深度学习在视觉定位与三维结构恢复中的研究进展

鲍振强，李艾华，崔智高，袁梦
火箭军工程大学，陕西 西安 710025

摘要 介绍了利用深度学习从图像或视频中恢复三维结构、进行深度估计和实现视觉传感器实时定位方面的研究与应用；对深度学习的研究概况进行了介绍；深入分析和比较了有无监督情况下具有代表性的深度学习算法和系统；对近年来深度学习方面的研究热点进行了讨论，并进行了总结和展望。

关键词 视觉光学；三维感知；视觉定位；深度估计；深度学习；研究进展
中图分类号 TP391.41 **文献标识码** A **doi**：10.3788/LOP55.050007

Research Progress of Deep Learning in Visual Localization and Three-Dimensional Structure Recovery

Bao Zhenqiang, Li Aihua, Cui Zhigao, Yuan Meng
Rocket Force University of Engineering, Xi'an, Shaanxi 710025, *China*

Abstract Research and application of deep learning in recovery of three-dimensional structure from image or video, depth estimation, and real-time localization of visual sensor are introduced. Research progress of deep learning is overviewed. According to whether there is supervision, some representative deep learning algorithms and systems are introduced individually with deep analysis and comparison. Finally, the research spots on deep learning in recent years are discussed, conclusions are presented, and some research tendencies are discussed.

Key words visual optics; three-dimensional perception; vision localization; depth estimation; deep learning; research progress
OCIS codes 330.7310; 140.6910; 140.1135; 200.4260

1 引 言

目前传统的信息表达方式(如文字、图片和视频等)早已不能满足人类的需要，对于很多应用(增强现实、自主导航机器人和无人驾驶等)，如何更好地感知周围真实世界、实现设备在环境中的实时精准定位显得尤为重要。比如，汽车要实现真正的无人驾驶，机器人要完成特定任务，增强现实技术使人足不出户就可以在真实世界中放置虚拟的物品、隔空试穿衣服等，这些都必须使设备能够感知和识别周围的物体，并且要实现准确定位。

通过视觉传感器就可以从二维图像或视频中认知真实世界、恢复三维(3D)空间结构和实现自身定位等。而在从二维图像中获取深度及相机姿态信息方面，常用的算法都过于依赖几何计算，前期的研究[1-3]大多利用几何线索的显式推理来优化 3D 结构，得到其中的深度信息。常用的深度估计与 3D 结构恢复的几何方法有立体视觉、SFM(Structure from Motion)，以及定位与地图构建(SLAM)技术。随着深度学习的发展，利用网络对海量图像进行学习可提取层次化特征，因此，深度学习可成功应用于深度估计、相机位姿估计等多个方面。

收稿日期：2017-10-09；**收到修改稿日期**：2017-11-16
基金项目：国家自然科学基金(61501470)、陕西省重点研发计划(2017GY-075)
作者简介：鲍振强(1991—)，男，硕士研究生，主要从事计算机视觉、视觉定位与导航方面的研究。E-mail：bzhenqiang@163.com
导师简介：李艾华(1966—)，男，博士，教授，主要从事模式识别、机器人技术和计算机视觉方面的研究。E-mail：L863@163.com

本文主要关注深度学习用于3D结构信息的感知、深度估计和视觉里程计方面的研究历史及进展，从有监督学习和无监督学习两个方面着手介绍了其中典型的算法系统，并针对各种算法的优缺点进行了分析与讨论；此外，阐述了深度学习的最新研究热点和发展趋势，并在此基础上进行了总结和展望。

2 深度学习在深度估计和视觉里程计中的应用概况

传统的基于几何的算法在缺乏纹理、几何条件复杂、结构单调和闭塞的情况下都难以发挥很好的作用。V-SLAM(Visual Simultaneous Localization and Mapping)技术是传统的实现定位与3D重构的方法，需要跟踪大量的关键帧及帧间的变换关系，非常繁琐，计算量大且误差相对较大。为此，将深度学习和几何结构相结合，以增强对现实世界场景的理解，提取其中的语义信息，提高现有算法的性能。

2.1 深度估计

Flack等[4]提出将机器学习的方法应用于深度图提取领域，该方法使用5个特征值(x，y，r，g，b)表示每幅图像的像素点，其中x和y表示每幅图像的像素点，r、g、b表示该点在颜色空间上的分量，像素点作为输入，输出则为该像素点的深度值，用灰度值表示；但是，受限于当时的科技水平，该方法没有充分考虑深度线索及帧间的运动信息，只考虑局部信息而忽略了全局结构。然而，该方法为之后用深度学习来估计深度、3D结构恢复等提供了一种全新的思想。

