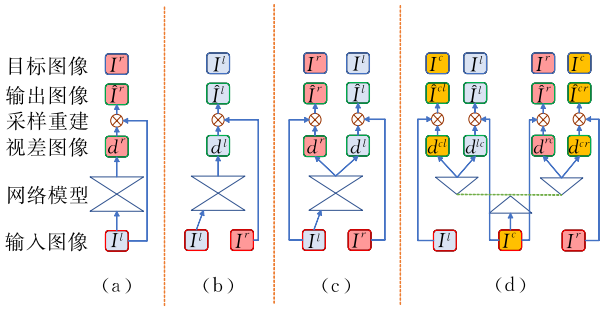


图 9 改进框架

文献[137]在对比实验中取得了优于监督学习的效果. 同时，由于可使用的数据更加广泛，他们的模型在泛化能力上也有出色的表现. 他们的工作激励了研究人员对无监督深度估计方法的研究.

这种类型的方法在训练时使用图像对，在训练完成后，只需接收单张 RGB 图像即可给出对该图像的深度估计结果.

虽然基于图像对的方法获得了可观的效果，但是，这种方法有着明显的缺陷：需要确定相机间的基线长度与相机焦距，且同时需要左右两视角下的图像. 这限制了可用于训练的数据集范围. 为了进一步减少限制条件，研究者通过进一步利用视频序列间的几何关系，提出了使用视频进行训练的无监督方法.

4.2 使用视频训练的方法

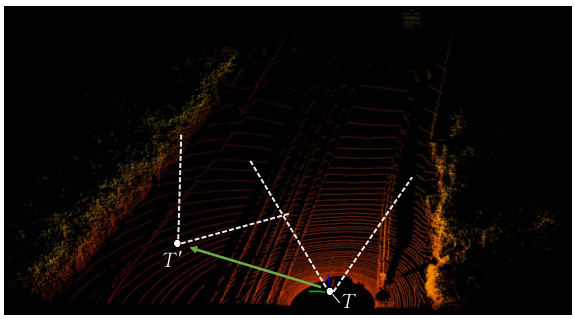
使用视频进行训练，其约束建立在相邻帧 I_n 与 I_{n-1} 之间的几何投影关系之上：

$$p_{n-1} \sim \boldsymbol{K} \boldsymbol{T}_{n \to n-1} D_n(p_n) K^{-1} p_n \tag{19}$$

其中 p_n 代表图像 I_n 中的像素，p_{n-1} 代表图像 I_{n-1} 中与 p_n 对应的像素，$\boldsymbol{K}$ 代表相机的内参矩阵，$D_n(p_n)K^{-1}$ 表示 p_n 处像素的深度，$\boldsymbol{T}_{n \to n-1}$ 表示相机的空间转换矩阵.

如图 10 所示，对于一张 RGB 图像(图 10(a))，通过逆投影 $D_n(p_n)K^{-1}$ 后即可形成相机 T_n 角度的空间点云(图 10(b))，接着左乘相机姿态转换矩阵

(a) RGB图像



(b) 点云图像

图 10　邻帧几何投影约束示意

$\boldsymbol{T}_{n\to n-1}$可以完成箭头所示的变换，最后左乘相机内参矩阵 $\boldsymbol{K}$ 即可将点云投影至相机 T'。由此，在确定 $D_n(p_n)$与 $\boldsymbol{T}_{n\to n-1}$的前提下，图像 I_n与 I_{n-1}之间便可以通过投影建立起联系.

在此基础之上，如图 11，Zhou 等人[139]设计了一个深度估计网络(DepthNet)从单张图片中估计深度 $\hat{D}_n$，另一个姿态估计网络(PoseNet)从邻帧图像对 I_n，I_{n-1}中估计相机转换矩阵 $\hat{\boldsymbol{T}}_{n\to n-1}$基于神经网络的输出，根据式(20)，相邻帧图片像素之间的对应关系可以建立起来：

$$p_{n-1}\sim \boldsymbol{K}\hat{\boldsymbol{T}}_{n\to n-1}\hat{D}_n(p_n)K^{-1}p_n \tag{20}$$

由于坐标可能落入栅格内部，作者使用双二线性插值保证像素点一一对应. 受文献[136]的启发，为了衡量内容差距，他们引入了类似的结构化损失，使用 L_1损失与 $SSIM$ 损失结合的重建损失 L_{rec}作为训练模型的监督信号：

$$L_{ap}=\frac{1}{N}\sum\alpha\frac{1-SSIM(I_n-\hat{I}_n)}{2}+(1-\alpha)|I_n-\hat{I}_n| \tag{21}$$

其中 α 是一个权衡因子，I_n与 $\hat{I}_n$分别表示原图像与重建图像.

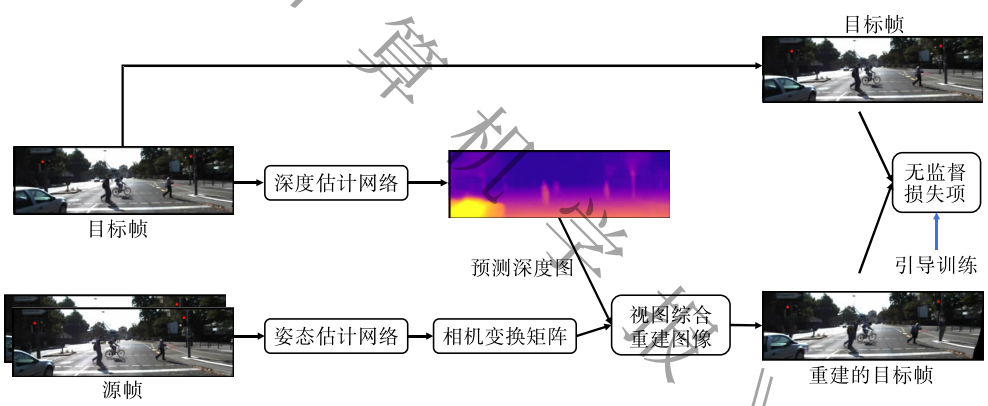


图 11　使用视频训练的无监督学习方法[139]

对于一个视频流，他们选择 t 时刻的图像作为重建目标(目标帧)，将附近几帧(I_{t-1}，I_{t+1})作为采样源(源帧)，使用每个图像对((I_{t-1}，I_t)，(I_t，I_{t+1}))的重建误差的均值训练网络. 但是如果目标帧中某点的对应像素可能由于遮挡等原因没有出现在源帧，该图像对产生的重建误差将偏大，进而影响整体的训练过程. 为了解决这一问题，Godard 等人[140]提出使用多个图像对重建损失的最小值训练网络，这样一个微小的改动明显地提高了深度估计结果. 此外同样受文献[136]的启发，他们也引入了类似的梯度平滑损失：

$$L_{smooth}=\frac{1}{N}\sum_{P}^{N}|\nabla D(p)|\times(e^{-|\nabla I(p)|})^T \tag{22}$$

其中 D 与 I 分别表示深度图和 RGB 图像，∇表示求梯度运算. 此项损失对图像起到了平滑作用. 为了增强深度估计结果的一致性，Mahjourian 等人[141]提出了三维空间中的重建损失，进一步与将场景的几何约束引入至模型的训练中.

式(20)～(22)构成了使用视频训练的无监督学习方法的基本框架. 虽然在训练阶段使用的是多帧视频，但在测试阶段，只需将 RGB 图像输入至深度估计神经网络就可以得到深度估计结果. 使用视频训练的无监督学习方法进一步降低了数据的使用限制，获得了可观的效果，是如今研究的热门方向. 我们将在接下来的几节中对相关的优化方法进行详细的讨论.

4.3　基于可解释掩膜的方法

由于缺乏真实深度值作为强有力的监督信号，无监督方法的准确率不高. 除此之外，一个主要原因是一些像素打破了投影重建算法的限制条件. 无监

督方法在相机转移时默认物体静止，对于运动物体，该假设被影响，重建效果变差，此时的误差难以有效地训练模型.对于一个在相机坐标系里与相机相对静止的物体，它在相机里的位置不变，对应的视差(逆深度)为零，在理论上会有无穷大的深度，模型的任何预测结果都将产生巨大误差，进而将严重影响模型的训练效果，最终导致无穷深度的出现.因此，掩膜(Mask)技术被研究者提出，此类方法的核心思想是过滤掉打破静态假设的像素点，以避免这些像素点对模型的训练产生负面影响.

文献[139,142]中提出的基于可解释性掩膜的方法的关键在于如下形式的损失函数：

$$L=\frac{1}{N}\sum_{P}^{N}M\left\|I_n(p)-\hat{I}_n(p)\right\| \tag{23}$$

其中 M 代表了通过一个掩膜网络预测到的可解释性掩膜，M 中每个值对应图像中的每个像素点的重建误差权重，通过不同的取值，掩膜可以调整各个像素点对模型的影响.由于网络将会倾向于预测 $M=0$ 以降低损失，故 Zhou 等人[139]同时使用了关于 M 的正则项误差 $L_{reg}(M)$ 鼓励网络掩膜网络做出 $M\neq 0$ 的预测. Vijayanarasimhan 等人[142]则是直接设计了一个更精细的掩膜网络对掩膜 M 进行估计.

虽然掩膜方法提升了准确率，但是它要求更高的计算量以及更复杂的网络训练方法.为了简化模型，一些研究人员提出了基于几何约束的掩膜方法[143-144].其中，文献[144]提出使用重叠掩膜(Overlap Mask)和空白掩膜(Blank Mask)减小物体重叠区域和空白区域对损失函数的影响.文献[143]提出了一个基于深度图和相邻图片的自发现掩膜(Self-Discovered Mask).这些方法降低了计算复杂度，获得了更精简的模型，同时深度估计更加健壮，效果也有所提升.更进一步地，Godard 等人[140]提出了在重建过程中即可计算的自动掩膜方法，如图 12 所示，上方的图片来自车载相机，与相机相对静止的车辆、天空被掩膜筛去，在二值图像中呈现黑色.下方的图片来自静止的监控摄像头，因此照片中大部分物体与相机相对静止，二值图像中大部分区域为黑色.

