文章编号: 1007-2853(2020) 01-0058-05

基于深度学习的单幅图像三维重建算法

朱 莉,陈 辉*
(上海电力大学 自动化工程学院,上海 200090)

摘要: 图像三维重建在逆向工程、人工智能等领域广泛应用基于深度学习利用单幅图像重构出三维模型,已经成为当前研究的热点.文章首先综述单幅图像三维重建的研究现状,重点研究基于体素表达的 3D-R2N2、基于点云表达的 PSGN、基于单片网格表达的 Pixel2Mesh 和基于多片三角形网格表达的 AtlasNet 四种算法,通过实验对比研究,来分析解决不同任务与输出模型不同表达方式的选择问题.

关 键 词: 深度学习; 单幅图像; 三维重建
中图分类号: TP 391.4 文献标志码: A DOI: 10.16039/j.cnki.cn22-1249.2020.01.013

三维重建是计算机视觉领域具有挑战性的问题之一,根据真实图像中的数据重建出具有精确几何信息的三维结构模型.目前流行的方法是多目图像重建,需要对于目标物体进行多角度测量,耗费资源,相比之下单幅图像的三维重建输入简单,更适合便捷式三维重建的应用场合,如逆向工程、模式识别、机器人导航[1]及无人驾驶领域中.传统的单幅图像三维重建主要包括基于几何外形重建和模型重建的算法.基于几何外形的方法以简单的方式提取表面信息,但对光照和灰度要求高,基于模型的方法利用先验知识在特定物体上取得较好的重建,但很难应用于所有类别.现研究中基于单幅图像的重建.问题分为物体和场景两种重建,本文中将对单个物体的重建进行探究.随着深度学习在计算机视觉领域的应用,目前数据的表达方式主要有体素、点云和网格三种,在不同的研究任务中,不同三维数据的表达方式各具优势.本文针对这个问题选择四种代表算法进行深入研究,并得出结论.

1 单幅图像三维重建研究现状

单幅图像三维重建的主要思想是从给定的单幅图像中提取目标的二维几何信息,并利用先验知识来推测出被遮挡的部分,重构出完整的三维结构,但重建过程中仍会存在图像自身的属性问题、重建的不适应问题、地面的模糊性以及类间差异和类内差异.

围绕深度学习的发展,基于体素表示的方法最先提出,将 CAD 模型进行体素化,表示为二值或实值的三维张量. 2015 年,Wu 等[2]提出 3D shapenets,利用卷积深度置信网络将 3D 几何外形表示为 3D 体素网格上二值变量的概率分布,通过吉布斯采样预测其表面形状类型,并填补未知区域生成三维体素模型.为克服缺乏纹理、镜面反射和基线等特征匹配问题,2016 年 Choy 等[3]提出 3D-R2N2,以端到端的方式自动学习,重建出单视图或多视图的三维体素模型.针对单幅图像重建的数据集中出现的类内和类间差异问题,Kanazawa 等[4]结合卷积神经网络提出 WarpNet 网络框架,利用细粒度数据集的结构在类别和姿势变化时,预测不同图像之间的对应关系,实现与监督方法相似质量的重构.为更好的利用先验知识,Wu 等[5]提出,MarrNet 模型,在真实图像上进行端到端的训练,顺序估计出 2.5D 草图和 3D 对象形状.为克服体素重建易受信息稀疏的缺点,Fan 等[6]提出 PSGN 生成点云表示的三维模型,利用条件采样器解决不确定性与固有模糊性,有很好的重建效果.体素和点云表示存在计算复杂和缺乏更精细的几何形状等问题,基于网格表示

收稿日期: 2019-10-10
基金项目: 国家自然科学基金资助项目(51705304) .
通讯作者简介: 陈 辉(1982-) ,女,辽宁葫芦岛人,上海电力大学讲师,博士,主要从事机器视觉方面的研究.