在深度估计方面，传统的做法是直接估计底层的3D几何，即在输入视图之间建立像素级别的对应关系，从输入视图直接合成新的视图，如文献[5-8]中所述。基于深度学习的方法是要对大量二维图像或视频进行训练学习，建立对应的模型系统，从而对新的输入进行预测。近年来，深度学习逐渐用于图像深度估计，主要分为多目[9-10]和单目[11-12]深度估计。基于多目图像的深度估计是根据同一场景在不同图像中的视差来进行估计，并对有监督的深度学习网络进行训练，得到图像间立体匹配的代价函数。Mayer等[13]提出了全卷积深度网络DispNet，通过最小化损失函数来构建回归模型，直接对图像间的一致性进行计算。而在单目深度估计方面，Saxena等[14]于2009年提出了监督型学习方法Make3D，该方法可以对单幅图像进行深度估计，其核心是马尔可夫随机场(MRF)，利用聚类算法将图像中属性(着色、纹理等)相似的部分分割出来，将图像划分为很多极小的区域(超像素块)。Liu等[15]基于超像素分割方法和卷积神经网络(CNN)提取各个超像素块上的特征，进而估计出超像素块的深度；该方法假设各个超像素块的深度都相同，减小了计算量但精度却不高，同时该方法在运动物体的深度提取上存在缺陷，不能准确提取场景中运动物体的深度信息。Eigen等[16]提出了一种多尺度的CNN，针对该网络的特征提取机制一般有2种：1)从图像全局角度出发，提取图像的整体特征；2)在全局特征的基础上精细提取图像各区域的局部特征。随后衍生出了很多改进的CNN模型，比如在CNN模型中加入语义信息[17]，将条件随机场(CRF)与CNN相结合[18]，利用更加稳健的损失函数进行训练[19]，从而进一步提高深度估计的精度。然而，上述网络模型都是基于真实深度进行训练的，即属于监督型学习方法。基于监督型学习方法可以得到真实设定和3D表征的定性或定量推断。从真实设定中获得理想效果的方法已经随着目前CNN实体的发展而进步，然而这是以增加直接3D监督为代价的，所以这种范式相当有限，同时获得这种大规模监督数据的成本也十分高昂，基于监督型学习方法进行单幅图像深度估计的明显缺点是需要昂贵的硬件和精确的标记[20-26]。

随着技术的发展，出现了很多可以从图像中学习真实的3D世界的无监督深度学习方法，如独立成分分析[27]、自编码器[28]、稀疏编码技术[29]和受限玻尔兹曼机[30]等。基于这些方法，无监督深度学习估计方法[31-38]应运而生。基于多目深度估计模型DeepStereo[39]，利用无监督深度学习方法合成新的视图，但是网络结构中包含多个复杂的处理阶段，训练比较繁琐，参数难以调节。Xie等[40]随后提出的深度估计模型Deep3D可以直接从视频数据中估计出深度信息，同样此网络的缺点也很明显，极大地增加了计算的内存，提高了对硬件设备的要求。2017年，Godard等[41]基于无监督深度学习方法实现单目视图的深度估计，并对损失函数(由图像误差、平滑误差和左右一致性误差三部分组成)进行了优化，从而增强了观测的平滑性，提高了深度估计的精度。

2.2 视觉里程计

视觉里程计也称为帧间估计，相比于传统的基于稠密或稀疏特征的帧间估计方法，基于深度学习的方法避免了特征提取、特征匹配和复杂的几何运算，使得基于深度学习的方法更加直观简洁。

Konda 等[42]提出的深度学习模型可以预测相机速度和方向的改变,其深度学习过程主要分为两步:1)提取图像序列的深度信息和运动信息;2)利用 CNN 进行学习,执行帧间估计。实验结果表明,该学习算法的执行速度较快,在 3.2 GHz CPU,24 GB RAM 和 GTX 680 GPU 配置的机器上,平均执行速度达到 0.026 s/frame,但是其估计精度还难以与传统视觉里程计方法相比。Handa 等[43]构建了包含全局变换、像素变换和 M 估计器在内的神经网络几何视觉(gvnn)软件库,网络系统包括 Siamse 网络层、位姿变换估计层、3D 网络生成层、投影层和双线性插值层;其中 Siamse 网络层的输入为两个连续的图像帧,输出为相机 6 自由度的位姿变换矩阵,基于此变换矩阵将上一帧图像投影到当前的位姿,并经过双线性插值生成预测图像,结合深度相机的深度信息来构造损失函数,并进行训练学习;该方法避免了传统神经网络在学习过程中单方面的像素丢失和各种运动模糊、强度变化和图像噪声对匹配的影响,大大减小了误差[44]。此外,Wang 等[45-46]将递归神经网络(RNN)与 CNN 结合到一起,形成递归卷积神经网络(RCNNs),并将其应用于视觉里程计问题中,取得了不错的效果,该网络可以同时进行深度和位姿估计。

3 视觉定位与 3D 结构恢复典型深度学习算法

3.1 监督型学习方法的代表性算法

3.1.1 Make3D 方法

Make3D 方法是 Saxena 等[14]于 2009 年提出的由监督学习方法从单幅图像估计出详细 3D 结构的算法,该监督型参数化学习方法的核心是 MRF,以此来描述图像中各点的深度及彼此间的空间关系。从实验效果上来看,Make3D 对非结构化、相机参数未知的大范围复杂场景深度估计有相对理想的效果,但不能准确提取场景中运动物体的深度信息。

Make3D 方法以超像素块为基本单元,如图 1 所示,图中 α 为平面参数,d_i 为超像素块上点 i 到相机中心的距离。利用聚类算法将图像中属性相似(着色和纹理等)的部分分割出来,将图像划分为很多极小的区域(超像素块)。兼顾各个超像素块的深度信息及超像素块之间的联系,各个超像素块所在的平面构成了 3D 模型的基本单元,综合得到一个反映实景的图像模型。建立模型时需考虑图像的 4 种属性:1)图像特征和深度,超像素块内包含深度信息的特征;2)关联结构,除去遮挡的情况,相邻的超像素块很可能互相连接;3)共面结构,两个相邻超像素块无连接且拥有相似特征,即共面;4)共线结构,二维图像中的长直线在 3D 模型中也可能是直线。图 2 为建立模型时考虑的部分属性,其中 S_i,S_j,S_i',S_j',S_i'',S_j''均为超像素块。

