深度学习在基于单幅图像的物体三维重建中的应用

陈 加 [1, 2] 张玉麒 [1] 宋 鹏 [3] 魏艳涛 [1] 王 煜 [4]

摘 要 基于单幅图像的物体三维重建是计算机视觉领域的一个重要问题, 近几十年来得到了广泛的关注. 随着深度学习的不断发展, 近年来基于单幅图像的物体三维重建取得了显著进展. 本文对深度学习在基于单幅图像的物体三维重建领域的研究进展及具体应用进行了综述. 首先介绍了基于单幅图像的三维重建的研究背景及其传统方法的研究现状, 其次简要介绍了深度学习并详细综述了深度学习在基于单幅图像的物体三维重建中的应用, 随后简要概述了三维物体重建的常用公共数据集, 最后进行了分析与总结, 指出了目前存在的问题及未来的研究方向.

关键词 三维重建, 深度学习, 计算机视觉, 单幅图像

引用格式 陈加, 张玉麒, 宋鹏, 魏艳涛, 王煜. 深度学习在基于单幅图像的物体三维重建中的应用. 自动化学报, 2019, **45**(4): 657−668

DOI 10.16383/j.aas.2018.c180236

Application of Deep Learning to 3D Object Reconstruction From a Single Image

CHEN Jia[1, 2] ZHANG Yu-Qi[1] SONG Peng[3] WEI Yan-Tao[1] WANG Yu[4]

Abstract 3D object reconstruction from a single image is an important topic in computer vision, which has attracted enormous attention during the past decades. With the further study in deep learning, remarkable progress of 3D object reconstruction from a single image has been obtained in recent years. In this paper, we review the applications of deep learning models in the field of 3D object reconstruction from a single image. First, we introduce the research background and the current state-of-the-art of traditional methods. Then, we provide a brief overview of typical deep learning models and we describe the applications of deep learning techniques in 3D object reconstruction from a single image. After that, we list several commonly used data sets for 3D object reconstruction. Finally, we discuss current challenges and further research directions.

Key words 3D reconstruction, deep learning, computer vision, single image

Citation Chen Jia, Zhang Yu-Qi, Song Peng, Wei Yan-Tao, Wang Yu. Application of deep learning to 3D object reconstruction from a single image. *Acta Automatica Sinica*, 2019, **45**(4): 657−668

收稿日期 2018-04-20 录用日期 2018-08-30
Manuscript received April 20, 2018; accepted August 30, 2018
国家自然科学基金 (61605054, 61502195), 湖北省自然科学基金 (2014CFB659), 华中师范大学中央高校基本科研业务费 (CCNU19QD007, CCNU19TD007, CCNU16JYKX039, CCNU15A05023) 资助
Supported by National Natural Science Foundation of China (61605054, 61502195), Hubei Provincial Natural Science Foundation (2014CFB659), and the Fundamental Research Funds for the central Universities of Central China Normal University (CCNU19QD007, CCNU19TD007, CCNU16JYKX039, CCNU15A05023)
本文责任编委 吴毅红
Recommended by Associate Editor WU Yi-Hong

1. 华中师范大学教育信息技术学院 武汉 430079, 中国 2. 英国萨里大学视觉、语音和信号处理中心 萨里 GU27XH, 英国 3. 瑞士联邦理工学院 (洛桑) 计算机图形学与几何实验室 洛桑 CH-1015, 瑞士 4. 香港科技大学机器人研究院 香港 999077, 中国
1. School of Educational Information Technology, Central China Normal University, Wuhan 430079, China 2. Centre for Vision, Speech and Signal Processing, University of Surrey, Surrey GU27XH, UK 3. Computer Graphics and Geometry Laboratory, École Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland 4. Robotics Institute, the Hong Kong University of Science and Technology, Hong Kong 999077, China

计算机视觉研究的主要目标之一是从二维图像复原三维结构[1]. 二维图像是当今时代极易获取的数据形式, 互联网上每时每刻都在产生海量的图像数据, 而三维模型相对难以获取. 目前已经有许多基于多幅图像的物体三维重建方法[2−4] 被提出, 而基于单幅图像的物体三维重建问题因输入形式的特殊性使其更具挑战性. 仅以单幅图像作为输入使得重建丢失了很多几何信息, 这就需要一些假设或先验知识, 亦或从已有的模型基于学习来进行重建. 此外针对重建对象的不同, 当前基于单幅图像的重建问题可分为对物体 (Object) 的重建和对场景 (Scene) 的重建[5]. 本文属于对物体的重建这一子类. 评价基于单幅图像的物体三维重建的重建精度目前没有一个固定的标准[6], 一些传统方法采用 Hausdorff 距离[7] 作为评价重建效果的标准. 随着深度学习的不断发展, IoU (Intersection over union) 值被引入[8], 且被很多论文采用为评价标准, 此外亦有论文采用更注重物体几何外形的 CD (Chamfer distance) 值

等[9].

基于图像的三维重建具有重要的实用价值和应用前景. 随着互联网及电子商务的发展, 很多商家或企业开始利用三维模型来帮助自己进行产品的展示与推广. 三维模型与二维图像相比, 因为多了一维信息, 所以更能将物体的真实感和细节的质感与纹理表现出来. 同时, 在诸如虚拟现实、城市数字化建模、文物数字化保护、医学 CT 器官重建、三维影视动漫制作等领域, 基于图像的三维重建也具有广泛的应用[5, 10−13]. 多目图像三维重建往往需要相机标定等额外操作, 相比之下基于单幅图像的三维重建因输入简单, 更适合需要便捷式三维重建的应用场合, 近年来逐渐成为一个新的学术研究热点问题.

然而基于单幅图像的三维重建常常面临以下几个方面的挑战:

1) 类内差异和类间差异. 不同的重建物体即使是同一个类型, 也会因为材料、外形等存在较大的差异性. 而不同类型的物体, 亦可能存在较大的相似性. 如果只是针对某个特定类别的物体进行三维重建往往会使重建系统缺乏一般性[14−15], 而针对多类别的重建系统则会因较大的类内差异和较小的类间差异使得重建精度不高[16−17], 如何构建既具有一般性又重建精度高的三维重建算法是目前研究的重点.

2) 图像自身属性. 真实世界视图中的物体往往存在遮挡、非刚性变形等现象, 且很多时候不满足理想的朗伯特反射模型[18], 这就使得待重建的物体存在较大的多样性, 因此对重建算法提出了更高的要求.

3) 不适定问题. 基于单幅图像的三维重建本身就是一个不适定问题 (Ill-posed problem)[5], 即由于输入形式为单幅图像, 深度信息不可避免地丢失, 如果不给定一些先验知识或假设, 重建结果是不唯一的. 如何根据一些假设和先验知识来重建最适合的模型, 以及如何提供最少的假设和先验, 这对三维重建工作提出了不小的挑战.

针对上述基于单幅图像物体三维重建问题, 许多文献提出了相应的解决方案. 文献 [19−20] 探讨了对特定种类物体进行重建的问题; 文献 [16−17] 针对图像自身属性诸如非理想朗伯特模型的重建提出了解决办法; 此外有许多假设被提出[21−24] 以解决不适定问题. 但上述方法仍未能很好地解决这些问题, 重建精度仍然有待提高. 随着深度学习技术的不断发展, 很多基于深度学习的三维重建方法[8−9, 25] 近几年开始被提出, 且重建效果更好, 逐渐成为该领域近年来研究的重点.