的深度学习模型被提出.但图像和网格之间的转换或渲染易产生离散操作阻碍反向传播. Wang 等[7]基于图形的卷积神经网络来表示三维网格提出 Pixel2Mesh,通过逐步变形椭球,利用从输入图像中提取的感知特征生成正确的三角形网格表示的三维形状. Groueix 等[8]提出 AtlasNet,通过将一组正方形映射到 3D 形状的表面来局部地近似目标表面,多个这样的正方形使用非磁盘拓扑对复杂的表面进行建模.

为便于比较三种不同表示方法的优劣性,以及适用于什么研究场合及方向,选择基于体素表达的 3D-R2N2 网络、基于点云表达的 PSGN 网络、基于单片网格表达的 Pixel2Mesh 网络和基于多片网格表达的 AtlasNet 网络这 4 种网络来进行定性定量的分析.

2　几种典型的不同三维表示的点云重建算法

2.1　基于体素表达的 3D-R2N2 网络

3D-R2N2 网络的主要思想是从大量的数据集中学习一个二维图像到对应三维底层模型的映射,利用数据驱动的方法构建的 3D 循环重建神经网络(3D-R2N2) .以多个或单个图像作为输入,输出三维像素块的对应 3D 重建模型. 3D－R2N2 网络都由三部分组成: Encoder(2D-CNN) 、LSTM (3D-LSTM) 以及 Decoder(3D-DCNN) .见图 1.

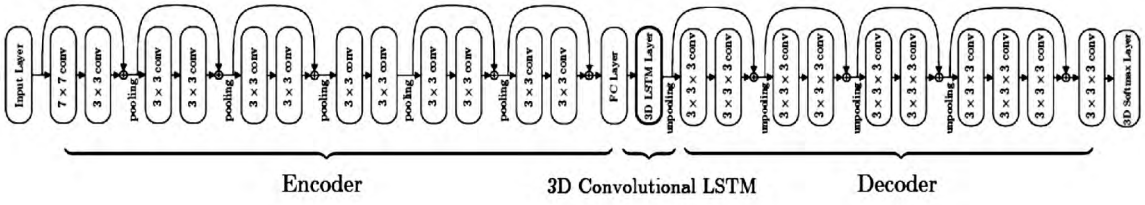

图 1　3D-R2N2 网络架构

Encoder: 主要是将输入的 127* 127 的 RGB 图像变换成低维的特征矩阵,然后送到 LSTM 单元进行循环学习,选择性地更新它们的单元状态或保持状态.其中 CNN 在每两个卷积层之间都会增加一个残余连接,来有效的加速深度神经网络的优化过程.

LSTM: 是 3D-R2N2 的核心部分,由一组具有受限连接的结构化 LSTM 单元组成,每个单元负责重建最终输出的特定部分,与循环神经网络(RNN) 类似,允许网络保留读取的内容,并在读取新图像时更新内存.

Decoder: 采用一个简单的解码器网络,包含 5 个卷积和一个深度残差网络,其中包含 4 个残余连接,最后是最终卷积,卷积层采取 3* 3* 3 大小的卷积核,来处理被赋予三维的结构信息的像素数据,解码 LSTM 隐藏状态,并生成最终的 3D 概率体素重建.

2.2　基于点云表达的 PSGN 网络

PSGN 网络的结构见图 2,主要包括点云生成模块、点云损失函数设计和生成多个预测模型三部分.

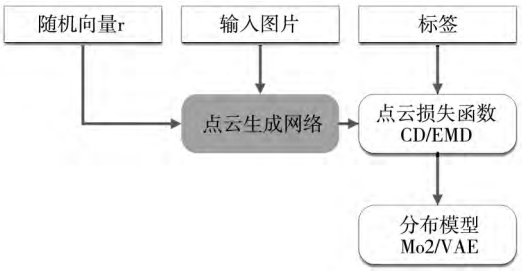

图 2　PSGN 分布式建模模块

点云生成网络: 具有编码器和预测器两个阶段,见图 3.编码器是卷积层和 ReLU 层的组合,输入包括输入图像 I 与随机向量 r,r 用来干扰 I 的预测,目的是将 I 和 r 的输入对映射到嵌入空间.预测器一种方案是由完全连接层组成,独立预测各点的坐标,描述复杂的结构时能表现出良好的性能,但该预测器不能充分利用自然几何统计.第二种网络具有两个预测分支,完全连接(fc) 分支和反卷积(deconv) 分支.改进的网络不仅在复杂结构中具有很高的灵活性,而且能更好的利用几何平滑性,反卷积分支通过权重共享减少参数,同时 Hourglass 卷积网络反复进行编解码,可以更好地联合全局和局部信息.