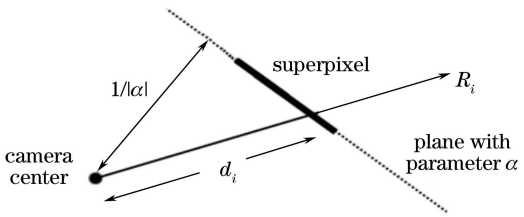

图 1 超像素块的特征向量

Fig. 1 Feature vector of superpixel block

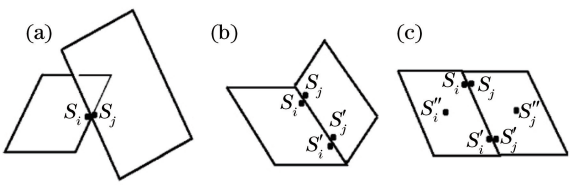

图 2 图像属性。(a)关联结构;(b)共线结构;(c)共面结构

Fig. 2 Image properties. (a) Connected structure; (b) co-linearity structure; (c) co-planarity structure

根据以上 4 种属性,通过原图像及对应的深度图得到 MRF 模型,实现 3D 重构,最终需要优化的能量函数为

$$P(\alpha \mid X,\nu,y,R;\theta)=\frac{1}{Z}\prod_i f_1\times(\alpha_i \mid \boldsymbol{X}_i,\nu_i,R_i;\theta)\times\prod_{i,j} f_2(\alpha_i,\alpha_j \mid y_{ij},R_i,R_j),\tag{1}$$

式中 f_1 函数表征超像素块的特征向量与空间平面参数之间的统计相关性;f_2 函数表征两个不同的超像素块 i,j 之间的关系(互连性、共面性、共线性和遮挡关系等);α_i 为超像素块 i 的平面参数,超像素块 i 中共有 S_i 个像素;$\boldsymbol{X}_i=\{x_{i,s_i}\in\mathbf{R}^{524}:s_i=1,2,\cdots,S_i\}$ 为超像素块 i 的特征向量,其中 x_{i,s_i} 表示超像素块 i 中像素 S_i 的特征;R_i 表示相机中心射向超像素块 i 的光束集合,$R_i=\{R_{i,s_i}:s_i=1,\cdots,S_i\}$;$\nu_i$ 表示超像素块 i 估计深度的置信度;θ 为模型参数;Z 为归一化常数;y_{ij} 为超像素块 i、j 对应的空间平面之间是否存在遮挡关系的标记。

3.1.2 几何-环境网络方法

2017 年 Kenall 等[25]在 CVPR(Computer Vision and Pattern Recognition)大会上提出了几

何-环境网络(GC-Net)方法,其网络结构模型是一种端到端的回归模型,如图 3 所示;GC-Net 方法的创新之处在于:用微分方式建立几何代价,并将其用于回归模型,无额外的预处理或正则化等操作,降低了工程设计的复杂性。该方法的优点包括:1) 不降低特征维度,使用深度特征组成成本向量,充分利用 3D 卷积提取上下文的纹理信息,并用 3D 卷积来传播信息,无需利用学习概率分布、成本函数或分类结果;2) 以一个更加宽泛的视野和更加有效的图像上下文信息回归拟合立体匹配视差,这远比只利用局部几何和外形信息效果要好。从图 4 的测试结果来看,该方法估计的深度已经非常接近真实值。

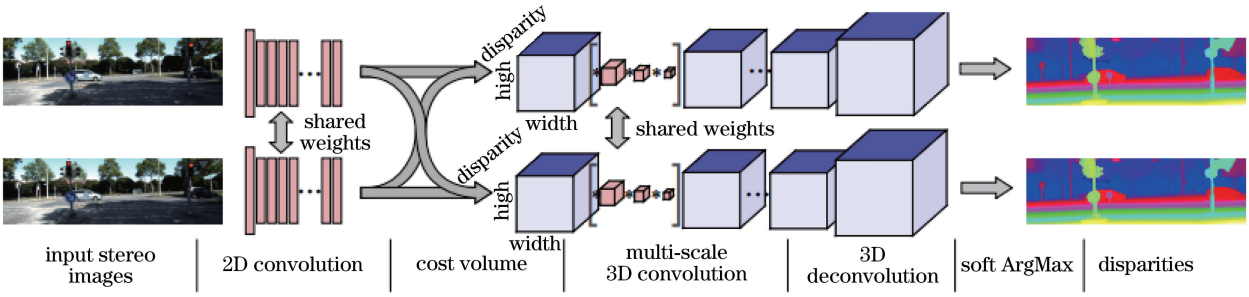

图 3 GC-Net 模型结构示意图

Fig. 3 Structural diagram of GC-Net model

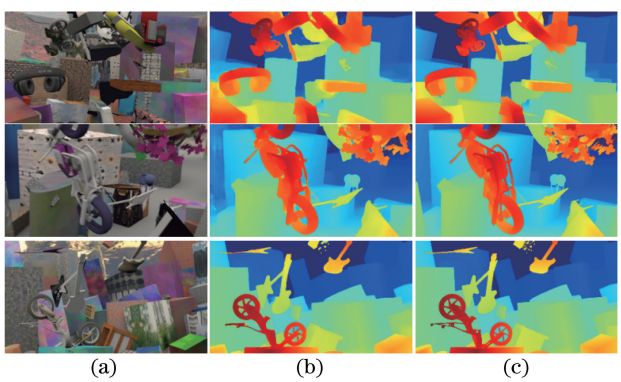

图 4 测试结果。(a)输入双目图的左边图像;(b)深度估计结果;(c)真实深度

Fig. 4 Test results. (a) Left image of input binocular image; (b) depth estimation result; (c) true depth