本文结构安排如下: 第 1 节简要介绍传统的基于单幅图像物体三维重建的研究成果; 第 2 节介绍深度学习算法模型及其近年来在基于单幅图像的三维重建领域的研究进展; 第 3 节介绍物体三维重建的常用公共数据集; 第 4 节对该方向目前仍存在的问题提出思考和展望, 并对文章进行总结, 分析基于深度学习算法的优缺点.

1 传统的基于单幅图像的物体三维重建方法

基于单幅图像的物体三维重建在计算机视觉领域是一个长期存在且具有挑战性的问题, 往往利用先验知识或引入合适的约束来进行重建. 按照重建方法的不同, 传统方法可以分为基于模型的重建方法和基于几何外形恢复的重建方法两类.

1.1 基于模型的重建方法

一般而言, 基于模型的重建方法由要表示对象的参数模型组成, 通过找到模型的投影和输入图像之间最佳拟合时模型的参数来完成重建[26]. 基于模型表示的物体重建反映了对模型表示的不同偏好. 在早期的工作中, 广义柱体[27] 对柱类外形进行了紧凑地描述, 而基于多面体模型的方法[28−29] 则只能针对一些方形物体进行重建, 此外还有超二次曲面模型[30], 一些只针对车辆的手工刚性三维模型[31−32] 等. 这些模型都能对某种外形进行一定的描述, 但是可描述的对象太具有局限性. 基于 CAD 模型的方法[33−35] 可以粗略地表示物体的近似外形, 通过给予一组对应点, 可以非常有效地确定近似实例的视点, 但生成的模型和训练的模型有较大的偏差. 此外近期还出现了基于 CAD 模型的类似实例进行非参数化重建的方法[36], 但是该方法仅限于对预先分割好的在线商品图像进行重建.

近期, 一些可以变形的模型因更具有表现力引起了更多学者的注意. 形变模型 (Morphable model) 常用于对人脸进行重建[14, 37−38], 它是一种线性组合模型, 通过图像光流算法来建立三维人脸点到点的稠密对应. 通过调节模型参数使输入图像与形变模型匹配. 因为形变模型的建立往往是通过三维扫描的形式来获取三维信息, 为了克服对三维数据的要求, Cashman 等[15] 提出了一种混合的方法, 使用单个三维模型和大量二维信息学习得到动物的形变模型来重建诸如海豚这种较为复杂的动物模型. Vicente 等[19] 针对 PASCAL VOC 数据集中的图像进行重建, 提出了一种新方法, 先在图像数据集中找到与输入的单幅图像同类型的相似视角的不同物体的图像, 然后使用可视外壳进行重建, 但是需要在测试的时候添加关键点注释, 且处于理想的分割状态, 无法应用于较为真实的图像. Kar 等[20] 更进一步, 利用物体检测数据集中的 2D 注释来训练学习可变形的模型, 仅在训练的时候使用了部分注释, 且可以重建真实图像中的物体, 利用自底向上的

模块来补充高频外形细节, 重建效果较之前的方法有一定的提升.

基于模型的方法在针对特定类别物体的重建上能取得较好的效果, 因为这种方法中的先验知识在模型的设计阶段就已经被设定好, 能够较好地对针对的物体提供更多的先验信息, 但是这类方法很难扩展到其他物体上. 可变形的模型也往往只能沿着特定类别的变化模式变化. 表 1 列出了近年来两种常用的基准算法和一种不针对具体重建类别的方法 (均采用 Hausdorff 距离[7] 作为评价参数)[39] 在 PASCAL VOC 数据集上的三维重建结果, 其中 Hausdorff 距离越小代表精度越高.

1.2　基于几何外形恢复的方法

根据二维图像中的三维信息来恢复物体三维几何外形的技术统称为 Shape from X, X 可以是: 阴影 (Shading)、纹理 (Texture)、运动 (Motion)、光度立体 (Stereo)、轮廓 (Silhouette) 等. 基于光度立体、运动和轮廓恢复三维外形的方法常用于多目重建. 基于纹理和阴影恢复三维外形的方法常用于针对单幅图像的三维重建.

从纹理中恢复外形 (Shape from texture)[40] 往往要求假定纹理满足某种先验性质, 例如假设纹理分布具有均一性[41], 或要求纹理由明确的纹理基元组成[42]. 从纹理中恢复外形的方法重建精度相对较低, 并且适用性窄, 实际应用相对较少.

从阴影中恢复外形 (Shape from shading, SFS)[43] 主要利用物体表面的明暗变化解析物体表面的矢量信息, 最后转化为深度信息. 通常是在假定理想光照下, 即满足朗伯特 (Lambertian) 反射模型的状态下进行重建, 但是在满足假定朗伯特反射模型状态下的 SFS 问题本身也是不适定的 (Ill-posed), 因此需要引入相应的附加条件对其正则化. Ikeuchi 等[22] 在 SFS 中加入平滑度约束, 即假定物体表面是光滑的, 以此使问题变为适定, 但这对于具有分形特征的自然景物三维外形恢复效果仍不太理想.

大多数传统的 SFS 方法是基于正交投影[44−45], 且假设光源都在无穷远处. 而透视投影因比正交投影更为精准, 慢慢被引入到 SFS 方法中[46−47]. 同时, 使用朗伯特模型的三维重建误差较大, 为了提高重建精度, 许多非朗伯特模型被提出来. Ahmed 等[17] 用 Ward 反射模型对三维外形恢复进行了研究, Bakshi 等[16] 将 SFS 方法应用到包含有漫反射和镜面反射两种情况的混合表面重建.

此外, 还有一些方法通过监督学习的方式来学习几何信息[48−50], 以此来预测深度图, 但是对深度图的估计往往针对的是场景三维重建[51], 而本文主要针对的是对物体三维重建的综述, 因此对此类方法以及一些其他基于场景的三维重建方法[52−54], 本文不再做详细论述.

基于几何外形恢复的方法往往具有更好的泛化性, 其重建不是针对特定类别的物体, 能够以较自然简单的方式提取物体的表面信息. 但同时该类方法往往对光照和灰度提出了较高的要求, 通过理想光源之类的一些约束来使重建的解唯一. 因此该方法往往难以对真实图像进行较好质量的重建.