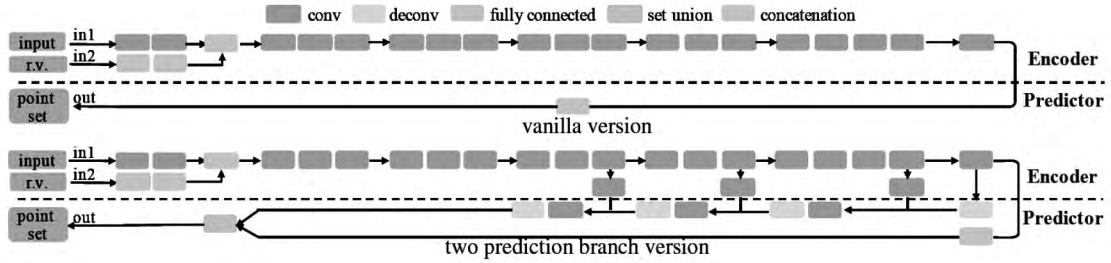

图 3 PSGN 点云的生成网络结构[6]

点云损失函数：选取了两种距离 Chamfer distance（CD）**和** Earth Mover0s distance（EMD）**作为候选，来比较预测的点云和地面真值.**

生成多个预测模型：实验中采用 Mon **损失或者** VAE **方法来进行不确定性的建模，在重建结果上，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.**

2.3 基于网格表达的网络

（1）基于单片网格表达的 Pixel2Mesh **网络**

Pixel2Mesh **网络是基于图形的卷积神经网络，通过逐步变形椭球，采用粗到细的策略，定义各种网格相关损失来捕捉不同层次的特征，从端到端生成三角形网格表示的三维形状的深度学习体系结构.**

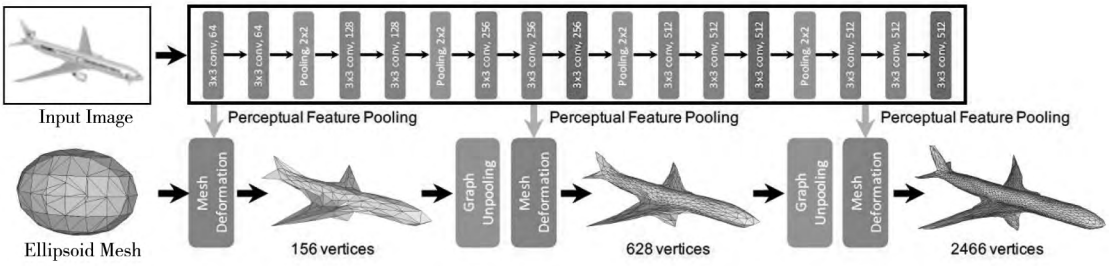

图 4 Pixel2Mesh 网络结构[7]

Pixel2Mesh **整个网络包括一个图像特征网络和一个级联的网格变形网络.图像特征网络是** 2D CNN **结构，见图** 4**，可从输入图像中提取感知特征，网格变形网络可利用该特征将椭圆形网格逐渐变形为所需的** 3D **模型.级联的网格变形网络是一个基于图的卷积网络（** GCN**），其中包含三个变形块.每个变形块都会获取一个表示当前网格模型的输入图，并在顶点上附加** 3D **形状特征，并生成新的顶点位置和特征.而图形解池层增加了顶点数量，以增加处理细节的能力，同时仍保持三角形网格拓扑.**