3.1.3 基于多视角监督学习的单视角重建方法

几何学是学习系统与多视角训练数据间的桥梁[47-49],通过学习系统和几何之间的相互作用,使得学习系统预测和多视角观察得到的几何一致,从而达到 3D 恢复的目的。经典的学习算法架构如图 5 所示。

如图 5 所示,在监督环境下,训练数据包含不同视角的多种观测结果,使用该训练数据来学习预测器 P;利用预测器 P 并根据单幅 2D 图像推断出 3D 结果;在预测器 P 和几何输出之间设置一个简单的策略网络检验器 V。一般地,训练步骤如下:1) 选取一个随机训练图像 I,此图像与从视角 C 观察到的结果 O 相关;2) 预测 3D 形状 $S=P(I)$,使用 V 来检测 (S,O,C) 的一致性;3) 更新 P,使用梯度下降法,使 S 与 (O,C) 更一致;4) 重复此过程直至其收敛。

利用多视角监督学习进行单视角预测的方法全部遵守上面的模板,核心在于预测器 P 与 (O,C) 无

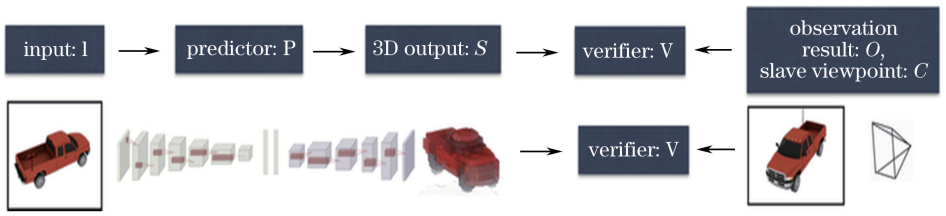

图 5 经典的学习算法架构

Fig. 5 Classical learning algorithm architecture

关，故会遍历所有可能与观察结果相一致的 S，从而得到很好的预测结果，但其劣势在于多视角观察结果 O 的种类过于单一。为此，文献[26]中利用经典射线一致性公式引入了一个一般的检验器，以衡量3D形状与不同种类2D观察结果间的一致性。该通用性公式通过利用不同种类的多视角观察结果来对3D预测结果（如前景模板、深度、彩色图像和语义等）进行学习。观察结果 O 中的每一个像素都对应着一条有相关信息的射线，通过计算3D形状 S 与射线 r 之间的一致性，来替代计算观察结果 O 与3D形状 S 之间的几何一致性，此方法的优势在于可以通过简单定义相应的代价函数进行多种观察（如颜色图片、前景等）。

3.2 无监督型学习方法的代表性算法

无监督学习是人工智能领域的研究热点，被认为是人工智能能够在真实世界中真正有效的进行自我学习的方法，在深度估计、位姿估计、3D重建、机器人自主导航和自动驾驶中都具有重要的应用价值；然而，无监督学习这种可以完全自我学习的特性，使其研究进展相对缓慢。

几何信息具有良好的可观测性和连续表达性，可以用于构建无监督学习模型[31-33,36-37]。文献[36]中使用了光流法，将得到的光流图像送入卷积网络学习，用无监督方法从几何信息中学习运动信息。文献[37]中采用无监督方法从几何信息中学习深度信息。

3.2.1 深度估计网络

文献[37]提出利用无监督型CNN进行图像的深度估计，该方法将两个距离固定的摄像头拍摄的图像对作为训练数据，通过双目摄像头测距原理，计算真实的物体与摄像机之间的距离，不需要人工标注数据，图6为该算法的总体流程图。

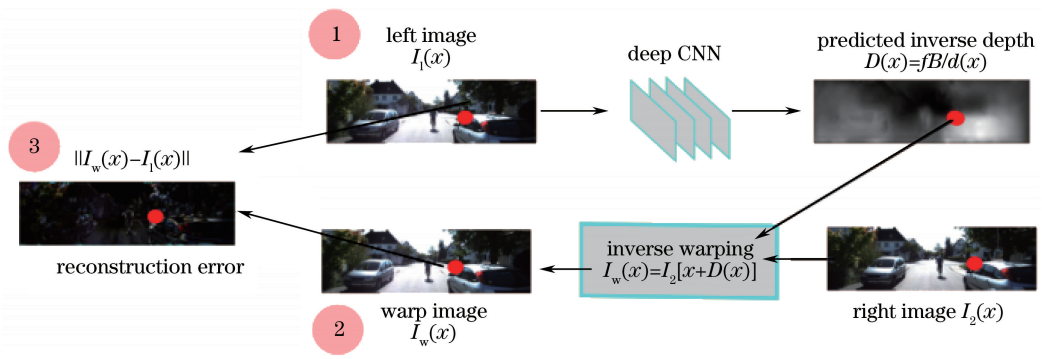

图6 算法的总体流程图

Fig. 6 Overview flow chart of algorithm

上述算法的具体流程为：将左边图像 $I_1(x)$ 作为输入，由深度网络输出深度图，然后将左边深度图与右边图像 $I_2(x)$ 结合，重构出图像 $I_w(x)$，最后以 $I_w(x)$ 与输入图像 $I_1(x)$ 的图像误差训练模型，直至收敛。其中 $D(x)$ 为双目相机的视差距离，f 为相机焦距，B 为双目相机的基线长度，$d(x)$ 为像素点深度值。