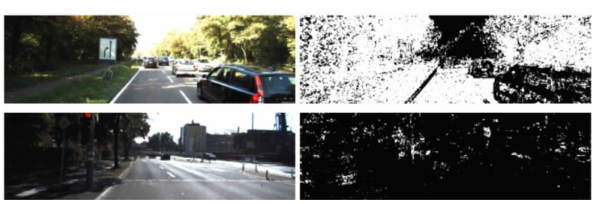


图 12 可解释性掩膜示例[140]

可解释掩膜类方法[139,141-145]有效地降低了打破静态假设的像素点对模型的影响，后续许多不同类型的方法也都引入了掩膜技术，比如在多任务框架中使用掩膜[146-147]、在生成式对抗网络框架中使用掩膜[148]等.随着可解释性掩膜方法的轻量级化，这种优化手段的普适化是必然趋势.

4.4 基于视觉里程计的方法

使用视频进行训练的无监督方法由深度估计网络和相机姿态估计网络两部分组成，使用准确率更高的相机姿态估计方法是一个提高准确性的方式.

通过视觉获取相机姿态的问题通常被称为视觉里程计(Visual Odometry)，随着该领域的发展，一些估计相机姿态的方法已经获得了很好的效果，可以替代文献[139]中的姿态神经网络.

按照这个思路 Wang 等人[149]提出使用直接测距法(Direct Vsual Odometry，DVO)[150]，通过最小化三帧图像之间的重建误差获得相机姿态转换信息.由于使用直接测距法可以获得更准确的相机姿态转换矩阵 $\hat{\boldsymbol{T}}_{n\to n-1}$，他们的模型得到了更准确的深度估计结果.此外，他们还通过实验证明了在计算损失之前，将深度值进行归一化可以得到更好的训练效果.

类似地，Dai 等人[151]采用 ORB-SLAM[152]算法全局优化相机姿态.选定初始帧后，使用随机抽样一致性(RANSAC)算法[153]进行图像的特征点匹配，继而计算两帧之间的相机变换矩阵.

鉴于深度估计与相机姿态估计之间的紧密联系，另一类的方法[55-56,116,147,152-153]通过联合训练单目深度估计和视觉里程计，在深度上增加约束，可以同时改进深度估计与视觉里程计结果.

其中，文献[154]最早提出了联合学习深度估计和视觉里程计的无监督框架.他们利用了图像在时空上的约束，改进了深度估计的性能，并且消除了尺度不确定性问题，同时获得了当时最好的无监督视觉里程计估计效果.在此基础之上，为了进一步获取上下文信息，文献[55]引入了循环神经网络.从图像生成角度出发，文献[116,155]引入了生成式对抗网络.

为更好地启发读者，进一步结合传统方法和深度学习方法，现将对三种经典的视觉里程计方法进行简单介绍.在此类传统视觉里程计方法的帮助下，深度估计的准确性可能会进一步提升. ORB-SLAM[152]属于纯特征的视觉里程计方法，是一个支持视觉、视觉加惯导、混合地图的系统.它是第一个基于特征的紧耦合的 VIO 系统，仅依赖于最大后验估计. SVO[156]属于一种半直接的视觉里程计方法，它跟踪图像关

键点(由 FAST 实现的角点)，然后像直接法那样，根据关键点周围的信息，估计相机运动以及他们的位置. DSO[157]是一种稀疏直接视觉里程计方法，不包含回环检测、地图复用的功能. 它通过最小化光度误差来计算相机位姿与地图点的关系，将数据关联和位姿估计放在了一个统一的非线性优化问题中. 由于直接法容易受到光照影响，DSO 提出了光度标定，提升了算法的鲁棒性.

使用更准确的相机姿态估计方法可以减少由于姿态估计网络产生的误差，将训练的重点放在深度估计网络上，因此从思路上来讲使用视觉里程计方法替代姿态估计网络是一个直观、行之有效的办法. 此外，由于视觉里程计问题与深度估计问题的相关性，将视觉里程计引入深度估计，实现深度与视觉里程计的联合估计，也有一定的发展空间.

4.5　引入辅助信息的方法

与使用监督学习的深度估计方法的发展思路相似，同样由于深度信息的匮乏，研究者关注于将其他相关的信息引入基础的深度估计任务中，比如：物体运动信息、语义分割信息、相机内参矩阵. 由此，不同任务之间的关系就可以作为额外的信号训练模型，提高模型深度估计的能力.

从物体运动信息角度出发，Yin 等人[158]提出了一个联合估计深度和光流(运动)的工作，网络包含刚性结构重建模块和非刚性运动定位模块，同时提升了深度估计和光流估计两个任务的准确率. 由于文献[158]中提出的方法使用深度估计网络和姿态估计网络对光流进行预测，深度估计网络和姿态估计网络的误差会传播到光流的预测中，这将降低结果的准确率，文献[159]提出了直接使用一个光流估计网络去预测光流值的方法，得到了更好的效果. 在引入光流信息的基础方法之上，Hur 等人[160]进一步使用带有空间信息的场景流信息辅助估计深度. 光流网络的结果可以帮助修正深度估计网络产生的误差，故此类方法的一个优势是可以借助运动信息减小运动物体对深度估计模型训练的影响.

为了更好地对运动物体建模，文献[160]提出使用一个神经网络对独立物体的运动进行估计，并联合训练深度估计网络. 更进一步地，文献[161]将物体的瞬时运动速度引入作为监督信息，试图解决无监督学习框架下的尺度模糊问题.

从语义信息角度出发，Guizilini 等人[161]使用了一个预先训练好的语义分割模型引导深度估计，预训练好的语义分割网络得到的语义信息通过像素自适应卷积引导模型进行深度估计. 与有监督学习引入语义分割信息的效果类似，此类方法可以获得边缘更为清晰的深度估计结果[161-163]. 引入语义信息的另一个出发点是消除动态物体对训练的影响，Klingner 等人[164]提出使用语义分割对生成视图进行预处理，只使用静止部分训练网络，进一步提升了深度估计的效果.

上述方法要求已知相机的内参矩阵，文献[165]提出的方法将姿态估计网络扩展，使其具有预测相机内参矩阵的能力，进一步减少了训练过程要求的已知参数. 目前，这个方向的研究比较火热，由于单目视频缺少深度信息，将其他相关信息引入到深度估计是一个可行的办法. 随着相关领域的进一步发展，这种类型的优化方法体现了巨大的潜力，等待着学者们提出更优秀的方法.

4.6　使用生成式对抗网络的方法

与监督学习方法相同，基于无监督的深度估计同样可以视为图像生成问题，生成式对抗网络也被引入到无监督学习中. 但是由于无监督任务中没有真实深度值标签，我们不能简单地将深度估计网络替换成生成器. 此类方法[56,168-171]的一个通用的框架如图 13 所示，模型中的生成器包含一个姿态估计网

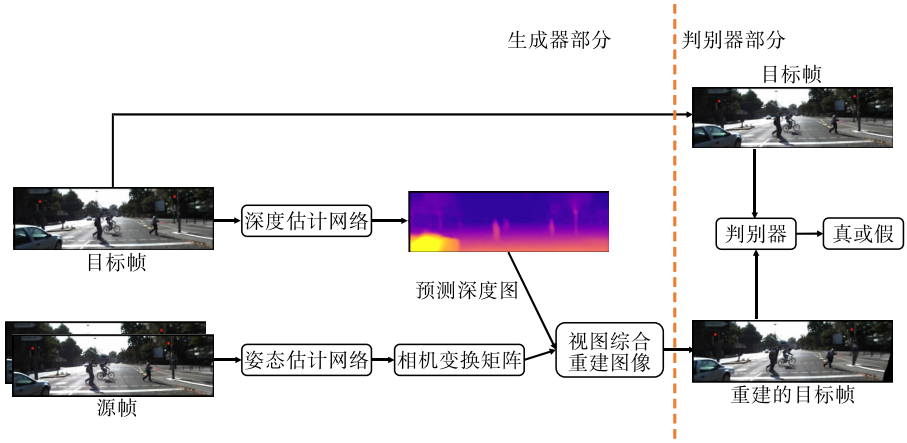


图 13　生成式对抗网络无监督深度估计方法[56]

络和一个深度估计网络，他们的结果用来通过视图重建生成图像，判别器则用来区分生成的重建图像和真实图像.通过对抗损失，生成器产生的重建图像逐渐逼近真实图像，深度估计结果更精确.考虑到视频是连续的，上下文的信息可能也会有助于深度估计，Li 等人[56]在生成式对抗网络的框架上将 LSTM 模块引入至姿态估计网络和深度估计网络获取上下文信息.他们的实验结果表明视频上下文蕴含的信息可以帮助估计深度，如何充分利用上下文信息、增强网络的抗遗忘性[57]是一个可以进一步研究的方向.

与有监督方法类似，生成式对抗网络也被应用于域迁移任务，但相比于有监督方法，无监督方法框架下的域迁移方法数量较少，我们在此主要对以下两个较新的方法进行介绍：Vankadari 等人[172]提出使用生成式对抗网络对在无监督方法框架下训练好的特征提取器进行域自适应，主要解决的问题是夜间场景的深度估计，将在白天场景训练的模型应用于夜间.Cheng 等人[173]解决的问题则和有监督域迁移的背景相同，即在生成数据上进行训练，将模型迁移到真实域上.他们提出基于语义一致性的迁移方法，提取不同类型场景下相同的语义信息对模型的迁移提供引导.