（2）基于多片网格表达 AtlasNet **网络**

AtlasNet **网络主要思想是基于** 3D **表面生成的，由可学习的参数化组成，受表面正式定义为局部类似于欧几里德平面的拓扑空间的启发，通过将一组正方形映射到** 3D **形状的表面来局部地近似目标表面，多个这样的正方形允许网络使用非磁盘拓扑对复杂的表面进行建模.**

AtlasNet **网络重点在于推理过程，分为两大模块，学习编码目标物体表面以及网格的生成.学习编码目标物体表面主要解决输入** 3D **点云时如何自动编码** 3D **形状以及输入** RGB **图像时如何重建** 3D **形状.对于自动编码器，采用** PointNet **的编码器，将输入点云转换为尺寸为** k = 1024 **的潜矢量，对于图像，采用** ResNet-18 **作为编码器，解码器采用** 4 **个完全连接的层，大小分别为** 1024、512、256、128.**训练中对学习到的参数化以及真值点云定期采样，以避免过度拟合.网格生成过程是将单位正方形上的规则网格转换为** 3D**，以** 3D **方式去连接之前以** 2D **方式连接的点，此方法可以生成高分辨率网格.为了避免网格不闭合或出现空洞或重叠，采用对表面进行密集采样并使用网格重建算法，最终生成闭合高分辨率网格.**

3 实例分析

在三维重建研究中，针对三种不同表示，至今

仍没有一个完全统一的评价指标[8].本实验中采用以下指标对这四种网络进行分析,其中体素表示和点云表示采用交并比 IoU 指标,网格表示采用 F-Score[9] 和 Metro 指标,最后采用倒角距离 CD 和地球移动距离 EMD 对三种三维表示的重建精度进行对比.

3.1　评价指标

(1) 交并比 IoU

测量预测形状体积与真值体积之间的交集与两个体积的并集的比率,因此处理基于表面重建的其他表示时需要对重建的和真值模型先进行体素化.

$$IoU = \frac{V^{\text{pred}} \cap V^{\text{gt}}}{V^{\text{pred}} \cup V^{\text{gt}}} = \frac{\sum_i \{I(V_i^{\text{pred}}) \cdot I(V_i^{\text{gt}})\}}{\sum_i \{I(V_i^{\text{pred}}) + I(V_i^{\text{gt}}) - I(V_i^{\text{pred}}) \cdot I(V_i^{\text{gt}})\}} \quad (1)$$

其中 $I(\cdot)$ 是指标函数, V_i^{pred} 是预测值的第 i 个体素, V_i^{gt} 是真值的第 i 个体素,并且是阈值,因此 IoU 值越高,重建精度越高.

(2) F-Score

从预测结果和真值中统一采样,这些点可以在特定阈值 τ 内找到彼此最近邻居,然后计算 F-score 作为精度.对于 F-Score,越大重建效果越好.

(3) Metro

为了说明网格的连通性,使用公开可用的 Metro 软件,使用 Metro 标准比较了输出网格和地面真实网格的平均欧式距离.

(4) 倒角距离 CD 和地球移动距离 EMD

S 将代表地面真值点云, $\hat{S}$ 代表预测点云.倒角距离 CD 是计算生成点云和真值点云之间平均的最短点距离,地球移动距离 EMD 用来表示真值模型分布与预测模型分布之间的距离.这两个指标都是越小,重建质量越高.

$$d_{CD}(S, S\hat{S}) = \sum_{x \in S} \min_{y \in \hat{S}} \| x - y \|_2^2 + \sum_{y \in S\hat{S}} \min_{x \in S} \| x - y \|_2^2 \quad (2)$$

$$d_{EMD}(S, \hat{S}) = \min_{\phi: S \to \hat{S}} \sum_{x \in S} \| x - \phi(x) \|_2 \quad (3)$$

3.2　实验结果分析

实验中所有环境配置见表 2,并可视化这几种算法在 shapenet 数据集的飞机类别上的运行结果见图 5,各种指标的结果比较见表 3.