该算法中，目标函数 E 由两部分组成，解决了视差不连续的问题（梯度），E 可表示为

$$E=\sum_{i=1}^{N}\left(E^{i}_{\text{recons}}+\gamma E^{i}_{\text{smooth}}\right), \tag{2}$$

式中 E^{i}_{recons} 为重构图像 $I_w(x)$ 与输入图像 $I_1(x)$ 的图像误差，即回归像素点的深度值；$\gamma E^{i}_{\text{smooth}}$ 为由视差不连续所产生的误差，其中 γ 为视差不连续误差的强度参数。利用相邻区域深度的一致性可提高深度估计的精度。

文献[50]中基于回归森林算法和深度图创建场景坐标标签，推断出实际坐标系中的每个像素对应的3D坐标，进而预测相机的姿态。位姿估计方法使用一种计量估计器结果与给定位姿的像素数量一致的能量函数，然后使用RANSAC算法得到相机的位姿。该方法的局限在于输入图像只能为RGB-D相机拍摄的深度图，并且只能用于室内场景。

3.2.2 PoseNet

PoseNet是剑桥大学Kendall等[35]于2015年提出的一种视觉定位算法，该算法利用深度学习来实现单目摄像机自身定位及环境场景识别。该算法通过训练从大量图片中自动生成标签，网络深度为23层，这表明深度卷积网络可用于解决复杂的图像回归问题，从而使得由大型分类数据进行转移学习成为可能。PoseNet算法的精度高，实时性强，可以用于室内和室外，在光照变化、运动模糊等情况下，

该算法较稳定。其不足之处在于：该算法将相机位置和姿态作为两个单独的回归目标进行学习，即其目标函数$l_{\text{oss}}(I)$由两部分组成：

$$l_{\text{oss}}(I)=\| \hat{x}-x \|_2+\beta \| \hat{q}-\frac{q}{\| q \|} \|_2, \tag{3}$$

式中x、$\hat{x}$分别为相机的真实位置和预测位置；q、$\hat{q}$分别为相机的真实姿态和预测姿态；β为比例因子，其作用是保持位置和姿态两者的误差大致相等。

针对 PoseNet 算法的不足，文献[38]对原 PoseNet 目标函数进行了改进，自动学习位置与姿态之间的权重。图像I对应的相机总误差$L_\sigma(I)$可表示为

$$L_\sigma(I)=L_x(I)\hat{\sigma}_x^{-2}+\lg\hat{\sigma}_x^2+L_q(I)\hat{\sigma}_q^{-2}+\lg\hat{\sigma}_q^2, \tag{4}$$

式中$L_x(I)$、$L_q(I)$分别为图像I对应的相机位置误差和相机姿态误差，x、q分别为相机位置和姿态，$\hat{\sigma}_x^2$、$\hat{\sigma}_q^2$分别为每帧图像I对应的相机位置与姿态的方差。此外，该算法还使用几何重投影误差对相机位置和姿态同时进行学习，利用真实世界的几何信息极大地提高了算法性能。优化的目标函数可表示为

$$L_g(I)=\frac{1}{|\psi'|}\sum_{g\in\psi'}\| \pi(x,q,g)-\pi(\hat{x},\hat{q},g) \|_\gamma, \tag{5}$$

式中$\pi(\cdot)$为空间点g与二维图像点(u, v)的映射函数，x、$\hat{x}$分别为相机位置的真实值和估计值，q、$\hat{q}$分别为姿态的真实值与估计值，ψ'表示空间点的集合。映射关系可表示为

$$\pi(x,q,g)\rightarrow\begin{pmatrix}u\\v\end{pmatrix}, \tag{6}$$

$$\begin{pmatrix}u'\\v'\\w'\end{pmatrix}=\boldsymbol{K}(\boldsymbol{R}g+x),\begin{pmatrix}u\\v\end{pmatrix}=\begin{pmatrix}u'/w'\\v'/w'\end{pmatrix}, \tag{7}$$

式中$\boldsymbol{K}$为相机的内参矩阵；$\boldsymbol{R}$为q映射到特殊正交群的旋转矩阵，即$q_{4\times1}\rightarrow\boldsymbol{R}_{3\times3}$；$u'$、$v'$、$w'$分别为空间点$g$的三维坐标值。