这类方法基于生成式对抗网络，灵感来源于无监督方法中生成式对抗网络产生的良好效果.由于受限于没有真实深度图作为监督信号，基于生成式对抗网络的无监督学习深度估计方法数量不多，能否有一种更巧妙地使用生成式对抗网络的模型架构，值得我们进一步思考.

从另一个角度来看，使用生成式对抗网络的方法也比较灵活，它可以结合其他优化手段，比如将语义分割、光流等相关信息引入基于生成式对抗网络、将视觉里程计引入生成式对抗网络框架[56]、为生成式对抗网络设计掩膜[148]等，都有一定的可行性.此外，由于生成式对抗网络具有不稳定性，增强模型的鲁棒性是此方向发展的一个关键点.

4.7 面向实时与轻量级的方法

由于深度估计在自动驾驶、移动端测距等任务下对模型有推理速度、规模量级的限制要求，在这些特定场景下，我们更关注于如何获得实时性更好、规模更小的深度估计模型.

深度学习本身具有规模大、数据多、效果好的特点，实时、轻量级与优秀的性能通常难以兼得.学者们尝试了许多方法克服深度学习的局限性.由于此类方法关注于网络结构、数量较少、关键思想与训练方式关联较小，我们将一些基于有监督和无监督的方法在此处一并进行总结.

在有监督方法中，比较有代表性的是文献[134]中提出的应用于嵌入式系统的神经网络模型.他们使用轻量级的 MobileNet 模型作为编码器，极大地减少了模型的复杂性.为了进一步减轻嵌入式系统的计算负担，他们在部署时采用了 NetAdapt 算法[182]对模型进行了剪枝.他们的算法可达 175 帧率.前文中详细介绍的 DORN[51]因为将回归任务转化为了分类任务，这种手段也提高了模型的推理速度.

在无监督方法中，Liu 等人[167]提出了使用一个 RNN 单元循环提取特征的方法来减小模型规模(无监督框架)，但相应的问题是推理时间的增加，他们的模型大小只有 0.217 M，相比于普通模型动辄 20 M 以上的规模，已经提升颇多，我们将一些经典模型的规模推理速度、准确率整理在表 6 中，希望给予读者更为直观的说明.Zhang 等人[57]则是使用了一个元学习算法，加快模型的迭代速度，通过在线训练的方式提升模型的实时性.

表 6 模型速度对比

方法	时间	帧率	*RMSE*/m	数据集	训练方式
Eigen et al.[9]	23	43	0.907	NYU	有监督
Eigen et al.[38]	96	10	0.753	NYU	有监督
Laina et al.[41]	237	6	0.604	NYU	有监督
Wofk et al.[134]	5.6	178	0.604	NYU	有监督
Zhou et al.[139]	19	52	6.856	KITTI	无监督
Yin et al.[158]	19	52	5.857	KITTI	无监督
Wang et al.[149]	18	55	5.583	KITTI	无监督
Godard et al.[137]	16	63	5.927	KITTI	无监督
Casser et al.[166]	19	52	4.750	KITTI	无监督
Godard et al.[140]	16	63	4.863	KITTI	无监督
Liu et al.[167]	12	81	4.483	KITTI	无监督

随着深度学习的进一步发展，模型的落地与部署越来越受人关注，对于深度估计，多数场景下，一个关键的要求就是模型的实时性与轻量级.受深度模型的特点影响，此类方法进展缓慢，目前只有较少的模型能够达到部署的要求，这是一个迫切需要被深入研究的方向.

4.8 无监督方法小结

无监督方法不使用真实深度值监督模型训练，学者们提出了使用图像对(4.1 节)与使用视频(4.2 节)的两种基本的无监督训练方法，图像对与视频帧之间的几何约束作为监督信息引导模型的训练.相比较之下，基于视频帧的方法无须已知相机的参数，是如今研究的热点.

由于没有真实深度值参与训练，无监督方法的准确率难以超过有监督学习，提升模型深度估计的准确率是研究人员们关注的重点.无监督领域的方

法主要通过引入视觉里程计(4.4 节)、引入辅助信息(4.5 节)提高模型的准确率.此外,由于基于无监督学习的方法存在着它的特殊性,一些违背静态假设的像素会对模型训练起到负面效果,一些基于掩膜的方法(4.3 节)被提出,这类方法用来提高训练的有效性,进而提高模型的准确率.类似有监督学习,生成式对抗网络(4.6 节)也被引入无监督深度估计领域,同样取得了可观的效果.值得关注的是,面向实时与轻量级的方法(4.7 节)是目前研究要克服的重点问题,模型的实时性和是否轻量级直接影响到深度估计方法的应用.

表 7 对现有的无监督方法进行分类,与有监督方法类似,此类方法也呈现出多种优化手段融合的趋势.更具体地,表 8 对每个无监督方法的主要贡献

表 7　无监督方法分类

改进方式		方法
训练方法	使用图像对训练的方法	[136,137,174,175]
	使用视频训练的方法	[55-57,117,118,139-141,143-149,151,153,154,158-166,168,172,173,176-181]
优化方法	基于可解释掩膜的方法	[140,141,143,144,146-148]
	基于视觉里程计的方法	[55,56,116,117,144,149,151,154,155]
	引入辅助信息的方法	[55,56,117,146-148,154,155,158-166,181]
	使用生成式对抗网络的方法	[56,116,117,155,168,172,173]
	面向实时与轻量级的方法	[57,134,167]

表 8　无监督方法总结

方法	提出年份	主要贡献	来源	KITTI
Garg et al.[136]	2016	使用图像对训练的基本框架	ECCV	5.285
Godard et al.[137]	2017	引入左右一致性	ICCV	5.093
Zhou et al.[139]	2017	使用视频训练的基本框架	ICCV	6.709
Yang et al.[182]	2018	扩展的深度网络	硕士论文	—
Mahjourian et al.[141]	2018	掩膜网络、引入几何约束	CVPR	5.912
Wang et al.[149]	2018	直接视觉测距	CVPR	5.583
Zhan et al.[154]	2018	联合视觉里程计	CVPR	5.585
Yin et al.[158]	2018	多任务框架	CVPR	5.857
Zou et al.[159]	2018	多任务框架	ECCV	—
Kumar et al.[168]	2018	GAN	ECCV	4.756
Ma et al.[178]	2019	残差稠密模块	小型微型计算机系统	4.797
Godard et al.[140]	2019	最小重建损失、前向计算掩膜	ICCV	4.863
Bian et al.[143]	2019	自发现掩膜	NIPS	5.439
Wang et al.[144]	2019	重叠掩膜、空白掩膜、精细掩膜	ICRA	4.290
Wang et al.[147]	2019	多任务框架、联合视觉里程计	CVPR	3.404
Li et al.[56]	2019	联合视觉里程计、GAN、LSTM	ICCV	5.564
Almalioglu et al.[155]	2019	联合视觉里程计、GAN	ICRA	5.448
Wang et al.[55]	2019	联合视觉里程计、RNN	CVPR	—
Feng et al.[116]	2019	联合视觉里程计、stacked GAN	RAL	4.003
Casser et al.[166]	2019	引入物体运动信息	AAAI	4.750
Chen et al.[165]	2019	相机内参估计的多任务框架	ICCV	4.743
Zhao et al.[174]	2020	特征金字塔	激光与光电子学进展	5.331
Ding et al.[177]	2020	金字塔结构	光学学报	5.562
Wang et al.[146]	2020	多任务框架、低平均值掩膜	ArXiv	5.255
Zhao et al.[148]	2020	尺度一致性、masked GAN	ArXiv	5.536
Dai et al.[151]	2020	ORB-SLAM	激光与光电子学进展	6.432
Hur et al.[160]	2020	多任务框架	ArXiv	4.877
Guizilini et al.[181]	2020	使用语义分割信息引导深度估计	ArXiv	4.381
Guizilini et al.[161]	2020	PackNet、引入瞬时速度	CVPR	4.601
Wang et al.[163]	2020	使用语义分割信息引导深度估计	计算机工程与应用	5.198
Liu et al.[162]	2020	使用视觉语义引导深度估计	硕士论文	—
Cen et al.[179]	2020	引入双注意力机制	广东工业大学学报	4.731
Zhang et al.[57]	2020	元学习、在线深度估计	CVPR	5.658
Johnston et al.[180]	2020	注意力机制、离散视差	CVPR	4.699
Klingner et al.[164]	2020	语义分割	ECCV	4.693
Vankadari et al.[172]	2020	夜间图像域迁移	ECCV	—
Cheng et al.[173]	2020	语义信息引导域迁移	ECCV	4.981
Liu et al.[167]	2020	轻量级网络	ISPRS	5.264
Zhang et al.[57]	2020	在线学习	CVPR	0.867
Watson et al.[176]	2021	多帧进行深度估计	CVPR	4.261

按时间顺序进行总结.同样地,我们也提供了每种方法的客观性能指标(*RMSE*).相比于有监督方法,无监督方法更多在 KITTI 数据集(截断到 80 m)上进行训练、测试,故此类方法更适用于室外场景.注意一些方法提供的是截断到 50 m 的结果,没有可比性,在表中用"—"表示.

5 半监督学习方法

由于监督学习获取真实标签成本过高,无监督学习又缺乏强有力的监督信号,学者们提出了半监督学习方法,希望能够取长补短,利用少量带有真实标签与大量不带有真实标签的数据训练模型,得到深度估计能力更强的模型.