2　基于深度学习的单幅图像三维重建

2.1　深度学习及其模型简介

深度学习的概念源于对人工神经网络 (Artificial neural network, ANN) 的研究. 它是一种特征学习的方法, 把低层次的原始数据通过一些简单而非线性的模型转化成为更高层次的表达[55]. 通过大量的转换组合, 得到更好的特征表示. 早在 1986 年, Rumelhart 等[56] 就提出人工神经网络的反向传播 (Back propagation, BP) 算法, 但这一时期人们普遍认为梯度下降会陷入局部极值, 且存在梯度消失、硬件条件不足等问题, 直到 2006 年, Hinton 等[57] 介绍了一种新的深度神经网络模型 DBN 及训练方法, 降低了深度神经网络的优化难度, 利用预训练方法缓解了局部极值问题, 从此深度学习受到学术界的关注. 之后 LeCun、Bengio、Ng 等对深度神经网络展开研究[58]. 随着一些新的模型训练方法的涌现, 深度学习在诸如语音识别[59−60]、自然语言处理[61−63]、图像识别和分割等[64−65] 多个领域都取得

表 1　不同方法对 PASCAL VOC 数据集图像中的物体重建的结果对比[20]

Table 1　Comparison of different methods on the PASCAL VOC[20]

方法	飞机	单车	轮船	公交	汽车	椅子	摩托	沙发	火车	电视	均值
Twarog 等[39]	9.73	10.39	11.68	15.40	11.77	8.58	8.99	8.62	23.68	9.45	11.83
Vicente 等[19]	5.07	6.03	8.80	8.76	4.38	5.74	4.86	6.49	17.52	8.37	7.60
Kar 等[20]	5.00	6.27	9.94	6.22	5.18	5.20	4.98	6.58	12.60	9.64	7.16

了较大的进展. 近年来, 深度学习在三维数据的分类、识别和重建上也取得了很大的进展[66−68]. 目前广泛应用的深度学习模型主要包括深度置信网络 (Deep belief network, DBN)[57, 69]、堆叠自动编码器 (Stacked auto-encoders, SAE)[70]、卷积神经网络 (Convolutional neural networks, CNN)[71]、循环神经网络 (Recurrent neural networks, RNN)[72] 等.

2.2 深度学习在基于单幅图像三维重建中的应用

相较于二维图像领域, 深度学习在三维外形重建上的研究起步较晚, 但在近三年内也取得了较大的进展. 本节依据三维外形的不同表示, 从基于体素表示和基于点云、网格表示两个方面介绍深度学习在三维重建中的研究现状.

2.2.1 基于体素表示的三维重建

随着深度学习在三维领域的不断扩展, 围绕深度学习研究基于体素的三维重建方法开始被提出, 利用体素化的方法将所有的 CAD 模型表示为二值或实值的三维张量, 保证了模型大小的相同. Wu 等[67] 建立的网络结构 3D shapenets 是较早提出的基于体素表示的三维重建网络, 其利用深度卷积置信网络 (CDBN) 将三维几何外形表示为三维体素上二值变量的概率分布, 输入深度图, 通过吉布斯采样 (Gibbs sampling) 不断预测外形类型和填补未知的体素来完成重建. 为了得到更好的训练效果, 其同时建立了大型的 CAD 模型数据集 ModelNet. Choy 等[8] 提出了一种基于标准 LSTM 的扩展网络结构 3D-R2N2 (3D recurrent reconstruction neural network), 使用该网络学习二维图像与三维外形间的映射, 网络以端到端的形式获取一个或多个对象实例的图像, 首先利用一个标准的 CNN 结构对原始输入图像进行编码, 用其提出的 3D-LSTM 进行过渡连接, 3D-LSTM 单元排列成三维网格结构, 每个单元接收一个从编码器中得到的特征向量, 并将他们输送到解码器中. 这样每个 3D-LSTM 单元重构输出体素的一部分. 再利用一个标准反卷积网络对其解码, 通过这样的网络结构建立了二维图像和三维模型的映射. 该方法还在单个框架中统一了单视图和多视图重建, 且不需要图像注释或分类标签进行训练, 克服了过去无法解决的缺乏纹理和宽基线特征匹配等问题的挑战. 通过以 IoU (Intersection-over-union) 作为评价重建效果指标的实验, 验证了在单幅图像的三维重建效果优于 Kar 等[20] 的传统方法, 但该方法在重建椅子细腿等方面存在断裂失真的问题. Girdhar 等[73] 提出了一种名为 TL-embedding network 的网络结构, 该网络的自编码器以 $20 \times 20 \times 20$ 的像素网格表示作为输入, 通过自编码学习三维模型的嵌入 (Embedding), 形成一个 64 维的嵌入空间 (Embedding space), 然后通过 ConvNets 输入二维图像, 找到对应的嵌入, 最后通过解码器得到体素表示的三维模型. 在重建结果上, 更能抓住重建的细节, 例如椅子的腿部和靠背, 重建效果优于 Kar 等[20] 的方法. Kar 等[74] 尝试在同一系统中统一单视图和多视图的三维重建, 提出了一种叫做立体学习机 (Learnt stereo machine, LSM) 的新系统, 其可以利用单视角和语义线索进行单视图三维重建, 同时也可以利用立体视觉整合来自不同视角的信息进行多视图重建. 该方法在编码部分提取特征后, 加入一个反投影模块, 将由前馈卷积神经网络获取的二维图像中的特征投影到三维网格中, 并使得结果根据极线约束在三维网格中对齐, 简化了特征匹配. 通过实验与 3D-R2N2[8] 的 IoU 值对比, 无论在单视图还是多视图均取得了更好的效果, 并且在实验中即便只给出飞机和椅子的数据, 还是可以完成汽车模型的重建, 因而具有较好的泛化能力. Wu 等[75] 等提出了一种叫 MarrNet 的网络模型, 在端到端生成重建结果的网络结构中加入了生成 2.5 D 草图的步骤, 增强了重建效果并使得网络可以更轻松地针对不同类别的物体进行重建.

早期的工作主要基于监督学习, 但获得大规模监督数据的成本往往是巨大的, 随着研究的深入, 一些基于生成模型的弱监督学习和无监督学习的方法逐渐被提出. Kanazawa 等[76] 提出了一种新的网络结构 WarpNet, 利用薄板样条插值 (Thin-Platespline) 进行转换, 从一幅鸟的图像变形得到另一幅鸟的图像, 得到一个人工的对应, 通过将这样的两幅图像作为原始图和目标图来学习其中的变化, 最后将通过网络学习得到的结果作为空间先验来匹配图像中的外表变化、视点和关节, 不需要部分注释来进行单视图重建. Tulsiani 等[77−78] 采用另一种监督形式, 通过学习单视角的三维结构来构建多视角观察 (Multi-view observations), 再通过多视角观察得到几何一致性, 其利用经典射线一致性公式引入了一个一般的检验器, 可以测量 3D 外形与不同种类观测结果间的一致性. Rezende 等[1] 首次提出了一个无监督的生成模型, 在二维图像上可以进行端到端的无监督训练, 不需要真实的三维标签, 证明了无监督生成模型学习三维表征的可能性. 在此基础上, Yan 等[79] 提出一个名为 Perspective transformer nets 的网络结构, 在传统的编解码卷积神经网络中加入了透视变换作为正则化, 在不知道对应的真实模型的情况下, 提出了一种轮廓损失函数, 通过透视变换, 将在不同特定视角下的二维物体轮廓和对应体素轮廓的距离作为新的损失函数, 该方法在无监督学习下的重建具有良好的泛化能力. 此外, 一些

学者利用生成对抗网络进行重建[80−83]. Wu 等[84] 提出了 3D-VAE-GAN 的网络结构, 输入单幅图像, 通过变分自编码网络的编码器得到图像的潜在向量 (Latent vector), 再通过生成对抗网络的生成器得到重建的物体. 使用生成对抗网络的优点是可以从高斯或均匀分布等概率表征空间中采样新的三维对象, 并且判别器 (Discrimitator) 带有三维物体识别的信息特征. 该方法与 TL-embedding network 的重建精度对比, 取得了更好的效果. Zhu 等[82] 对图像中物体的二维轮廓使用了更简单的标注, 对 TL-embedding network 和 3D-VAE-GAN 网络进行了微调, 重建取得了更好的效果. Gadelha 等[85] 提出了一种 Projective GANs (PrGANs) 的生成对抗网络, 在生成器上加入了投影模块, 投影模块通过给定视角呈现体素形状来捕获三维表示, 而后转化为二维图像再传递给判别器, 通过判别器判定输入图像是生成的还是真实的. 通过反复训练, 调整生成器, 改进了生成的三维体素外形. 增加投影模块使该方法与之前 3D-VAE-GAN 网络需要联合三维数据相比, 在学习阶段不使用任何标注、三维信息或视角信息来推断潜在的三维外形分布. Rosca 等[81] 对 AE-GANs 的网络结构进行了改进, 提出了一种新的变分自编码器和 GANs 结合的方法 α-GAN, 融合两种方法的优势, 构建新的优化目标函数, 重建也取得了较好的效果.