3.2.3 深度估计 CNN 结合位姿估计 CNN

Zhou 等[51]采用了无监督方法对视频数据进行训练，并对单张图片的深度以及连续帧之间的车辆运动进行估计。如图 7 所示，该方法将用于单一视角深度估计的深度估计 CNN 与用于帧间估计的位姿估计 CNN 相结合，并对两个网络进行联合训练。其中I_t，I_{t-1}，I_{t+1}分别为在时刻t，$t-1$，$t+1$拍摄的图像，p_t，p_{t-1}，p_{t+1}分别为同一空间点在三张图像上的投影点，$\hat{\boldsymbol{T}}_{t\rightarrow t-1}$，$\hat{\boldsymbol{T}}_{t\rightarrow t+1}$分别表示时刻$t$到时刻$t-1$和$t+1$的相机位姿转换矩阵，$\hat{D}_t(p)$为预测的相机在时刻$t$的深度点。

基于合成视图方法的网络结构如图 8 所示。对于深度估计网络，输入为单张图像，网络结构在

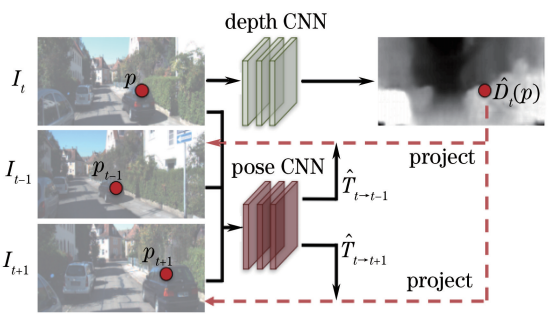

图 7 基于合成视图方法的总体流程图

Fig. 7 Overall flow chart based on composite view method

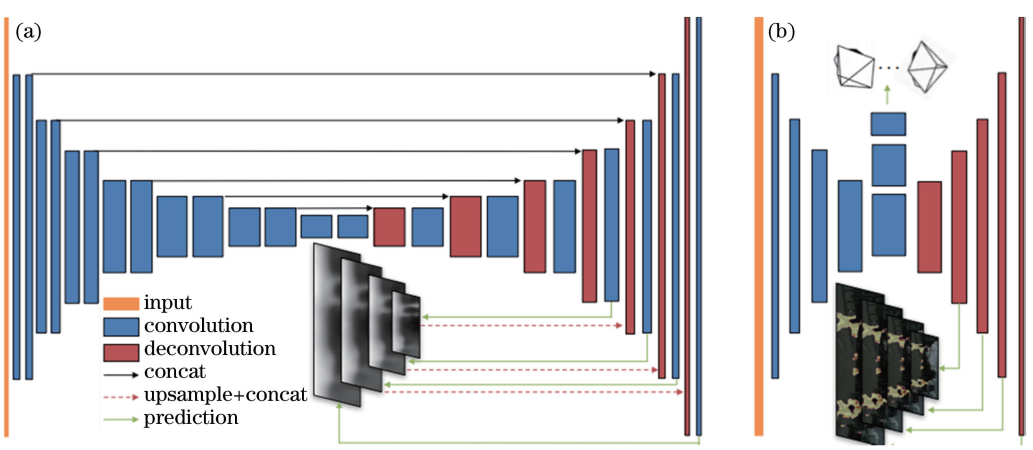

图 8 网络结构图。(a)深度估计网络；(b) 位姿估计/解释性网络

Fig. 8 Network architectures. (a) Depth estimation network; (b) pose estimation/explainability network

DispNet 网络的基础上加了多尺度输出；对于位姿估计/解释性网络，输入为连续帧的切片，两个网络共享前 5 层卷积，预测出 6 自由度的帧间转换位姿之后，再进行反卷积，输出不同尺度下的可解释性模版。

综合来讲，文献[51]的亮点在于：提出了一种无监督的深度学习方法，使网络可以用更多的场景进行训练；使用了光流的处理方法，保证得到正确的深度估计和帧间估计以使目标函数最优；利用了卷积-反卷积的网络结构，引入平滑损失的机制增强了深度预测的平滑稳定性，引入多尺度估计机制提高了训练的稳定性；引入可解释性模版，滤除场景中的运动物体，解决图像随视角变化的遮挡问题等。

对文献[51]中所采用的系统在 KITTI 和 Cityscapes 数据集上进行训练，从最终评估性能来看，此无监督训练系统的效果非常好，深度估计模型可以与基于监督的模型效果相媲美，而摄像机的位姿估计也与目前相当完备的 SLAM 系统相当。

4 研究热点讨论

构建深度学习网络时，或多或少都存在一些理想条件下的设定，在应用上有明显的局限性。然而，现实环境是复杂多变的，因此要真正实现无人驾驶等目标，解决以下问题意义重大。

1) 构建深度网络时需要更好地考虑动态物体及遮挡问题。在实际应用场景中，难以避免地存在运动物体，建模时大多都将这些动态特征进行了滤除。但是，当这些运动物体在图像中所占的比重较大时，难以保证滤除效果。此外，遮挡现象也是不可回避的问题，其直接影响重构和定位的效果。在构建网络时，如何很好地引入动态物体检测和遮挡检测，值得深入探讨，这对提高算法的稳定性及定位精度至关重要。

2) 损失函数。现有损失函数的构建方法的效果不是很好，与传统基于点特征的方法尚存在一定的差距。因此，如何利用场景几何信息构建更加合理的损失函数，并将其用于网络训练，以实现更加精确的定位与深度估计，也将是研究的热点。

3) 从连续视频中精确还原 3D 场景。目前的一些深度学习算法虽然已经可以直接从视频中恢复出深度信息，但这些深度信息相当模糊，对于需要精确恢复 3D 场景的应用(如增强现实等)，还远不能满足要求。