5.1 使用图像对训练的方法

在文献[137]提出的框架基础上,Kuznietsov 等人[183]将传感器得到的稀疏深度图引入模型作为额外的监督信号,实现了监督学习与半监督方法的结合.他们的主要贡献是提出了如下形式的损失函数:

$$L_\theta(I_l,I_r,Z_l,Z_r)=\lambda_t L_\theta^S(I_l,I_r,Z_l,Z_r)+\gamma L_\theta^U(I_l,I_r)+L_\theta^R(I_l,I_r) \tag{24}$$

$$L_\theta^S=\sum_{x\in\Omega_Z}\|\rho_\theta x^{-1}-Z(x)\|_\delta \tag{25}$$

式(24)中 λ_t 和 γ 是权衡监督学习损失 L_θ^S 与无监督损失学习损失 L_θ^U 的权重,L_θ^R 是正则化损失.θ 是预测逆深度 ρ_θ 的神经网络的参数.式(25)是有监督损失项的计算方式,$\|\cdot\|_\delta$ 是式(8)中描述的 berHu 损失,在训练时,他们将 δ 也就是式(8)中的 c 设置为 $0.2\max_{x\in\Omega_Z}(|\rho(x)^{-1}-Z(x)|)$.这项有监督学习损失将真实深度信息作为监督信号引入,进一步提高了模型的深度估计能力.无监督损失项和正则项沿用了文献[137]中的形式,不同的是他们在无监督项中额外添加了高斯平滑处理 G,并用实验证明了使用平滑后的图像作为监督信号能够进一步改善模型的训练效果.

他们提出的模型框架与训练流程如图 14 所示,通过与图 8 所示的无监督学习框架对比,可以进一步体会此类工作的主要贡献.在文献[183]的工作基础之上,Amiri 等人[184]引入了文献[137]中的左右一致性损失,获得了更准确的深度估计模型.为了减少输入数据的限制,文献[185]提出使用 Deep3D[186]模型生成输入的左视图对应的右视图,再将左右视图输入神经网络(Dispnet[26])进行深度估计的方法.由于 Deep3D 模型的训练属于无监督学习范畴,Dispnet 网络的训练属于有监督学习范畴,故他们提出的方法可以归类于半监督方法.

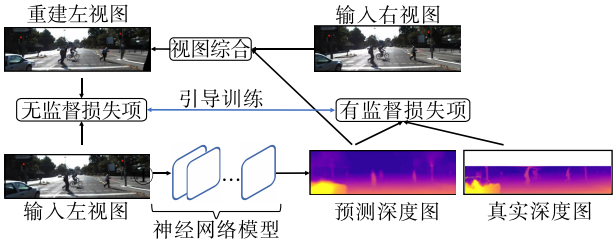


图 14 使用图像对训练的半监督学习方法框架[183]

使用图片对训练的半监督学习方法主要基于无监督学习框架,在损失函数中增添有监督损失项,将真实深度信息引入,引导模型训练.

5.2 使用视频训练的方法

与使用图像对进行训练的半监督学习方法思路类似,使用视频进行训练的半监督方法主要在对应的无监督学习框架上进行修改,将真实深度图作为监督信息引入,通过使用更强大的监督信号进行训练,获得更准确的深度估计模型.

一个较基础的框架如图 15 所示,通过与图 11 中所示使用视频训练的无监督方法框架进行对比,我们可以归纳出此类方法的主要贡献是引入了虚线框所示的有监督损失项.所以,与使用图像对进行训练的方法类似,此类方法主要通过修改损失函数,引入有监督损失项实现有监督与无监督的融合.

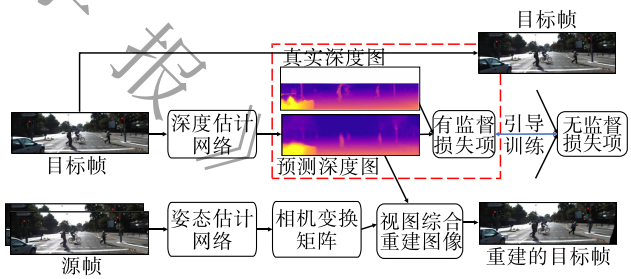


图 15 使用视频训练的半监督学习方法框架[187]

顺应此思路,文献[187]将真实深度信息作为监督信号引入,实现监督学习与无监督学习的结合,提出如下形式的损失函数:

$$L(I_t,S)=L_{photo}\otimes M_{photo}+\lambda_{smooth}\times L_{smooth}+\lambda_{rep}\times L_{rep} \tag{26}$$

$$L_{rep}(\hat{D}_t,D_t)=\frac{1}{V}\sum_{i\in V_t}\|\hat{p}_S^i-p_S^i\|+\frac{1}{V}\sum_{i\in V_t}\|\pi_S(\hat{d}_t^iK^{-1}u_t^i)-\pi_S(d_t^iK^{-1}u_t^i)\| \tag{27}$$

式(26)中无监督损失项 L_{photo} 为文献[139]中损失函数的形式,在训练时,他们使用文献[140]提出的计算重建误差最小值的训练方法和掩膜方法.此

外，此式中平滑项 L_{smooth} 使用了文献[137]提出的梯度平滑误差. 为了更充分地利用姿态转换与深度估计结果，他们提出了式(27)中的重投影距离损失 L_{rep}. 如图 16 所示，p 是三维空间中的一个物体，简单的损失函数将直接计算预测深度 p 与真实深度 d_{gt} 之间的差值，而他们提出的方法则是将预测出的点 $\hat{p}$ 与 p_1 先重投影至源相机 O_s 平面，再计算投影点之间的距离. 与目前有监督学习领域的 berHu 损失(式(8))和有序回归损失(式(9))类似，重投影距离损失也为深度引入了权重，一个直观的解释是图中对 p_1 和 p_2 的预测有着不相等的简单损失，但是拥有相同大小的重投影距离损失，他们的方法对图像中部像素施加了较大的权重，希望模型能够在这个区域得到更准确的深度估计结果. 此外，他们的方法可以直接使用姿态估计网络结果，与前两者式(8)、式(9)相比无须人工设定参数，更容易训练.

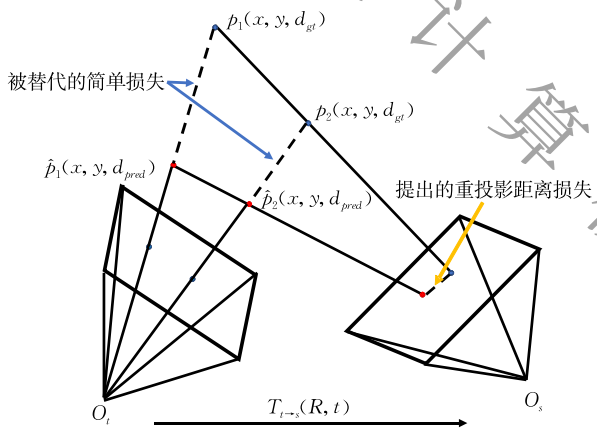


图 16　重投影距离损失示意

使用视频训练的半监督学习方法也主要基于无监督学习框架，在损失函数中增添有监督损失项，将真实深度信息引入，引导模型训练.

5.3　使用生成式对抗网络的方法

类似于有监督方法和无监督方法，生成式对抗网络也被引入至半监督学习领域，完成深度图像生成任务.

Ji 等人[188]最早提出基于 Patch GAN[189]框架的半监督学习方法，他们使用的网络模型如图 17 所示，生成器(Generator)接收没有对应真实深度值的 RGB 图像并给出对应的预测深度图，预测的深度图和数据中原有的真实深度图送入深度判别器(Depth Discriminator)中进行判别，预测深度图与原 RGB 图像连接起来、真实深度图与对应 RGB 图像连接起来共同送入图像对判别器(Pair Discriminator)中进行判别.

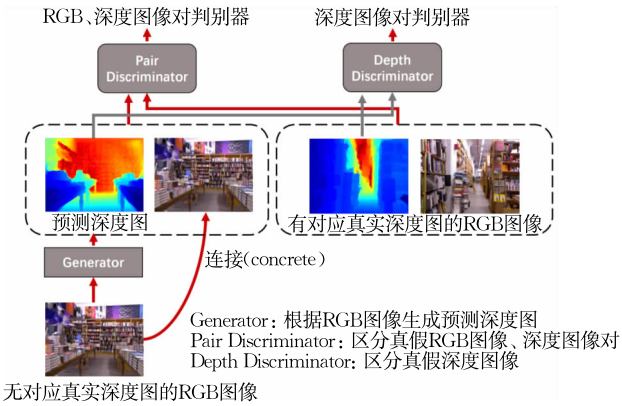


图 17　Ji 等人[188]提出的网络框架

通过生成器与判别器之间的对抗损失对模型进行训练，可以得到有深度估计能力的生成器. 因为模型中同时有一个生成器、两个判别器，所以整体的训练可以归结为一个三元最小最大化过程：

$$\min_G \max_{PD,DD} V(G, PD, DD) \tag{28}$$

其中 G、PD 和 DD 分别代表生成器、图像对判别器和深度图判别器，V 代表各组对抗期望. 训练结束后，生成器可以直接给出输入 RGB 图像对应的深度估计结果.

通过总结使用生成式对抗网络的监督学习和无监督学习方法，我们可以推断，将语义信息融入至生成式对抗网络模型、使用生成数据进行训练都是此类方法可行的发展方向.

5.4　引入语义信息的方法

此类方法使用多任务框架将语义信息引入深度估计模型. 借助包含丰富边缘信息的语义分割结果，模型可以对物体边缘做出更准确的深度估计. 此外，语义分割模型同时可以划分出动态物体与静态物体，也可以帮助模型减小动态物体造成的影响. 多任务框架中的语义分割任务带有真实标签，属于有监督学习范畴，深度估计任务不带有真实标签，属于无监督学习范畴，此类方法将两者结合，被归类于半监督学习方法.