表 2　实验环境配置

参数	操作系统	GPU	CUDA	CuDNN	Tensorflow-gpu
数值	Ubuntu 16.04	GTX 1080Ti	8.0	6.0	1.2.0

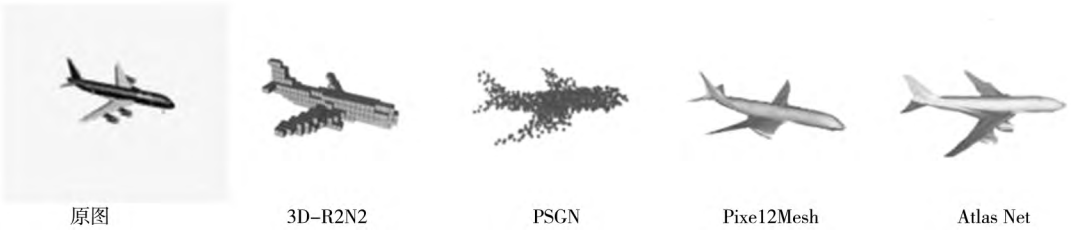

图 5　三种不同表示的网络重建结果

在图 5 中,第一个为原始输入图像,从左至右依次为 3D-R2N2 网络、PSGN 网络、Pixel2Mesh 网络和 AtlasNet 网络生成图像.从图中的视觉外观中,基于体素表示的 3D-R2N2 和基于网格表示的 Pixel2Mesh 和 AtlasNet 更注构造物体表面这个曲面上的点,而基于点云表示的 PSGN 更倾向于生成 3D 形状体积内的点.3D-R2N2 模型的分辨率较低,PSGN 生成的模型精细度不够,错过物体的薄特征比如飞机的机翼上的四个发动机,相比之下,PSGN 的目标函数鼓励保留精细结构.但 PSGN 的模型由于点云之间缺少关联信息,导致表面信息的模糊性,而 Pixel2Mesh 和 AtlasNet 的模型相对前两种表示而言,更注重表面信息,因此生成的模型更加逼真,更接近于原来的真实图像.表 3 中显示了评估标准数据,从表格中可以看出 PSGN 的 CD 和 EMD 值小于 Pixel2Mesh,但是由

于点云具有最大的自由度而这种自由容易导致较小的 CD 和 EMD，但从图中发现 Pixel2Mesh 的重构模型更逼真.

表 3 不同表示的网络重建评估数值

评价指标	3DR2N2	PSGN	Pixel2 Mesh	Atlasnet
Iou	0.542	0.601	–	–
F-Score ($\tau=10^{-4}$)	41.46	68.2	71.12	–
Metro	–	–	–	0.127
CD	0.895	0.430	0.477	0.254
EMD	0.606	0.396	0.579	–

根据实验的可视化结果得出以下结论.

(1) 着重研究物体表面的课题研究，一般选用体素表达与网格表达. 体素和点云这两种表示都失去了重要的表面细节，而且重构曲面模型也是很重要的，相比之下网格，能够建模形状细节.

(2) 针对研究带孔和精细细节的对象时一般选用体素表达. 由于点云的稀疏性与点云之间的不关联性，不适用基于点云表示的模型，网格通常是用渲染管线或可微分的渲染器进行调整以匹配图像的统计信息，基于网格表示的模型难以一致的方式生成，此时更适宜选用体素表达.

(3) 目标物体物体结构比较复杂时优先选择体素表达. 网格通常由三角形网格构成，因此网格能够很好地描述目标对象的细节层次，局限于网络拓扑的可变性，使得网络对目标物体的复杂性比较敏感，当目标物体的结构比较复杂时，网格重构的精确度和效率降低. 基于体素的网络获取三维物体的特征也比较多，因此可优先选择基于体素表达的网络.

(4) 研究方向更倾向于重建物体的变换或应用于其他任务时，可优先选择点云表示.