体素表示的三维物体相较于二维图像, 计算量更大, 需求内存更多, 往往因计算和内存的限制, 分辨率主要为 $32 \times 32 \times 32$ 以下. 针对这一问题, 一些基于八叉树的卷积神经网络被提出[86−88]. Riegler 等[89] 提出了一种卷积网络 OctNet, 取代体素部分, 将三维空间分割成一组不平衡八叉树, 每个八叉树根据数据的密度来分割三维空间. 其充分利用了三维输入数据的稀疏性, 从而能够更加合理地使用内存及计算. 受此启发, Häne 等[90] 提出了一个叫做层次表面预测 (Hierarchical surface prediction, HSP) 的通用框架, 将体素分为占用、未占用和边界三类. 使用这种方法, 在一个八叉树中分层次地预测从粗到细多分辨率的体素块, 只要保证在那些标记为边界的区域有相对较高的分辨率即可. 通过迭代, 可以层进地预测出分辨率为 $256\times256\times256$ 的体素表示. 同样是使用八叉树结构, 与 Riegler 等[89] 提出的方法中需要假设在测试期间八叉树结构为已知的不同, Tatarchenko 等[88] 提出了一种称作 OGN (Octree generating networks) 的网络结构, 通过网络学习预测八叉树的结构, 同时在网络的解码初期预测大量的输出空间, 而直到网络的某一层, 密集的网格才被八叉树替代, 从而节省了后续高分辨率计算需要的内存, 并且可以将分辨率提升为 $512 \times 512 \times 512$. Sun 等[87] 提出了一种称作 CVN (Colorful voxel network) 的网络结构, 这是第一个基于深度学习的能够端到端同时从单一图像恢复三维外形和表面颜色的网络结构, 设计了一种新的损失函数 MSFCEL (Mean squared false cross-entropy loss) 用于解决体素表示的稀疏问题, 从而能够生成更高分辨率的结果.

ShapeNet 团队组织了一次基于单幅图像物体三维重建的挑战赛[6], 共 3 支队伍参加, 包括上文提到的 HSP[90] 和 α-GAN[81]. 每个队伍从测试图像重建出分辨率为 $256 \times 256 \times 256$ 的三维模型, 挑战赛采用 IoU 和 CD 两种评价标准. 在与 3D-R2N2[8] 结果的对比中, HSP 在基于 IoU 的评价标准中赢得第一, 而 α-GAN 在基于 CD 的评价标准中赢得第一, ShapeNet 团队猜测原因是 gan 损失比交叉熵损失更有助于描绘几何的正确性[6].

2.2.2 基于点云和网格表示的三维重建

基于点云和网格的单幅图像三维重建工作目前还比较少, 原因在于相较于可以直接用于卷积神经网络中的体素表示, 点云和网格表示则需要一定的改变. 例如, 系统若需处理点云时一般需要维持点顺序不变. 随着一些基于深度学习和点云形式的物体识别的相关工作的出现[91−93] 和相关研究的不断推进, Fan 等[9] 提出了一个点集生成网络, 这是第一个用深度学习研究点云表示点集产生的网络结构. 它有多个平行的预测分支, 网络结构中包含卷积模块、反卷积模块、全连接模块. 这样复杂的模型具有高度的灵活性, 在描述复杂结构方面表现出色, 而由于卷积层和反卷积层引起的空间连续性, 其对大光滑表面更友好. 而该网络引入了 Hourglass 卷积网络结构[94] 反复进行的编解码操作, 使该方法具有更强的表示能力, 可以更好地联合全局和局部信息. 其系统地探讨了点云生成网络的损失函数设计, 选取了两种距离 Chamfer distance (CD) 和 Earth Mover′s distance (EMD) 作为候选. 在重建结果上, 该方法能产生多个可能的输出来解决单幅图像三维重建的不适定问题, 在与 3D-R2N2 方法的结果对比中, 该方法在所有类别中均能获得更高的 IoU 值, 拥有更好的重建效果, 但是在输入图像中有多个对象的情况下, 由于网络还没有采取任何检测或注意力机制, 网络会产生扭曲的输出. 并且其所需的可学习参数与三维点预测的数量成线性比例且不能很好地缩放, 使用三维距离度量作为优化标准对于大量点来说仍是困难的. Lin 等[95] 针对上述问题, 在网络结构中使用了二维卷积运算来捕获生成的点云之间的相关性并以更易于计算的方式进行优化, 生成的点云具有更好的精度.

图像和网格之间的转换产生的离散操作会阻碍反向传播的过程, 导致基于深度学习重建网格表示的三维模型面临不小的挑战. Kato 等[25] 针对这个问题提出了一种渲染网格的近似梯度, 将该部分作为一个渲染器集成到神经网络中. 经过渲染器处理, 其使用轮廓图像监督来执行单图像 3D 网格重建. 通过对比 Yan 等[79] 基于体素的重建方法, 验证了其在视觉和 IoU 值方面均超过了基于体素的方法, 但该方法存在一个明显的不足, 即不能生成一个具有各种拓扑的对象. Pontes 等[96] 提出了一个新的学习框架, 通过学习框架推断网格表示的参数来解决基于网格重建所面临的问题, 其在面对输入为真实世界的单幅图像时表现更好. Wang 等[97] 将网络分为特征提取和网格变形两个部分, 先由 2D CNN 部分提取特征, 再利用提取的特征通过 GCN (Graph convolutional network) 来解决网格结构无法直接作用于 CNN 的问题, 最后生成重建模型. 该文章对比了基于体素的 3D-R2N2[8]、Fan 等[9] 基于点云及 Kato[25] 基于网格的方法, 实验中重建效果均高于上述三种方法, 但仍存在只能生成相同拓扑网格的局限性.

3 基于单幅图像三维重建的常用数据集

为了更好地研究基于单幅图像的物体三维重建, 构建大规模的三维模型数据集成为必然要求. 目前有多个三维模型的公共数据集供科研人员使用.

1) PASCAL 3D+ 数据集[98]

PASCAL VOC 数据集是在图像识别、图像分割和目标检测等领域经常使用的大型数据集, 它的广泛使用也推动了计算机视觉领域的不断发展. 而 PASCAL 3D+ 正是基于 PASCAL VOC 2012[99] 的 12 种刚体类别的图像, 为它们添加了三维模型标注的数据集, 其每一类通过 ImageNet[100] 扩展得到更多的图像, 最终每一类平均有 3 000 左右的物体. 该数据集图像物体变化较大, 且包含遮挡和截断等情况, 能够更好地反映真实世界中的变化.