Ramirez 等人[190]最早提出基于半监督学习的多任务框架，他们的方法在使用图像对进行训练的无监督学习方法的基础上引入语义分割任务，这两个任务共享相同的底层特征. 他们在有监督部分使用语义分割任务中常见的损失函数形式：

$$L_s = H(p_t, \hat{p}_t) + KL(p_t, \hat{p}_t) \tag{29}$$

其中 p_t 代表真实的语义分割结果，$\hat{p}_t$ 代表预测的语义分析结果，H 与 KL 分别代表信息熵与 KL 散度. 此外，他们还设计了一个跨域不连续项损失(Cross-Domain Discontinuity Term)：

$$L_{sdd}=\frac{1}{N}\sum_{i,j}\mathrm{sgn}(|\delta_x sem^l_{i,j}|)e^{-\left\|\frac{\delta_x d^l_{i,j}}{d^l_{i,j}}\right\|}+\mathrm{sgn}(|\delta_y sem^l_{i,j}|)e^{-\left\|\frac{\delta_y d^l_{i,j}}{d^l_{i,j}}\right\|} \tag{30}$$

其中 sem 代表真实的语义分割结果图像，d 表示预测的视差图. 类比式(18)，这个跨域不连续损失目的是在提高语义标签变化处深度图的梯度，因为语义分割边缘一般也是物体边缘，所以通过此项，可以进一步提高深度估计模型预测物体边缘深度的能力.

此外，为了进一步使用语义分割信息引导深度估计，他们提出了语义引导的视差平滑误差：

$$L_s=\|d-f_{\rightarrow}(d)\|\otimes(1-\|\varphi(s)-f_{\rightarrow}(\varphi(s))\|) \tag{31}$$

其中，s 与 d 分别表示语义分割结果与深度估计结果，φ 表示最大值置 1 其余值置 0，$f_{\rightarrow}$ 表示将图片移动一个像素距离. $\otimes$表示像素点乘. 与式(30)描述的跨域不连续损失相反，式(31)中的语义分割信息引导损失关注于语义标签相同的物体区域，希望预测的深度值在相同物体区域平滑.

此类方法主要以使用图片对进行训练的无监督学习方法为基础，引入额外的语义信息作为监督信号. 类比有监督学习和无监督学习，引入更多的相关信息辅助深度估计是一个可行的发展方向，此外，将这种框架引入至使用视频训练的无监督框架中也有一定的可行性.

5.5 半监督学习方法小结

在无监督方法的基础之上，借助部分真实标签提供的监督信号，半监督学习方法一定程度上地解决了无监督方法中存在的尺度模糊问题，获得了准确率高于无监督学习方法的深度估计结果.

目前基于半监督学习的方法数量较少，有较大的发展空间. 目前大部分的有监督和无监督方法会使用 KITTI 数据集进行训练并进行测试，KITTI 数据集中的真实 LIDAR 深度值服务于有监督损失，视频服务于无监督损失. 从此角度观之，半监督学习相当于两者的叠加，由于有监督学习和无监督学习的提升空间仍然很大，半监督方法得到的提升有限，受到的关注较小. 半监督方法的另一个思考角度是使用一个较为庞大的无标签数据集进行无监督训练，在有标签数据上进行有监督训练，两者相辅相成，各有提升. 但是由于相机参数等因素对模型的巨大影响，数据集间的统计差异会影响模型的性能，尽管目前已有一些域迁移方法尝试克服此困难，但仍然有可观的进步空间. 此外，评测标准尚不一致，目前仍缺少统一的基准线(benchmark)对整体的框架流程进行统一，对于半监督方法，使用多少数据进行有监督训练、多少数据进行无监督训练，需要严格规定以保证公平的对比.

因为此类方法基于无监督框架，所以也主要分为使用图像对(5.1 节)和使用视频(5.2 节)进行训练的两种方法. 类似地，使用生成式对抗网络(5.3 节)的优化方法. 此类方法的数量远小于前两类. 多数方法从无监督框架出发，在部分有标签数据基础上增加监督项损失.

表 9 对现有的半监督方法进行分类. 更具体地，我们在表 10 中对每个半监督方法的主要贡献按时间顺序进行总结. 同样地，我们也提供了每种方法的客观性能指标($RMSE$). 相比于有监督方法，无监督方法更多在 KITTI 数据集(截断到 80 m)上进行训练、测试，故此类方法更适用于室外场景.

表 9 半监督方法分类

	改进方式	方法
训练方法	使用图像对训练的方法	[183-185]
	使用视频训练的方法	[187]
优化方法	使用生成式对抗网络的方法	[188]
	引入语义信息的方法	[190,191]

表 10 无监督方法总结

方法	提出年份	主要贡献	来源	KITTI
Kuznietsov et al.[183]	2017	使用图像对训练的半监督框架	CVPR	4.621
Ramirez et al.[190]	2018	语义信息	ACCV	6.526
Luo et al.[185]	2018	Deep3D、Dispnet	ICCV	4.252
Ji et al.[188]	2019	patch GAN	TPAMI	4.405
Amiri et al.[184]	2019	图像对的左右一致性约束	ROBIO	3.464
Guizilini et al.[187]	2020	使用视频训练的半监督框架、重投影距离损失	PMLR	3.097
Wimbauer et al.[191]	2021	语义信息、视觉里程计真值	CVPR	4.261

6 总结与展望

在本篇综述中，我们根据不同的训练方式总结了基于深度学习的单目深度估计方法，讨论了各种现有方法的特点与改进思路. 经过总结，我们发现单目深度估计问题历史悠久，存在很多经典算法，而基于深度学习的方法在短短的几年内就大放异彩，这

体现了深度神经网络强大的特征表达与学习能力.最后,本节总结基于深度学习的单目深度估计领域的发展趋势与难点问题:

(1)以往许多工作主要关注于使用不同的网络架构提升精度,一些经典的结构比如全卷积网络(FCN)、深度残差网络(ResNet)都展示出了强大的深度估计能力.因此,随着深度神经网络的发展,不断尝试使用新的网络结构是自然的发展趋势.随着新的网络结构与损失函数的提出,超参数的选择成为了巨大的挑战,直接使用神经网络对超参数进行估计是一个可行的方法[92].

(2)由于语义分割信息包含明确的边界信息,其常用来引导深度估计,如何更充分地使用语义分割,如何将其更有效地引入语义分割信息是一个研究方向.顺承这个思路,一些其他任务包含了有助于深度估计的信息,使用多任务框架联合估计深度的方法也有很大的发展空间.

(3)现在的方法大多采用离线学习的训练方式,即事先对模型进行训练,训练完成后固定模型并开始预测.那么,就必然存在训练数据与要预测的数据之间有域偏差的问题,域偏差将导致模型的泛化能力大大降低.引入相机参数估计和使用域自适应的方法可以缓解该问题,是目前研究的一大方向.此外,从另一个角度出发,如 Zhang 等人[57]人提出结合元学习实现在线学习的方法,也有很大的开发潜力.

(4)因为有监督学习依赖大量真实深度数据,所以发展更大规模的数据集也有重大意义.此外,因为生成数据数量较多、更易获得,通过提升模型的泛化能力,将使用生成数据训练的模型应用于实际也是解决真实数据匮乏的方法.

(5)深度估计模型被广泛应用于自治系统,用于感知周围环境的三维信息.因为一些系统可以通过硬件设备感知部分场景的深度信息,所以根据稀疏真实值补全深度图是一个很有价值的研究方向.此外,一些系统要求深度估计要有足够的鲁棒性,所以如何处理带噪的输入图像也是深度估计领域的一个关键问题.

(6)深度神经网络在嵌入式设备上的实时运行能力将对其应用产生巨大影响,而往往神经网络要求更多的时间进行计算,故轻量级深度估计网络也是一个发展方向,目前对此方向的研究不多,但很有现实意义.

(7)深度估计面对的一大难点是动态物体和遮挡问题的处理.动态物体和遮挡问题会打破基于视频的单目深度估计方法的静态假设,进而影响深度估计效果,在 3.3 节中我们讨论了部分现有的解决方法,虽然它们取得了一定的进展,但是如何进一步优化此类像素位置的深度仍有很多工作要做.

(8)另一个困难是获得高分辨率的深度图输出.当前方法得到的输出结果的分辨率较低且受限于数据集的质量,难以满足增强现实(AR)等任务的需要,将深度估计的结果进行超分辨率可能是一个解决方法,但如何直接通过深度估计模型获得高分辨率图像仍是研究人员面对的一大问题.

(9)此外,模型的可解释性也很关键.基于深度学习的方法普遍有黑盒特性,缺乏可解释性.文献[26,60]提出神经网络主要从边缘信息估计深度,这与人的感知并不完全相似.进一步挖掘深度估计的可解释性能够帮助提升模型的准确率、泛化能力.如何解释网络模型的工作过程、工作原理,也等待着研究人员的进一步挖掘.