4 结 论

基于深度学习的单个物体三维重建中，不同三维表示的选择在不同的任务以及与神经网络的结合上会产生不同的效果，也直接影响着网络的选择构建、损失函数的设计以及输出模型的精度，因此三维表示的选择至关重要，应根据研究内容合理选择三维表示方式.

参考文献:

[1] 王影，梁凯，刘麒，等. 基于 Zigbee 网络的智能播种机器人关键技术研究[J]. 吉林化工学院学报(自然科学版)，2019(7)：48-51.

[2] Wu Z，Song S，Khosla A，et al. 3dshapenets：A deep representation for volumetric shapes [C]. Proceedings of the IEEE conference on computer vision and pattern recognition，2015：1912-1920.

[3] Choy C B，Xu D，Gwak J Y，et al.3D-R2N2：A Unified Approach for Single and Multi-view 3D Object Reconstruction [C]. European Conference on Computer Vision，2016：628-644.

[4] Kanazawa A，Jacobs D W，Chandraker M，WarpNet：Weakly Supervised Matching for Single-View Reconstruction [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.2016：3253-3261.

[5] Wu J，Wang Y，Xue T，et al. MarrNet：3D Shape Reconstruction via 2.5D Sketches [C]. Advances in neural information processing systems.2017：540-550.

[6] Fan H，Hao S，Guibas L. A Point Set Generation Network for 3D Object Reconstruction from a Single Image [C]. IEEE Conference on Computer Vision & Pattern Recognition.2017：2463-2471.

[7] Wang N，Zhang Y，Li Z，et al. Pixel2Mesh：Generating 3D Mesh Models from Single RGB Images [C]. European Conference on Computer Vision.2018：55-71.

[8] Groueix T，Fisher M，Kim V G，et al. AtlasNet：A Papier-Mâché Approach to Learning 3D Surface Generation [EB/OL]. (2018-7-20) [2019-10-23]. arXiv preprint arXiv：1802.05384.

[9] Yi L，Shao L，Savva M，et al. Large-scale 3d shape reconstruction and segmentation fromshapenet core55 [EB/OL]. (2017-10-27) [2019-10-23]. arXiv preprint arXiv：1710.06104.

[10] Knapitsch A，Park J，Zhou Q Y，et al. Tanks and temples：benchmarking large-scale scene reconstruction [J]. ACM Transactions on Graphics，2017，36(4)：1-13.

(下转第 67 页)

Double Closed Loop Sliding Mode Control of Ball and Plate System Based on Particle Swarm Optimization

ZHAO Jule,HAN Guangxin*

(College of Information and Control Engineering,Jilin Institute of Chemical Technology,Jilin 132022,China)

Abstract: Ball and plate system is a high order nonlinear system with the character of open loop instability. Based on the Lyapunov stability theory and dual closed-loop control strategy,a sliding mode variable structure position controller is designed,and the particle swarm optimization algorithm is used to optimize the sliding mode parameters with the aim of improving the control precision and reducing the sliding mode chattering elimination.Simulation results show that the sliding mode variable structure control algorithm has better tracking and robust performance compared with traditional dual closed-loop PID control.

Key words: Ball and plate system; Dual closed loop control; Sliding mode variable structure control; Particle swarm optimization

(上接第 62 页)

SingleImage 3D Reconstruction Algorithm Based on Deep Learning

ZHU Li,CHEN Hui*

(College of Automation Engineering,Shanghai University of Electric Power,Shanghai,200090,China)

Abstract: Image 3D reconstruction is widely used in reverse engineering,artificial intelligence and other fields. It is a hot research topic based on deep learning to reconstruct a 3D model using a single image.The article first reviews the research status of 3D reconstruction of single image,focusing on 3D-R2N2 based on voxel expression,PSGN based on point cloud expression,Pixel2Mesh based on monolithic grid expression and AtlasNet based on multi-piece triangle mesh expression. The algorithm analyzes the problem of different expressions of different tasks and output models through experimental comparative research.

Key words: deep learning; single image; 3D reconstruction