2) ShapeNet 数据集[101]

该数据集由物体的三维 CAD 模型组成, 是目前为止包含丰富注释的最大的三维模型数据集. 其在 WordNet[102] 分类下进行组织, 为每一个三维模型提供丰富的语义注释, 包括物理尺寸、关键字等, 注释可通过基于 Web 的界面提供, 以实现对象属性的数据可视化. ShapeNet 共包含超过 300 万个模型, 其中 22 万个模型被归类为 3 135 个类别.

3) Online Products 数据集[103]

该数据集包含在线销售的 23 000 个物体的图像. 由于存在宽基线的问题, 传统的 MVS 和 SFM 方法无法通过这些图像进行重建.

4) ModelNet 数据集[67]

该数据集是当前规模较大、模型类别较多的一个大型 CAD 数据集, 收集了各类 3D CAD 网站, 3D Warehouse 以及 Princeton Shape Benchmark[104] 660 种共计 151 125 个 CAD 模型.

5) IKEA Dataset 数据集[33]

该数据集收集了来自 Google 3D Warehouse 的 225 个 IKEA 的三维模型和从 Flickr 得到的 800 幅图像, 分为 IKEA 家具和 IKEA 房间两个部分, 主要集中了室内家具的模型, 模型类别及数量相对较少, 同时部分图像存在遮挡. 该数据集的每一幅图像都标注其关联的三维模型, 可以借此评估三维重建的效果.

4 思考、展望与结论

随着深度学习的不断发展和三维数据集的不断完善, 基于单幅图像的三维重建取得了较大的进展, 表 2 展示了目前代表性传统方法[20] 和 3D-R2N2 在 PASCAL 3D+ 数据集上以 IoU 值作为重建评价标准的重建效果对比. 可以看出与传统手工设计的方法相比, 基于深度学习的端到端的训练方法能够直接以单幅图像作为输入, 并以重建的三维模型作为输出, 提取特征效率更高, 重建效果更好. 同时深度学习使用诸如 dropout 等稀疏化网络参数的方法来防止过拟合, 以此来利用大规模的数据, 具有更好的泛化性. 正如人看到二维图像即可联想到它的三维表示, 基于深度学习的单幅图像重建也越来越趋向于与人类认知三维物体方法相同的无监督学习[1, 79, 84], 也有越来越多的网络融合了单幅图像和

表 2 现有的传统方法与 3D-R2N2 重建结果的对比[8]

Table 2 Comparison of traditional methods and 3D-R2N2[8]

方法	飞机	单车	轮船	公交	汽车	椅子	摩托	沙发	火车	电视	均值
Kar 等[20]	0.298	0.114	0.188	0.501	0.472	0.234	0.361	0.149	0.249	0.492	0.318
Choy 等[8]	0.544	0.499	0.560	0.816	0.699	0.280	0.649	0.332	0.672	0.574	0.571

多幅图像两种方式, 使得重建能够更加灵活. 同时基于深度学习的方法也不断地在各种三维表示形式上进行着尝试, 表 3 对比了目前基于体素、点云、网格的主流方法在 ShapeNetCore[77] 数据集上以平均 IoU 值作为重建评价准则的重建精度.

综上所述, 基于深度学习的方法相较于传统的方法拥有较多的优势, 并且在这一领域逐渐取得了显著进展, 但是同时在这一领域也存在如下问题:

1) 公共数据集较小. 对于一个三维重建任务来说, 增加训练数据的种类和规模可以增加学习的泛化能力. 但是与目前千万级的二维图像数据集相比, 三维公共数据集规模小、种类少. 即使是近年来发布的较大的数据集 ModelNet 也仅包含了来自 662 个类的 127 915 个三维外形. 相信随着深度学习在三维领域的不断深入, 在未来会涌现出更大规模的三维公共数据集.

2) 重建分辨率及精度问题. 三维物体相较于二维多了一个维度, 基于体素的重建随着重建分辨率的增加, 物体体积成立方体增长, 使其受限于计算和内存, 重建物体常见的分辨率是 $32 \times 32 \times 32$. 这样分辨率的重建结果是非常粗糙的, 离真实物体还有较大差距. 即使有针对这一问题提出的改进方法[90], 改进后仍然无法达到较为精密的重建效果. 而在以主要依赖于大规模多样性标记数据集的监督学习的方法中, 在实验中与真实模型对比, 重建精度也未达到 0.85 以上. 要提高基于体素重建的分辨率, 还要考虑三维体素的稀疏性, 未来针对如何在基于体素的重建中提升计算效率, 避免在未占用的部分浪费过多内存, 提高重建的分辨率以及如何改善网络结构以提高重建效果, 能够恢复更多细节, 这些仍然是未来值得关注的问题.

3) 基于点云和网格重建的问题. 图像是结构化的, 可以表示为二维平面上的一个矩阵, 基于体素的重建使模型通过体素化变为二值模式, 也保证了大小的相同. 但三维点云和网格都是不规则的数据形式, 这使得学习方法的应用存在问题, 由于欧几里德卷积运算不能直接应用, 这些数据表示不能很好地适应传统的 CNN. 目前针对该问题仅有少数前期探索工作[9], 主要思路有: a) 先将三维点云或网格数据转化成二维图像, 再在神经网络中提取特征; b) 设计适应原始三维数据特点的网络模型, 例如结合 GCN 的网络模型; c) 在三维外形上手工提取低级特征, 再采用深度学习模型提取高级特征. 但总体而言, 该问题仍未得到有效解决. 如何设计能适应原始三维数据特点的深度学习模型, 以及如何设计点云生成网络的损失函数仍是未来一个值得研究的课题. 而基于网格的重建, 如何生成具有不同拓扑的对象, 是一个具有重要意义的研究方向.

4) 单幅图像重建的不适定问题. 正如在传统方法中提到的, 对一幅图像的三维重建, 特别是对一幅来自真实世界的图像 (区别于从 CAD 模型中生成的二维图像), 其不可见部分的几何外形需要猜测, 一幅图像往往可能对应多个不同的重建模型, 且均可以看作是该幅图像的重建结果. 从统计的角度来看, 输入图像的合理预测形成一个分布. 反映在训练集中, 两个看起来相似的图像可能具有相当不同的重建结果. 如果将这个问题看作是一个回归问题, 就与传统的只有一个对应真实模型的三维重建大有不同, 定义正确的损失函数就显得尤为重要. 针对这一问题, Fan 等[9] 通过 VAE 网络结构和其定义的 MoN 损失使得网络能对单幅图像生成多种可能的重建结果, 该方法进行了一次有益的尝试, 但在实现细节和准确度上仍有提高的空间.

5) 三维模型的表示形式和评价指标. 与深度学习在二维图像中的应用不同, 目前人们仍然还在探索什么样的三维表示是最准确有效的, 因此基于体素、网格、点云表示的方法也仍然在不断涌现. 而在对基于单幅图像的三维重建的评价标准上, 至今也仍没有一个完全统一的定论[6], 哪种评价指标最能够反映重建的效果, 仍然有待进一步的研究.