参　考　文　献

[1] Izadi S, Kim D, Hilliges O, et al. KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera//Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology. Santa Barbara, USA, 2011: 559-568

[2] Labayrade R, Aubert D, Tarel J P, et al. Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation//Proceedings of the Intelligent Vehicle Symposium. Versailles, France, 2002: 646-651

[3] Gupta S, Davidson J, Levine S, et al. Cognitive mapping and planning for visual navigation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2616-2625

[4] Ullman S. The interpretation of structure from motion. Proceedings of the Royal Society of London. Series B, Biological Sciences, 1979, 203(1153): 405-426

[5] Scharstein D, Szeliski R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision, 2002, 47(1): 7-42

[6] Marr D, Poggio T. A computational theory of human stereo vision. Proceedings of the Royal Society of London. Series B, Biological Sciences, 1979, 204(1156): 301-328

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

[8] Gu C J, Cong Y, Sun G. Three birds, one stone: Unified laser-based 3-D reconstruction across different media. IEEE Transactions on Instrumentation and Measurement, 2020, 70: 1-12

[9] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 2366-2374

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

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

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

[13] Qin H T, Gong R H, Liu X L, et al. Forward and backward information retention for accurate binary neural networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 2250-2259

[14] Amodei D, Ananthanarayanan S, Anubhai R, et al. Deep speech 2: End-to-end speech recognition in English and mandarin//Proceedings of the International Conference on Machine Learning. New York City, USA, 2016: 173-182

[15] Young T, Hazarika D, Poria S. Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine, 2018, 13(3): 55-75

[16] Bi Tian-Teng, Liu Yue, Weng Dong-Dong, et al. Survey on supervised learning based depth estimation from a single image. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(8): 1383-1393(in Chinese)
(毕天腾, 刘越, 翁冬冬等. 基于监督学习的单幅图像深度估计综述. 计算机辅助设计与图形学学报, 2018, 30(8): 1383-1393)

[17] Bhoi A. Monocular depth estimation: A survey. arXiv preprint arXiv:1901.09402, 2019

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

[19] Huang Jun, Wang Cong, Liu Yue, et al. The progress of monocular depth estimation technology. Journal of Image and Graphics, 2019, 24(12): 2081-2097(in Chinese)
(黄军, 王聪, 刘越等. 单目深度估计技术进展综述. 中国图象图形学报, 2019, 24(12): 2081-2097)

[20] Guo Ji-Feng, Bai Cheng-Chao, Guo Shuang. A review of monocular depth estimation based on deep learning. Unmanned Systems Technology, 2019, 2(2): 12-21(in Chinese)
(郭继峰, 白成超, 郭爽. 基于深度学习的单目视觉深度估计研究综述. 无人系统技术, 2019, 2(2): 12-21)

[21] Silberman N, Fergus R. Indoor scene segmentation using a structured light sensor//Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops. Barcelona, Spain, 2011: 601-608

[22] Saxena A, Sun M, Ng A Y. Make3D: Learning 3D scene structure from a single still image. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 31(5): 824-840

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

[24] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 3213-3223

[25] Li Z, Snavely N. MegaDepth: Learning single-view depth prediction from internet photos//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, 2018: 2041-2050

[26] Mayer N, Ilg E, Hausser P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 4040-4048

[27] Vasiljevic I, Kolkin N, Zhang S, et al. DIODE: A dense indoor and outdoor depth dataset. arXiv preprint arXiv: 1908.00463, 2019

[28] Hua Y, Kohli P, Uplavikar P, et al. Holopix50k: A large-scale in-the-wild stereo image dataset. arXiv preprint arXiv: 2003.11172, 2020

[29] Tan J, Lin W, Chang A X, et al. Mirror3D: Depth refinement for mirror surfaces//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 15990-15999

[30] Cruz L, Lucio D, Velho L. Kinect and RGBD images: Challenges and applications//Proceedings of the 25th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials. Ouro Preto, Brazil, 2012: 36-49

[31] Pavlakos G, Zhou X, Derpanis K G, et al. Coarse-to-fine volumetric prediction for single-image 3D human pose//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 7025-7034

[32] Jiao J, Cao Y, Song Y, et al. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 53-69

[33] Bao W, Lai W S, Ma C, et al. Depth-aware video frame interpolation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3703-3712

[34] Tateno K, Tombari F, Laina I, et al. CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 6243-6252

[35] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 2012, 25: 1097-1105

[36] Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv preprint arXiv: 1602.07360, 2016

[37] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014

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

[39] Chen W, Fu Z, Yang D, et al. Single-image depth perception in the wild. arXiv preprint arXiv:1604.03901, 2016

[40] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[41] Laina I, Rupprecht C, Belagiannis V, et al. Deeper depth prediction with fully convolutional residual networks// Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV). Stanford, USA, 2016: 239-248

[42] Iandola F, Moskewicz M, Karayev S, et al. DenseNet: Implementing efficient ConvNet descriptor pyramids. arXiv preprint arXiv:1404.1869, 2014

[43] He L, Wang G, Hu Z. Learning depth from single images with deep neural network embedding focal length. IEEE Transactions on Image Processing, 2018, 27(9): 4676-4689

[44] Yuan Jian-Zhong, Zhou Wu-Jie, Pan Ting, et al. Road scene depth estimation based on deep convolutional neural networks. Laser & Optoelectronics Progress, 2019, 56(8): 171-179(in Chinese)
(袁建中，周武杰，潘婷等. 基于深度卷积神经网络的道路场景深度估计. 激光与光电子学进展，2019，56(8)：171-179)

[45] Dai J, Li Y, He K, et al. R-FCN: Object detection via region-based fully convolutional networks. arXiv preprint arXiv:1605.06409, 2016

[46] Shelhamer E, Barron J T, Darrell T. Scene intrinsics and depth from a single image//Proceedings of the IEEE International Conference on Computer Vision Workshops. Santiago, Chile, 2015: 37-44

[47] Mancini M, Costante G, Valigi P, et al. Fast robust monocular depth estimation for obstacle detection with fully convolutional networks//Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Daejeon, Korea, 2016: 4296-4303

[48] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122, 2015

[49] Li B, Dai Y, He M. Monocular depth estimation with hierarchical fusion of dilated CNNs and soft-weighted-sum inference. Pattern Recognition, 2018, 83: 328-339

[50] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916

[51] Fu H, Gong M, Wang C, et al. Deep ordinal regression network for monocular depth estimation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 2002-2011

[52] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks//Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada, 2013: 6645-6649

[53] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[54] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 1126-1135

[55] Wang R, Pizer S M, Frahm J M. Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5555-5564

[56] Li S, Xue F, Wang X, et al. Sequential adversarial learning for self-supervised deep visual odometry//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 2851-2860

[57] Zhang Z, Lathuilière S, Ricci E, et al. Online depth learning against forgetting in monocular videos//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 4494-4503

[58] Xia Z, Sullivan P, Chakrabarti A. Generating and exploiting probabilistic monocular depth estimates//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 65-74

[59] Jaritz M, De Charette R, Wirbel E, et al. Sparse and dense data with CNNs: Depth completion and semantic segmentation //Proceedings of the 2018 International Conference on 3D Vision (3DV). Verona, Italy, 2018: 52-60

[60] Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models. Advances in Neural Information Processing Systems, 2015, 28: 3483-3491

[61] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015

[62] Zoph B, Vasudevan V, Shlens J, et al. Learning transferable architectures for scalable image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, 2018: 8697-8710

[63] Pilzer A, Lathuiliere S, Sebe N, et al. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 9768-9777

[64] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv preprint arXiv:1706.03762, 2017

[65] Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction. arXiv preprint, 2103.13413, 2021

[66] Bhat S F, Alhashim I, Wonka P. AdaBins: Depth estimation using adaptive bins//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 4009-4018

[67] Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017

[68] Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition//Proceedings of the ICML Deep Learning Workshop: Volume 2. Sydney, Australia, 2015

[69] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules. arXiv preprint arXiv:1710.09829, 2017

[70] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828

[71] Wang L, Zhang J, Wang O, et al. SDC-depth: Semantic divide-and-conquer network for monocular depth estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 541-550

[72] Liu J, Wang Y, Li Y, et al. Collaborative deconvolutional neural networks for joint depth estimation and semantic segmentation. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(11): 5655-5666

[73] Xu D, Ouyang W, Wang X, et al. PAD-Net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, 2018: 675-684

[74] Ye J, Ji Y, Wang X, et al. Student becoming the master: Knowledge amalgamation for joint scene parsing, depth estimation, and more//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 2829-2838

[75] Chen L, Yang Z, Ma J, et al. Driving scene perception network: Real-time joint detection, depth estimation and semantic segmentation//Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe, USA, 2018: 1283-1291

[76] Qi X, Liao R, Liu Z, et al. GeoNet: Geometric neural network for joint depth and surface normal estimation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, 2018: 283-291

[77] Yin W, Liu Y, Shen C, et al. Enforcing geometric constraints of virtual normal for depth prediction//Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach, USA, 2019: 5684-5693

[78] Palafox P R, Betz J, Nobis F, et al. SemanticDepth: Fusing semantic segmentation and monocular depth estimation for enabling autonomous driving in roads without lane lines. Sensors, 2019, 19(14): 3224

[79] Zhang Z, Cui Z, Xu C, et al. Joint task-recursive learning for semantic segmentation and depth estimation//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 235-251

[80] Nekrasov V, Dharmasiri T, Spek A, et al. Real-time joint semantic segmentation and depth estimation using asymmetric annotations//Proceedings of the 2019 International Conference on Robotics and Automation (ICRA). Montreal, Canada, 2019: 7101-7107

[81] Lin Z, Cohen S D, Wang P, et al. Joint depth estimation and semantic segmentation from a single image. Google Patents, 2018

[82] Ochs M, Kretz A, Mester R. SDNet: Semantically guided depth estimation network//Proceedings of the German Conference on Pattern Recognition. Dortmund, Germany, 2019: 288-302

[83] Atapour-Abarghouei A, Breckon T P. Monocular segment-wise depth: Monocular depth estimation based on a semantic segmentation prior//Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP). Taipei, China, 2019: 4295-4299

[84] Yan H, Zhang S, Zhang Y, et al. Monocular depth estimation with guidance of surface normal map. Neurocomputing, 2018, 280: 86-100