4) 深度网络的整合。目前，在视觉定位和深度恢复方面，深度网络均包括深度估计 CNN 和位姿估计 CNN 两个单独的网络，即使二者之间存在相对融合的部分。如果把这两个训练网络整合为一个网络，系统的实时性、稳定性和定位精度都将得到很大的提升。

5) 深度网络可解释性方面的探索。面对深度学习这个黑匣子，研究其可解释性，通过层层网络最终得出场景深度和位姿等信息，而不是从神经网络角度去理解或是避开这个问题，这是一件十分有意义的事，如果对深度网络进行了真正阐释，深度学习方法得到了理论支撑，深度学习或许会变得更加简单。

5 总结与展望

主要阐述了深度学习在图像、视频 3D 结构恢复及定位中的应用，介绍了其发展历程，总结了最新研究进展。根据有无监督进行分类，分别重点介绍了这两类方法中的一些代表性算法，并从创新性、效果和稳定性等方面进行了分析和比较，立足实际应用，简要探讨如下几点发展趋势。

1) 有监督模式向无监督模式的转变。目前虽然有监督的学习方式效果不错，但前提是要拥有大量标记的训练数据，收集和标记大型数据集的过程非常耗时，而且容易出错。当数据集的规模增加时，这些问题便更加突出。而无监督的学习方式具有巨大的优势，因此逐步从有监督学习模式转变为无监督学习模式是大势所趋。

2) 设备小型化。深度学习模型是一个高维的模型，优化这个模型对硬件设备的要求很高，要得到精确恢复的 3D 模型且做到精确定位，需要利用高性能计算工作站(甚至几十个 GPU)来完成大量数据的训练工作。从人的需求和用户体验角度考虑，设备小且智能非常重要，如拥有高性能计算设备的智能手机、平板电脑等。因此，研究并开发自身携带强大计算功能的视觉设备也是亟待解决的问题。

3) 多传感器融合。仅依靠视觉传感器具有其局限性，将其他传感器数据也整合进深度学习网络中，共同构建网络模型，可以得到更高的精度和稳定性。

参 考 文 献

[1] Roberts L G. Machine perception of three-dimensional solids [M]. Cambridge: Massachusetts Institute of Technology, 1965: 31-39.

[2] Barrow H G, Tenenbaum J M. Interpreting line drawings as three-dimensional surfaces[J]. Artificial Intelligence, 1981, 17: 75-116.

[3] Tian Y B, Bai J, Huang Z. Depth estimation with a panoramic stereo imaging system [J]. Acta Optica Sinica, 2013, 33(6): 0611002.
田延冰, 白剑, 黄治. 基于全景环带立体成像系统的

深度信息估计[J]. 光学学报, 2013, 33(6): 0611002.

[4] Flack J, Fox S. Rapid 2D-to-3D conversion[C]. SPIE, 2002, 4660: 78-86.

[5] Chen S E, Williams L. View interpolation for image synthesis[C]. Conference on Computer Graphics and Interactive Techniques, 1993: 279-288.

[6] Fitzgibbon A, Wexler Y, Zisserman A. Image-based rendering using image-based priors[J]. International Journal of Computer Vision, 2005, 63(2): 141-151.

[7] Seitz S M, Dyer C R. View morphing[C]. Conference on Computer Graphics and Interactive Techniques, 1996: 21-30.

[8] Zitnick C L, Kang S B, Uyttendaele M, *et al*. High-quality video view interpolation using alayered representation[C]. ACM Transactions on Graphics, 2004, 23(3): 600-608.

[9] L′ubor L, Häne C, Pollefeys M. Learning the matching function [J]. Computer Science, 2015: arXiv.

[10] Zbontar J, LeCun Y. Stereo matching by training a convolutional neural network to compare image patches[J]. Journal of Machine Learning Research, 2016, 17(65): 1-32.

[11] Xu L, Zhao H T, Sun S Y. Monocular infrared image depth estimation based on deep convolutional neural networks[J]. Acta Optica Sinica, 2016, 36(7): 0715002.
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[J]. 光学学报, 2016, 36(7): 0715002.

[12] Wu S C, Zhao H T, Sun S Y. Depth estimation from monocular infrared video based on bi-recursive convolutional neural network[J]. Acta Optica Sinica, 2017, 37(12): 1215003.
吴寿川, 赵海涛, 孙韶媛. 基于双向递归卷积神经网络的单目红外视频深度估计[J]. 光学学报, 2017, 37(12): 1215003.

[13] Mayer N, Ilg E, Häusser P, *et al*. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 4040-4048.

[14] Saxena A, Sun M, Ng A Y. Make3D: learning 3D scene structure from a single still image[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(5): 824-840.

[15] Liu F, Shen C, Lin G, *et al*. Learning depth from single monocular images using deep convolutional neural fields [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(10): 2024-2039.

[16] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C]. International Conference on Neural Information Processing Systems, 2014: 2366-2374.

[17] Shi J, Pollefeys M. Pulling things out of perspective [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2014: 89-96.

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

[19] Laina I, Rupprecht C, Belagiannis V, *et al*. Deeper depth prediction with fully convolutional residual networks[C]. Fourth IEEE International Conference on 3D Vision, 2016: 239-248.

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

[21] Fan X, Zheng K, Lin Y, *et al*. Combining local appearance and holistic view: dual-source deep neural networks for human pose estimation [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2015, 8753: 1347-1355.