本文综述了近年来深度学习在单幅图像三维重建中的应用和展望. 首先说明了传统的基于单幅图像的三维重建的方法和常用的公共数据集, 然后重点介绍了深度学习方法在基于单幅图像的三维重建的最新应用进展, 最后对深度学习在基于单幅图像的三维重建进行了分析, 对未来的发展趋势进行了思考与展望. 总体而言, 深度学习为解决基于单幅图像的三维重建提供了新的技术, 取得了较为显著的科研成果, 但其研究大部分仍存在大量的问题, 未来基于深度学习的单幅图像的三维重建仍然是一个亟

表 3　不同方法以平均 IoU 值作为评价标准的重建精度对比

Table 3　3D reconstruction comparison with different methods using IoU

	Choy 等[8]	Yan 等[79]	Kar 等[74]	Fan 等[74]	Kato 等[74]
IoU 均值	0.556	0.574	0.605	0.640	0.602

待研究的重点方向.

致谢

感谢英国萨里大学视觉、语音和信号处理中心 Evren Imre 博士 (现工作于动作捕捉公司 Vicon) 对本文提出的建设性意见.

References

1 Rezende D J, Ali Eslami S M, Mohamed S, Battaglia P, Jaderberg M, Heess N. Unsupervised learning of 3D structure from images. In: Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016). New York, USA: Curran Associates, Inc., 2016. 4996−5004

2 Haming K, Peters G. The structure-from-motion reconstruction pipeline — a survey with focus on short image sequences. *Kybernetika*, 2010, **46**(5): 926−937

3 Lhuillier M, Quan L. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, **27**(3): 418−433

4 Habbecke M, Kobbelt L. A surface-growing approach to multi-view stereo reconstruction. In: Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, MN, USA: IEEE, 2007. 1−8

5 Oswald M R, Töppe E, Nieuwenhuis C, Cremers D. A review of geometry recovery from a single image focusing on curved object reconstruction. *Innovations for Shape Analysis*. Berlin, Germany: Springer-Verlag, 2013. 343−378

6 Yi L, Shao L, Savva M, Huang H B, Zhou Y, Wang Q R, et al. Large-scale 3D shape reconstruction and segmentation from ShapeNet Core55. arXiv preprint arXiv: 1710.06104, 2017.

7 Aspert N, Santa-Cruz D, Ebrahimi T. MESH: measuring errors between surfaces using the Hausdorff distance. In: Proceedings of the 2002 IEEE International Conference on Multimedia and Expo. Lausanne, Switzerland: IEEE, 2002. 705−708

8 Choy C B, Xu D F, Gwak J Y, Chen K, Savarese S. 3D-R2N2: a unified approach for single and multi-view 3D object reconstruction. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer, 2016. 628−644

9 Fan H Q, Su H, Guibas L. A point set generation network for 3D object reconstruction from a single image. In: Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA: IEEE, 2017. 2463−2471

10 Kemelmacher-Shlizerman I, Basri R. 3D face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, **33**(2): 394−405

11 Wang H K, Stout D B, Chatziioannou A F. Mouse atlas registration with non-tomographic imaging modalities-a pilot study based on simulation. *Molecular Imaging and Biology*, 2012, **14**(4): 408−419

12 Dworzak J, Lamecker H, Von Berg J, Klinder T, Lorenz C, Kainmüller D, et al. 3D reconstruction of the human rib cage from 2D projection images using a statistical shape model. *International Journal of Computer Assisted Radiology and Surgery*, 2010, **5**(2): 111−124

13 Baka N, Kaptein B L, De Bruijne M, Van Walsum T, Giphart J E, Niessen W J, et al. 2D-3D shape reconstruction of the distal femur from stereo X-ray imaging using statistical shape models. *Medical Image Analysis*, 2011, **15**(6): 840−850

14 Blanz V, Vetter T. A morphable model for the synthesis of 3D faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. New York, USA: ACM Press, 1999. 187−194

15 Cashman T J, Fitzgibbon A W. What shape are dolphins? Building 3D morphable models from 2D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(1): 232−244

16 Bakshi S, Yang Y H. Shape from shading for non-Lambertian surfaces. In: Proceedings of the 1st International Conference on Image Processing. Austin, TX, USA: IEEE, 1994. 130−134

17 Ahmed A, Farag A. Shape from shading for hybrid surfaces. In: Proceedings of the 2007 IEEE International Conference on Image Processing. San Antonio, TX, USA: IEEE, 2007. II-525−II-528

18 Jin H L, Soatto S, Yezzi A J. Multi-view stereo reconstruction of dense shape and complex appearance. *International Journal of Computer Vision*, 2005, **63**(3): 175−189

19 Vicente S, Carreira J, Agapito L, Batista J. Reconstructing PASCAL VOC. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 41−48

20 Kar A, Tulsiani S, Carreira J, Malik J. Category-specific object reconstruction from a single image. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 1966−1974

21 Prasad M, Zisserman A, Fitzgibbon A W. Fast and controllable 3D modelling from silhouettes. In: Proceedings of the 2005 Eurographics. Hamburg, Federal Republic of Germany: Elsevier Science Publishing Company, 2005. 9−12

22 Ikeuchi K, Horn B K P. Numerical shape from shading and occluding boundaries. *Artificial Intelligence*, 1981, **17**(1−3): 141−184

23 Prasad M, Fitzgibbon A. Single view reconstruction of curved surfaces. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR′06). New York, NY, USA: IEEE, 2006. 1345−1354

24 Daum M, Dudek G. On 3-D surface reconstruction using shape from shadows. In: Proceedings of the 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Santa Barbara, CA, USA: IEEE, 1998. 461−468

25 Kato H, Ushiku Y, Harada T. Neural 3D mesh renderer. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 37−44

26 Rother D, Sapiro G. Seeing 3D objects in a single 2D image. In: Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009. 1819−1826

27 Nevatia R, Binford T O. Description and recognition of curved objects. *Artificial Intelligence*, 1977, **8**(1): 77−98

28 Gupta A, Efros A A, Hebert M. Blocks world revisited: image understanding using qualitative geometry and mechanics. In: Proceedings of the 11th European Conference on Computer Vision. Heraklion, Crete, Greece: Springer-Verlag, 2010. 482−496

29 Xiao J X, Russell B C, Torralba A. Localizing 3D cuboids in single-view images. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA: Curran Associates Inc., 2012. 746−754

30 Pentland A P. Perceptual organization and the representation of natural form. *Artificial Intelligence*, 1986, **28**(3): 293−331

31 Haag M, Nagel H H. Combination of edge element and optical flow estimates for 3D-model-based vehicle tracking in traffic image sequences. *International Journal of Computer Vision*, 1999, **35**(3): 295−319

32 Koller D, Daniilidis K, Nagel H H. Model-based object tracking in monocular image sequences of road traffic scenes. *International Journal of Computer Vision*, 1993, **10**(3): 257−281

33 Lim J J, Pirsiavash H, Torralba A. Parsing Ikea objects: fine pose estimation. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, NSW, Australia: IEEE, 2013. 2992−2999

34 Satkin S, Rashid M, Lin J, Hebert M. 3DNN: 3D nearest neighbor. *International Journal of Computer Vision*, 2015, **111**(1): 69−97