[85] Srinivasan P P, Garg R, Wadhwa N, et al. Aperture supervision for monocular depth estimation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, 2018: 6393-6401

[86] Chang J, Wetzstein G. Deep optics for monocular depth estimation and 3D object detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach, USA, 2019: 10193-10202

[87] Miangoleh S M H, Dille S, Mai L, et al. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 9685-9694

[88] Mathew A, Mathew J. Monocular depth estimation with SPN loss. Image and Vision Computing, 2020, 100: 103934

[89] Barron J T. A general and adaptive robust loss function//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4331-4339

[90] Chen Zong-Hai, Hong Yang, Wang Ji-Kai, et al. Monocular visual odometry based on recurrent convolutional neural networks. Robot, 2019, 41(2): 147-155(in Chinese)
(陈宗海, 洪洋, 王纪凯等. 基于循环卷积神经网络的单目视觉里程计. 机器人, 2019, 41(2): 147-155)

[91] Wang L, Zhang J, Wang Y, et al. CLIFFNet for monocular depth estimation with hierarchical embedding loss//Proceedings of the European Conference on Computer Vision. Scottish Event Campus, 2020: 316-331

[92] Lee J H, Kim C S. Multi-loss rebalancing algorithm for monocular depth estimation//Proceedings of the 16th European Conference on Computer Vision (ECCV 2020), Part XVII 16, Scottish Event Campus. Glasgow, UK, 2020: 785-801

[93] Zoran D, Isola P, Krishnan D, et al. Learning ordinal relationships for mid-level vision//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 388-396

[94] Xian K, Zhang J, Wang O, et al. Structure-guided ranking loss for single image depth prediction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 611-620

[95] Xue F, Zhuo G, Huang Z, et al. Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications//Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas, USA, 2020: 2330-2337

[96] Cao Y, Wu Z, Shen C. Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28(11): 3174-3182

[97] Huynh L, Nguyen-Ha P, Matas J, et al. Guiding monocular depth estimation using depth-attention volume//Proceedings of the European Conference on Computer Vision. Scottish Event Campus, 2020: 581-597

[98] Saha S, Obukhov A, Paudel D P, et al. Learning to relate depth and semantics for unsupervised domain adaptation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 8197-8207

[99] Lee B U, Lee K, Kweon I S. Depth completion using plane-residual representation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 13916-13925

[100] Diaz R, Marathe A. Soft labels for ordinal regression//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4738-4747

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

[102] Liu F, Shen C, Lin G, et al. Learning depth from single monocular images using deep convolutional neural fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(10): 2024-2039

[103] Li B, Shen C, Dai Y, et al. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1119-1127

[104] Wang P, Shen X, Lin Z, et al. Towards unified depth and semantic prediction from a single image//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 2800-2809

[105] Mousavian A, Pirsiavash H, Košecká J. Joint semantic segmentation and depth estimation with deep convolutional networks//Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV). Stanford, USA, 2016: 611-619

[106] Xu D, Wang W, Tang H, et al. Structured attention guided convolutional neural fields for monocular depth estimation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, 2018: 3917-3925

[107] Ramamonjisoa M, Du Y, Lepetit V. Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 14648-14657

[108] Roy A, Todorovic S. Monocular depth estimation using neural regression forest//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 5506-5514

[109] Wang X, Hou C, Pu L, et al. A depth estimating method from a single image using foe CRF. Multimedia Tools and Applications, 2015, 74(21): 9491-9506

[110] Hua Y, Tian H. Depth estimation with convolutional conditional random field network. Neurocomputing, 2016, 214: 546-554

[111] Mahmood F, Durr N J. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. Medical Image Analysis, 2018, 48: 230-243

[112] Ricci E, Ouyang W, Wang X, et al. Monocular depth estimation using multi-scale continuous CRFs as sequential deep networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(6): 1426-1440

[113] Wang Q, Zhao H, Hu Z, et al. Discrete convolutional CRF networks for depth estimation from monocular infrared images. International Journal of Machine Learning and Cybernetics, 2021, 12(1): 187-200

[114] Goodfellow I, Pouget-abadie J, Mirza M, et al. Generative adversarial nets. Advances in Neural Information Processing Systems. Montreal, Quebec, Canada, 2014: 2672-2680

[115] Jung H, Kim Y, Min D, et al. Depth prediction from a single image with conditional adversarial networks//Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP). Beijing, China, 2017: 1717-1721

[116] Feng T, Gu D. SGANVO: Unsupervised deep visual odometry and depth estimation with stacked generative adversarial networks. IEEE Robotics and Automation Letters, 2019, 4(4): 4431-4437

[117] Li Y, Qian K, Huang T, et al. Depth estimation from monocular image and coarse depth points based on conditional GAN//Proceedings of the MATEC Web of Conferences: Volume 175. Virtual: EDP Sciences, 2018: 03055

[118] Arslan A T, Seke E. Face depth estimation with conditional generative adversarial networks. IEEE Access, 2019, 7: 23222-23231

[119] Li Y, Wang N, Liu J, et al. Demystifying neural style transfer. arXiv preprint arXiv:1701.01036, 2017

[120] Atapour-Abarghouei A, Breckon T P. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, 2018: 2800-2810

[121] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of the IEEE International Conference on Computer Vision. Hawaii Convention Center, USA, 2017: 2223-2232

[122] Chen Y, Li W, Chen X, et al. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1841-1850

[123] Zheng C, Cham T J, Cai J. T2Net: Synthetic-to-realistic translation for solving single-image depth estimation tasks//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 767-783

[124] Guo R, Ayinde B, Sun H, et al. Monocular depth estimation using synthetic images with shadow removal//Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC). Yokohama Auckland, New Zealand, 2019: 1432-1439

[125] Lopez-Rodriguez A, Mikolajczyk K. DESC: Domain adaptation for depth estimation via semantic consistency. arXiv preprint arXiv:2009.01579, 2020

[126] Zhao S, Fu H, Gong M, et al. Geometry-aware symmetric domain adaptation for monocular depth estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 9788-9798

[127] Chen Y C, Lin Y Y, Yang M H, et al. CrDoCo: Pixel-level domain transfer with cross-domain consistency//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1791-1800

[128] Akada H, Bhat S F, Alhashim I, et al. Self-supervised learning of domain invariant features for depth estimation. arXiv preprint arXiv:2106.02594, 2021

[129] Liao Y, Huang L, Wang Y, et al. Parse geometry from a line: Monocular depth estimation with partial laser observation //Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA). Toulon, France, 2017: 5059-5066

[130] Chen Z, Badrinarayanan V, Drozdov G, et al. Estimating depth from RGB and sparse sensing//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 167-182

[131] Guizilini V, Ambrus R, Burgard W, et al. Sparse auxiliary networks for unified monocular depth prediction and completion//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 11078-11088

[132] Imran S, Liu X, Morris D. Depth completion with twin surface extrapolation at occlusion boundaries//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 2583-2592

[133] Wang T H, Wang F E, Lin J T, et al. Plug-and-Play: Improve depth prediction via sparse data propagation//Proceedings of the 2019 International Conference on Robotics and Automation (ICRA). Montreal, Canada, 2019: 5880-5886

[134] Wofk D, Ma F, Yang T J, et al. FastDepth: Fast monocular depth estimation on embedded systems//Proceedings of the 2019 International Conference on Robotics and Automation (ICRA). Montreal, Canada, 2019: 6101-6108

[135] Eldesokey A, Felsberg M, Holmquist K, et al. Uncertainty-aware CNNs for depth completion: Uncertainty from beginning to end//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 12014-12023

[136] Garg R, Bg V K, Carneiro G, et al. Unsupervised CNN for single view depth estimation: Geometry to the rescue//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 740-756

[137] Godard C, Mac Aodha O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 270-279

[138] Poggi M, Tosi F, Mattoccia S. Learning monocular depth estimation with unsupervised trinocular assumptions//Proceedings of the 2018 International Conference on 3D Vision (3DV). Verona, Italy, 2018: 324-333

[139] Zhou T, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1851-1858

[140] Godard C, Mac Aodha O, Firman M, et al. Digging into self-supervised monocular depth estimation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach, USA, 2019: 3828-3838

[141] Mahjourian R, Wicke M, Angelova A. Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 5667-5675

[142] Vijayanarasimhan S, Ricco S, Schmid C, et al. SfM-Net: Learning of structure and motion from video. arXiv preprint arXiv:1704.07804, 2017

[143] Bian J, Li Z, Wang N, et al. Unsupervised scale-consistent depth and ego-motion learning from monocular video. Advances in Neural Information Processing Systems, 2019, 32: 35-45

[144] Wang G, Wang H, Liu Y, et al. Unsupervised learning of monocular depth and ego-motion using multiple masks//Proceedings of the 2019 International Conference on Robotics and Automation (ICRA). Montreal, Canada, 2019: 4724-4730

[145] Xu Lu, Zhao Hai-Tao, Sun Shao-Yuan. Monocular infrared image depth estimation based on deep convolutional neural networks. Acta Optica Sinica, 2016, 36(7): 188-197 (in Chinese)
(许路, 赵海涛, 孙韶媛. 基于深层卷积神经网络的单目红外图像深度估计. 光学学报, 2016, 36(7): 188-197)

[146] Wang G, Zhang C, Wang H, et al. Unsupervised learning of depth, optical flow and pose with occlusion from 3D geometry. IEEE Transactions on Intelligent Transportation Systems, 2020, 23(1): 308-320

[147] Wang Y, Wang P, Yang Z, et al. UnOS: Unified unsupervised optical-flow and stereo-depth estimation by watching videos//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 8071-8081