[22] Ummenhofer B, Zhou H, Uhrig J, *et al*. DeMoN: depth and motion network for learning monocular stereo[C]. 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017: 5622-5631.

[23] Kuznietsov Y, Stückler J, Leibe B. Semi-supervised deep learning for monocular depth map prediction [C]. 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017: 2215-2223.

[24] Liu B, Gould S, Koller D. Single image depth estimation from predicted semantic labels[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2010: 1253-1260.

[25] Kendall A, Martirosyan H, Dasgupta S, *et al*. End-to-end learning of geometry and context for deep stereo regression [C]. 16th IEEE International Conference on Computer Vision, 2017: 66-75.

[26] Tulsiani S, Zhou T, Efros A A, *et al*. Multi-view supervision for single-view reconstruction via

differentiable ray consistency[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 209-217.

[27] Bell A J, Sejnowski T J. The " independent components" of natural scenes are edge filters[J]. Vision Research, 1997, 37(23): 3327-3338.

[28] Bourlard H, Kamp Y. Auto-association by multilayer perceptrons and singular value decomposition[J]. Biological Cybernetics, 1988, 59(4/5): 291-294.

[29] Olshausen B A, Field D J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images[J]. Nature, 1996, 381(6583): 607-609.

[30] Salakhutdinov R, Hinton G. Deep Boltzmann machines [J]. Journal of Machine Learning Research, 2009, 5 (2): 1967-2006.

[31] Gadelha M, Maji S, Wang R. Shape generation using spatially partitioned point clouds [J]. Computer Science, 2016: arXiv:1707.06267.

[32] Rezende D J, Eslami S M A, Mohamed S, *et al*. Unsupervised learning of 3D structure from images [J]. Advances in Neural Information Processing Systems, 2016: 4997-5005.

[33] Yan X, Yang J, Yumer E, *et al*. Perspective transformer nets: learning single-view 3D object reconstruction without 3D supervision[J]. Advances in Neural Information Processing Systems, 2016: 1696-1704.

[34] Jayaraman D, Grauman K. Learning image representations tied to ego-motion[C]. IEEE International Conference on Computer Vision, 2015: 1413-1421.

[35] Kendall A, Grimes M, Cipolla R. PoseNet: a convolutional network for real-time 6-DOF camera relocalization[C]. IEEE International Conference on Computer Vision, 2015: 2938-2946.

[36] Agrawal P, Carreira J, Malik J. Learning to see by moving [C]. IEEE International Conference on Computer Vision, 2015: 37-45.

[37] Garg R, Vijay K B G, Carneiro G, *et al*. Unsupervised CNN forsingle view depth estimation: geometry to the rescue [C]. 14th European Conference on Computer Vision, 2016, 9912: 740-756.

[38] Kendall A, Cipolla R. Geometric loss functions for camera pose regression with deep learning[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6555-6564.

[39] Flynn J, Snavely K, Neulander I, *et al*. Deepstereo: learning to predict new views from real world imagery: US20160335795[P]. 2018-03-13.

[40] Xie J, Girshick R, Farhadi A. Deep3D: Fully automatic 2D-to-3D video conversion with deep vonvolutional neural networks[C]. 14th European Conference on Computer Vision, 2016, 9908: 842-857.

[41] Godard C, Aodha O M, Brostow G J. Unsupervised monocular depth estimation with left-right consistency [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6602-6611.

[42] Konda K, Memisevic R. Learning visual odometry with a convolutional network [C]. International Conference on Computer Vision Theory and Applications, 2015: 486-490.

[43] Handa A, Bloesch M, Pătrăucean V, *et al*. gvnn: neural network library for geometric computer vision [C]. 14th European Conference on Computer Vision, 2016, 9915: 67-82.

[44] Zhao Y, Liu G L, Tian G H, *et al*. A survey of visual SLAM based on deep learning[J]. Robot, 2017, 39(6): 889-896.
赵洋, 刘国良, 田国会, 等. 基于深度学习的视觉SLAM综述[J]. 机器人, 2017, 39(6): 889-896.

[45] Wang S, Clark R, Wen H, *et al*. DeepVO: towards end-to-end visual odometry with deep recurrent convolutional neural networks [C]. IEEE International Conference on Robotics and Automation, 2017: 2043-2050.

[46] Li R, Wang S, Long Z, *et al*. UnDeepVO: monocular visual odometry through unsupervised deep learning[J]. Computer Science, 2017: arXiv: 1709.06841.

[47] Vijayanarasimhan S, Ricco S, Schmid C, *et al*. SfM-Net: learning of structure and motion from video[J]. Computer Science, 2017: arXiv:1704.07804.

[48] Gadelha M, Maji S, Wang R. 3D shape induction from 2D views of multiple objects[J]. Computer Science, 2016: arXiv:1612.05872.

[49] Arora R, Livescu K. Multi-view learning with supervision for transformed bottleneck features[C]. IEEE International Conference on Acoustics, Speech and Signal Processing, 2014: 2499-2503.

[50] Shotton J, Glocker B, Zach C, *et al*. Scene coordinate regression forests for camera relocalization in RGB-D images[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2013: 2930-2937.

[51] Zhou T, Brown M, Snavely N, *et al*. Unsupervised learning of depth and ego-motion from video[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6612-6619.