35 Pepik B, Stark M, Gehler P, Ritschel T, Schiele B. 3D object class detection in the wild. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Boston, MA, USA: IEEE, 2015. 1−10

36 Huang Q X, Wang H, Koltun V. Single-view reconstruction via joint analysis of image and shape collections. *ACM Transactions on Graphics (TOG)*, 2015, **34**(4): Article No. 87

37 Liu F, Zeng D, Li J, Zhao Q J. Cascaded regressor based 3D face reconstruction from a single arbitrary view image. [Online], available: https://arxiv.org/abs/1509.06161v1, March 25, 2019

38 Blanz V, Vetter T. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003, **25**(9): 1063−1074

39 Twarog N R, Tappen M F, Adelson E H. Playing with puffball: simple scale-invariant inflation for use in vision and graphics. In: Proceedings of the 2012 ACM Symposium on Applied Perception. Los Angeles, California, USA: ACM, 2012. 47−54

40 Aloimonos J. Shape from texture. *Biological Cybernetics*, 1988, **58**(5): 345−360

41 Marinos C, Blake A. Shape from texture: the homogeneity hypothesis. In: Proceedings of the 3rd International Conference on Computer Vision. Osaka, Japan: IEEE, 1990. 350−353

42 Loh A M, Hartley R I. Shape from non-homogeneous, non-stationary, anisotropic, perspective texture. In: Proceedings of the 2005 British Machine Vision Conference. Oxford, UK: BMVC, 2005. 69−78

43 Horn B K P. *Obtaining Shape from Shading Information*. Cambridge: MIT Press, 1989. 123−171

44 Robles-Kelly A, Hancock E R. An eigenvector method for shape-from-shading. In: Proceedings of the 12th International Conference on Image Analysis and Processing. Mantova, Italy: IEEE, 2003. 474−479

45 Cheung W P, Lee C K, Li K C. Direct shape from shading with improved rate of convergence. *Pattern Recognition*, 1997, **30**(3): 353−365

46 Yang L, Han J Q. 3D shape reconstruction of medical images using a perspective shape-from-shading method. *Measurement Science and Technology*, 2008, **19**(6): Article No. 065502

47 Tankus A, Kiryati N. Photometric stereo under perspective projection. In: Proceedings of the 10th IEEE International Conference on Computer Vision. Beijing, China: IEEE, 2005. 611−616

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

49 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*, 2009, **31**(5): 824−840

50 Delage E, Lee H, Ng A Y. A dynamic Bayesian network model for autonomous 3D reconstruction from a single indoor image. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR′06). New York, USA: IEEE, 2006. 2418−2428

51 Tulsiani S, Kar A, Carreira J, Malik J. Learning category-specific deformable 3D models for object reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(4): 719−731

52 Wang Wei, Gao Wei, Zhu Hai, Hu Zhan-Yi. Rapid and robust piecewise planar reconstruction of urban scenes. *Acta Automatica Sinica*, 2017, **43**(4): 674−684
(王伟, 高伟, 朱海, 胡占义. 快速鲁棒的城市场景分段平面重建. 自动化学报, 2017, **43**(4): 674−684)

53 Miao Jun, Chu Jun, Zhang Gui-Mei, Wang Lu. Dense multi-planar scene reconstruction from sparse point cloud. *Acta Automatica Sinica*, 2015, **41**(4): 813−822
(缪君, 储珺, 张桂梅, 王璐. 基于稀疏点云的多平面场景稠密重建. 自动化学报, 2015, **41**(4): 813−822)

54 Zhang Feng, Shi Li-Min, Sun Feng-Mei, Hu Zhan-Yi. An image based 3D reconstruction system for large indoor scenes. *Acta Automatica Sinica*, 2010, **36**(5): 625−633
(张峰, 史利民, 孙凤梅, 胡占义. 一种基于图像的室内大场景自动三维重建系统. 自动化学报, 2010, **36**(5): 625−633)

55 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, **521**(7553): 436−444

56 Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors. *Nature*, 1986, **323**(6088): 533−536

57 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. *Science*, 2006, **313**(5786): 504−507

58 Jiao Li-Cheng, Yang Shu-Yuan, Liu Fang, Wang Shi-Gang, Feng Zhi-Xi. Seventy years beyond neural networks: retrospect and prospect. *Chinese Journal of Computers*, 2016, **39**(8): 1697−1716
(焦李成, 杨淑媛, 刘芳, 王士刚, 冯志玺. 神经网络七十年: 回顾与展望. 计算机学报, 2016, **39**(8): 1697−1716)

59 Feng X, Zhang Y D, Glass J. Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE, 2014. 1759−1763

60 Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, Canada: IEEE, 2013. 6645−6649

61 Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland: ACM, 2008. 160−167

62 Huang E H, Socher R, Manning C D, Ng A Y. Improving word representations via global context and multiple word prototypes. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea: Association for Computational Linguistics, 2012. 873−882

63 Mikolov T, Chen K, Corrado G S, Dean J. Efficient estimation of word representations in vector space. [Online], available: http://www.oalib.com/paper/4057741, March 25, 2019

64 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA: Curran Associates Inc., 2012. 1097−1105

65 Le Q V. Building high-level features using large scale unsupervised learning. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, Canada: IEEE, 2013. 8595−8598

66 Socher R, Huval B, Bath B, Manning C D, Ng A Y. Convolutional-recursive deep learning for 3D object classification. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA: Curran Associates Inc., 2012. 656−664

67 Wu Z R, Song S R, Khosla A, Yu F, Zhang L G, Tang X O, et al. 3D shapeNets: a deep representation for volumetric shapes. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 1912−1920

68 Gupta S, Girshick R, Arbeláez P, Malik J. Learning rich features from RGB-D images for object detection and segmentation. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer-Verlag, 2014. 345−360

69 Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006, **18**(7): 1527−1554

70 Schölkopf B, Platt J, Hofmann T. Greedy layer-wise training of deep networks. In: Proceedings of the 19th International Conference on Neural Information Processing Systems. Canada: MIT Press, 2006. 153−160

71 LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**(11): 2278−2324

72 Williams R J, Zipser D. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1989, **1**(2): 270−280

73 Girdhar R, Fouhey D F, Rodriguez M, Gupta A. Learning a predictable and generative vector representation for objects. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer-Verlag, 2016. 484−499

74 Kar A, Hane C, Malik J. Learning a multi-view stereo machine. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017). New York, USA: Curran Associates, Inc., 2017. 364−375

75 Wu J J, Wang Y F, Xue T F, Sun X Y, Freeman W T, Tenenbaum J B. MarrNet: 3D shape reconstruction via 2.5D sketches. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017). New York, USA: Curran Associates, Inc., 2017. 8−15

76 Kanazawa A, Jacobs D W, Chandraker M. WarpNet: weakly supervised matching for single-view reconstruction. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 3253−3261

77 Tulsiani S, Zhou T H, Efros A A, Malik J. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, Hawaii, USA: IEEE, 2017. 209−217

78 Tulsiani S. Learning Single-view 3D Reconstruction of Objects and Scenes [Ph. D. dissertation], UC Berkeley, USA, 2018

79 Yan X C, Yang J M, Yumer E, Guo Y J, Lee H. Perspective transformer nets: learning single-view 3D object reconstruction without 3D supervision. In: Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016). New York, USA: Curran Associates, Inc., 2016. 1696− 1704

80 Gwak J Y, Choy C B, Garg A, Chandraker M, Savarese S. Weakly supervised generative adversarial networks for 3D reconstruction. arXiv preprint arXiv: 1705.10904, 2017. 263−272

81 Rosca M, Lakshminarayanan B, Warde-Farley D, Mohamed S. Variational approaches for auto-encoding generative adversarial networks. arXiv preprint arXiv: 1706.04987, 2017.