[148] Zhao C, Yen G G, Sun Q, et al. Masked GAN for unsupervised depth and pose prediction with scale consistency. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(12): 5392-5403

[149] Wang C, Buenaposada J M, Zhu R, et al. Learning depth from monocular videos using direct methods//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 2022-2030

[150] Steinbrücker F, Sturm J, Cremers D. Real-time visual odometry from dense RGB-D images//Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). Barcelona, Spain, 2011: 719-722

[151] Dai Ren-Yue, Fang Zhi-Jun, Gao Yong-Bin, et al. Unsupervised monocular depth estimation by fusing dilated convolutional network and SLAM. Laser & Optoelectronics Progress, 2020, 57(6): 106-114 (in Chinese)
(戴仁月, 方志军, 高永彬等. 融合扩张卷积网络与 SLAM 的无监督单目深度估计. 激光与光电子学进展, 2020, 57(6): 106-114)

[152] Mur-Artal R, Montiel J M M, Tardos J D. ORB-SLAM: A versatile and accurate monocular SLAM system. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163

[153] Shi G, Xu X, Dai Y. SIFT feature point matching based on improved RANSAC algorithm//Proceedings of the 2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, China, 2013: 474-477

[154] Zhan H, Garg R, Weerasekera C S, et al. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 340-349

[155] Almalioglu Y, Saputra M R U, De Gusmao P P, et al. GANVO: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks//Proceedings of the 2019 International Conference on Robotics and Automation (ICRA). Montreal, Canada, 2019: 5474-5480

[156] Forster C, Pizzoli M, Scaramuzza D. SVO: Fast semi-direct monocular visual odometry//Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA). Banff, Canada, 2014: 15-22

[157] Engel J, Koltun V, Cremers D. Direct sparse odometry. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(3): 611-625

[158] Yin Z, Shi J. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1983-1992

[159] Zou Y, Luo Z, Huang J B. DF-Net: Unsupervised joint learning of depth and flow using cross-task consistency//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 36-53

[160] Hur J, Roth S. Self-supervised monocular scene flow estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 7396-7405

[161] Guizilini V, Ambrus R, Pillai S, et al. 3D packing for self-supervised monocular depth estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 2485-2494

[162] Liu Guo-Dong. Research on Depth Estimation, Instance Segmentation and Reconstruction Based on Visual Semantics [M. S. dissertation]. University of Chinese Academy of Sciences, Shenzhen, 2020 (in Chinese)
(刘国栋. 基于视觉语义的深度估计、实例分割与重建[硕士学位论文]. 中国科学院大学, 深圳, 2020)

[163] Wang Xin-Sheng, Zhang Gui-Ling. Monocular depth estimation based on convolutional neural network. Computer Engineering and Applications, 2020, 56(13): 143-149 (in Chinese)
(王欣盛, 张桂玲. 基于卷积神经网络的单目深度估计. 计算机工程与应用, 2020, 56(13): 143-149)

[164] Klingner M, Termöhlen J A, Mikolajczyk J, et al. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance//Proceedings of the European Conference on Computer Vision. Scottish Event Campus, 2020: 582-600

[165] Chen Y, Schmid C, Sminchisescu C. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 7063-7072

[166] Casser V, Pirk S, Mahjourian R, et al. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos//Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 8001-8008

[167] Liu J, Li Q, Cao R, et al. MiniNet: An extremely lightweight convolutional neural network for real-time unsupervised monocular depth estimation. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 166: 255-267

[168] Kumar A C S, Bhandarkar S M, Prasad M. Monocular depth prediction using generative adversarial networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Virtual, 2018: 300-308

[169] Pilzer A, Xu D, Puscas M, et al. Unsupervised adversarial depth estimation using cycled generative networks//Proceedings of the 2018 International Conference on 3D Vision (3DV). Verona, Italy, 2018: 587-595

[170] Aleotti F, Tosi F, Poggi M, et al. Generative adversarial networks for unsupervised monocular depth prediction//Proceedings of the European Conference on Computer Vision (ECCV) Workshops. Munich, Germany, 2018

[171] Bhattacharyya S. Efficient Attention Guided GAN Architecture for Unsupervised Depth Estimation[Ph. D. dissertation]. The University of North Carolina at Charlotte, 2019

[172] Vankadari M, Garg S, Majumder A, et al. Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation//Proceedings of the European Conference on Computer Vision. Scottish Event Campus, 2020: 443-459

[173] Cheng B, Saggu I S, Shah R, et al. S^3Net: Semantic-aware self-supervised depth estimation with monocular videos and synthetic data//Proceedings of the European Conference on Computer Vision. Scottish Event Campus, 2020: 52-69

[174] Zhao Shuan-Feng, Huang Tao, Xu Qian, et al. Unsupervised monocular depth estimation for autonomous flight of drones. Laser & Optoelectronics Progress, 2020, 57(2): 137-146(in Chinese)
(赵栓峰，黄涛，许倩等. 面向无人机自主飞行的无监督单目视觉深度估计. 激光与光电子学进展，2020，57(2)：137-146)

[175] Xiang Qing-Xiang. Unsupervised Depth Prediction Based on Convolutional Neural Network and Binocular Parallax [M. S. dissertation]. Beijing University of Technology, Beijing, 2018(in Chinese)
(杨青相. 基于卷积神经网络和双目视差的无监督深度预测[硕士学位论文]. 北京工业大学，北京，2018)

[176] Watson J, Mac aodha O, Prisacariu V, et al. The temporal opportunist: Self-supervised multi-frame monocular depth//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 1164-1174

[177] Ding Meng, Jiang Xin-Yan. Scene depth estimation based on monocular vision in advanced driving assistance system. Acta Optica Sinica, 2020, 40(17): 131-139(in Chinese)
(丁萌，姜欣言. 先进驾驶辅助系统中基于单目视觉的场景深度估计方法. 光学学报，2020，40(17)：131-139)

[178] Ma Li, Cao Yi-Ming, Niu Bin. Unsupervised depth estimation from a single image using residual dense network. Journal of Chinese Computer Systems, 2019, 40(11): 2439-2444(in Chinese)
(马利，曹一铭，牛斌. 应用残差稠密网络的无监督单幅图像深度估计. 小型微型计算机系统，2019，40(11)：2439-2444)

[179] Cen Shi-Jie, He Yuan-Lie, Chen Xiao-Cong. A Monocular depth estimation combined with attention and unsupervised deep learning. Journal of Guangdong University of Technology, 2020, 37(4): 35-41(in Chinese)
(岑仕杰，何元烈，陈小聪. 结合注意力与无监督深度学习的单目深度估计. 广东工业大学学报，2020，37(4)：35-41)

[180] Johnston A, Carneiro G. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 4756-4765

[181] Guizilini V, Hou R, Li J, et al. Semantically-guided representation learning for self-supervised monocular depth. arXiv preprint arXiv:2002.12319, 2020

[182] Yang T J, Howard A, Chen B, et al. NetAdapt: Platform-aware neural network adaptation for mobile applications//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 285-300

[183] Kuznietsov Y, Stuckler J, Leibe B. Semi-supervised deep learning for monocular depth map prediction//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 6647-6655

[184] Amiri A J, Loo S Y, Zhang H. Semi-supervised monocular depth estimation with left-right consistency using deep neural network//Proceedings of the 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO). Dali, China, 2019: 602-607

[185] Luo Y, Ren J, Lin M, et al. Single view stereo matching//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2018: 155-163

[186] Xie J, Girshick R, Farhadi A. Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 842-857

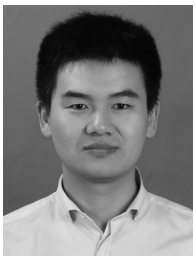
[187] Guizilini V, Li J, Ambrus R, et al. Robust semi-supervised monocular depth estimation with reprojected distances//Proceedings of the Conference on Robot Learning. Osaka, Japan, 2020: 503-512

[188] Ji R, Li K, Wang Y, et al. Semi-supervised adversarial monocular depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(10): 2410-2422

[189] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1125-1134

[190] Ramirez P Z, Poggi M, Tosi F, et al. Geometry meets semantics for semi-supervised monocular depth estimation//Proceedings of the Asian Conference on Computer Vision. Perth, Australia, 2018: 298-313

[191] Wimbauer F, Yang N, Von stumberg L, et al. MonoRec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 6112-6122



JIANG Jun-Jun, Ph. D., professor. His research interests include computer vision, image processing, remote sensing image processing and analysis.

LI Zhen-Yu, M. S. candidate. His research interests include computer vision and image processing.

LIU Xian-Ming, Ph.D., professor. His research interests include artificial intelligence, image processing, and computational imaging.

Background

Depth estimation plays a crucial role in scene perception and understanding, which aims to predict distances between camera and real-world pixels from single or multiple images. It is a popular research field in computer vision, applying as an important step in many practical tasks such as 3D reconstruction. In recent years, depth estimation methods have drawn increasing attention and intensive research as a low-level vision task. Traditional methods use lidar to obtain high-precision depth information but cannot be widely used in practice due to the high cost of obtaining dense and accurate depth maps. In contrast, image-based depth estimation methods directly estimate depth based on input RGB images without expensive equipment, yielding more favor in applications.

In this paper, we aim to do a comprehensive survey about deep-learning based monocular depth estimation methods. There are few overviews having been conducted. However, most of them are out of date or vaguely in classification. In this paper, we elaborately summarize all kinds of methods in detail and classify the methods by the training manner, which is more clear and easy for readers to understand. Moreover, we also analyse the progress of monocular depth estimation and propose some of the most important problems and challenges in this field, hoping to enlight readers for further research. We hope this paper can help readers dig into this field easier, and further push forward the development of this area.