82 Zhu R, Galoogahi H K, Wang C Y, Lucey S. Rethinking reprojection: closing the loop for pose-aware shape reconstruction from a single image. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 57−65

83 Liu J, Yu F, Funkhouser T. Interactive 3D modeling with a generative adversarial network. In: Proceedings of the 2017 International Conference on 3D Vision (3DV). Qingdao, China: IEEE, 2018. 126−134

84 Wu J J, Zhang C K, Xue T F, Freeman W T, Tenenbaum J B. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In: Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016). New York, USA: Curran Associates, Inc., 2016. 82−90

85 Gadelha M, Maji S, Wang R. 3D shape induction from 2D views of multiple objects. In: Proceedings of the 2017 International Conference on 3D Vision (3DV). Qingdao, China: IEEE, 2017. 402−411

86 Wang P S, Liu Y, Guo Y X, Sun C Y, Tong X. O-CNN: octree-based convolutional neural networks for 3D shape analysis. *ACM Transactions on Graphics (TOG)*, 2017, **36**(4): Article No. 72

87 Sun Y B, Liu Z W, Wang Y, Sarma S E. Im2avatar: Colorful 3D reconstruction from a single image. [Online], available: https://arxiv.org/abs/1804.06375, March 25, 2019

88 Tatarchenko M, Dosovitskiy A, Brox T. Octree generating networks: efficient convolutional architectures for high-resolution 3D outputs. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2107−2115

89 Riegler G, Ulusoys A O, Geiger A. Octnet: learning deep 3D representations at high resolutions. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, Hawaii, USA: IEEE, 2017. 6620−6629

90 Häne C, Tulsiani S, Malik J. Hierarchical surface prediction for 3D object reconstruction. In: Proceedings of the 2017 International Conference on 3D Vision (3DV). Qingdao, China: IEEE, 2017. 76−84

91 Charles R Q, Su H, Mo K, Guibas L J. PointNet: deep learning on point sets for 3D classification and segmentation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, Hawaii, USA: IEEE, 2017. 77−85

92 Qi C R, Yi L, Su H, Guibas L J. Pointnet++: deep hierarchical feature learning on point sets in a metric space. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). New York, USA: Curran Associates, Inc., 2017. 5099−5108

93 Klokov R, Lempitsky V. Escape from cells: deep Kd-networks for the recognition of 3D point cloud models. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 863−872

94 Newell A, Yang K Y, Deng J. Stacked hourglass networks for human pose estimation. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer, 2016. 483−499

95 Lin C H, Kong C, Lucey S. Learning efficient point cloud generation for dense 3D object reconstruction. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, California, USA: AAAI, 2017. 3−11

96 Pontes J K, Kong C, Sridharan S, Lucey S, Eriksson A, Fookes C. Image2mesh: A learning framework for single image 3D reconstruction. [Online], available: https://arxiv.org/abs/1711.10669v1, March 25, 2019

97 Wang N Y, Zhang Y D, Li ZW, Fu Y W, Liu W, Jiang Y G. Pixel2mesh: Generating 3D mesh models from single rgb images. [Online], available: https://arxiv.org/abs/1804.01654v1, March 25, 2019

98 Xiang Y, Mottaghi R, Savarese S. Beyond PASCAL: a benchmark for 3D object detection in the wild. In: Proceedings of the 2014 IEEE Winter Conference on Applications of Computer Vision. Steamboat Springs, CO, USA: IEEE, 2014. 75−82

99 Everingham M, Van Gool L, Williams C K I, Winn J, Zisserman A. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 2010, **88**(2): 303−338

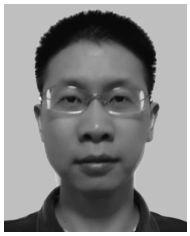
100 Deng J, Dong W, Socher R, Li L J, Li K, Li F F. ImageNet: a large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009. 248−255

101 Chang A X, Funkhouser T, Guibas L, Hanrahan P, Huang Q X, Li Z M, et al. Shapenet: An information-rich 3d model repository. [Online], available: https://arxiv.org/abs/1512.03012v1, March 25, 2019

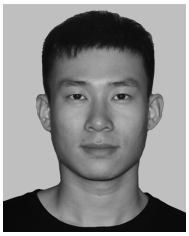
102 Miller G A. WordNet: a lexical database for English. *Communications of the ACM*, 1995, **38**(11): 39−41

103 Song H O, Xiang Y, Jegelka S, Savarese S. Deep metric learning via lifted structured feature embedding. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 4004−4012

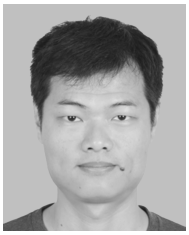
104 Shilane P, Min P, Kazhdan M, Funkhouser T. The princeton shape benchmark. In: Proceedings of the 2004 Shape Modeling Applications. Genova, Italy: IEEE, 2004. 167−178



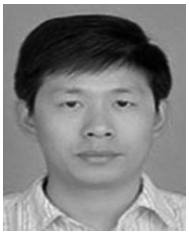
陈 加 华中师范大学教育信息技术学院讲师. 主要研究方向为可视计算, 运动捕捉, 三维重建, 教育信息技术.
E-mail: jc@mail.ccnu.edu.cn
(**CHEN Jia** Lecturer at the School of Educational Information Technology, Central China Normal University. His research interest covers visual computing, motion capture, 3D reconstruction, and educational information technology.)



张玉麒 华中师范大学教育信息技术学院硕士研究生. 主要研究方向为深度学习, 三维重建.
E-mail: ZYQ2046@mail.ccnu.edu.cn
(**ZHANG Yu-Qi** Master student at the School of Educational Information Technology, Central China Normal University. His research interest covers deep learning and 3D reconstruction.)



宋 鹏 瑞士联邦理工学院 (洛桑) 计算机图形学与几何实验室博士后. 主要研究方向为计算机图形学, 三维重建.
E-mail: peng.song@epfl.ch
(**SONG Peng** Postdoctor at the Computer Graphics and Geometry Laboratory, École Polytechnique Fédérale de Lausanne, Switzerland. His research interest covers computer graphics and 3D reconstruction.)



魏艳涛 华中师范大学教育信息技术学院副教授. 主要研究方向为深度学习, 计算机视觉. 本文通信作者.
E-mail: weiyantaoccnu@163.com
(**WEI Yan-Tao** Associate professor at the School of Educational Information Technology, Central China Normal University. His research interest covers deep learning and computer vision. Corresponding author of this paper.)

王 煜 香港科技大学机器人研究院院长, 教授. 主要研究方向为几何建模与设计, 机器人学. E-mail: mywang@ust.hk
(**WANG Yu** Director (professor) at the Robotics Institute, the Hong Kong University of Science and Technology. His research interest covers geometric modeling and design, robotics.)