计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications [doi: 10.15888/j.cnki.csa.008685]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于深度学习的单视图三维重建①

邹泞键, 冯　刚, 陈卫东

(华南师范大学 计算机学院, 广州 510631)
通信作者: 冯　刚, E-mail: 593645056@qq.com; 陈卫东, E-mail: 599141116@qq.com

摘　要: 单视图三维重建在计算机视觉领域中是一个具有挑战性的问题. 为了提升现有三维重建算法重建后三维模型的精度, 本文除了提取图像全局特征之外还提取图像局部特征, 结合全局特征和局部特征并选取 SDF (signed distance function) 作为重建后的三维物体表达方式, 不仅提高了模型的精度, 生成了更高质量的 3D 形状, 还增强了模型的泛化能力, 使得深度模型可以以较高质量重建出其他物体种类. 实验结果表明, 本文提出的深度网络结构和 3D 形状表示方法与当今最先进的重建算法相比, 无论在重建后三维模型的效果还是新型物体的泛化中都有更好的表现.

关键词: 三维重建; 单视图; 泛化能力; 深度学习; 隐性表面

引用格式: 邹泞键,冯刚,陈卫东.基于深度学习的单视图三维重建.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/8685.html

Single-view 3D Reconstruction Based on Deep Learning

ZOU Ning-Jian, FENG Gang, CHEN Wei-Dong

(School of Computer Science, South China Normal University, Guangzhou 510631, China)

Abstract: Single-view 3D reconstruction is a challenging problem in computer vision. To improve the accuracy of the 3D model reconstructed by the existing 3D reconstruction algorithm, this study extracts both global and local features of the image. On this basis, the signed distance function (SDF) is used to describe the reconstructed 3D objects. In this way, high-quality 3D shapes are generated, and the model has higher accuracy and enhanced generalization capability, which enables the deep model to reconstruct other types of objects with high quality. Experiments demonstrate that compared with the most advanced reconstruction algorithm at present, the proposed deep network and the method for representing 3D shapes have better performance in the effects of reconstructed 3D models and the generalization of new objects.

Key words: 3D reconstruction; single-view; generalization ability; deep learning; implicit surface

1　引言

单视图三维重建一直以来是计算机视觉领域里的一个热点. 随着深度学习的发展和大型 3D 模型数据集 ShapeNet[1] 的出现, 基于深度学习的三维重建成为主流. 编码器-解码器结构是目前基于深度学习三维重建的主要架构, 输入单一图片进编码器提取图像特征而后由解码器将特征向量还原成三维模型, 此结构是训练神经网络学习二维图像和三维物体之间的映射关系. 本文从 MarrNet 方法[2] 中得到启发, 输入单一图片不直接生成特征向量而是生成 2.5D 中间表达 (深度图、表面法向量贴图、轮廓图), 对比之下中间表达可以从图像中获得更多信息, 从而具有从复杂背景的图片中重建出三维模型的能力. 另外之前的研究都只提取了图像的全局特征, 因此重构后的三维模型精准度不高, 针对这个问题本文提出局部特征模块, 结合局部特征和全局特征生成更高质量的三维模型.

① 基金项目: 国家自然科学基金 (61370003)
收稿时间: 2021-12-10; 修改时间: 2022-01-10; 采用时间: 2022-01-28; csa 在线出版时间: 2022-06-16

近年来, 单视图三维重建工作使用的 3D 表达方式主要有点云, 体素, 网格. 基于点云[3–5] 的三维表达虽然易于使用 CNN (convolutional neural networks) 但分辨率十分受限. 基于体素[6–8] 的表达因其使用太多的体素去描述物体内部看不到的部分而增加了计算量. 基于网格[9–11] 的方法重构出来的三维物体较为精细, 不过这种方法受限于其固定的拓扑结构.

针对上述 3 种表达方式的局限性, 隐式曲面表达[12,13] 逐渐受到重视. Mescheder 等人[14] 则使用隐式曲面表达, 他们预测体积网格中每个单元被占用或未占用的概率, 在缓解体积网格分辨率受限的问题的同时也使最后生成的模型更平滑. Chen 等人[15] 同样使用 SDF 来完成三维重建的任务, 虽然生成了连续高质量的三维模型, 但重建细节欠佳, 不能很好的重建如三维模型孔洞处的细节. Wang 等人[16] 提出的 DISN 使用 SDF 作为表示, 因不仅提取了输入图片的特征更提取了局部特征使重建细节方面得到提升, 但其缺点在于只能处理纯净背景的图片, 处理复杂背景时其精度会大大降低. Thai 等人[17] 提出的 SDFNet 讨论重建精度的同时也研究了模型的泛化能力, 训练出的模型能很好的重建出训练集中未出现过的物体种类. Kleinberg 等人[18] 将 SDF 与 GAN (generative adversarial networks) 结合生成精度较好的三维模型.

本文提出的模型创新点有: (1) 加入 2.5D 草图预测模块使模型可以从复杂背景图片中重建三维模型, 结合局部特征提取模块和 SDF 隐式表面表达可以生成更逼真的模型. (2) 讨论模型的泛化能力, 研究模型是否可以重建出训练集中未出现过的物体种类, 实验结果证明加入 2.5D 草图预测模块有助于提升泛化能力. (3) 本文采用隐式表面表达方法中最常用的 SDF (signed distance function) 作为三维模型的表达方式不仅估计了采样点的正负值 (该点在表面外还是表面内) 还估计了其到表面的距离, 可以提取任意值的等值面.

2 网络结构

本文方法概况如下: 先通过采样策略为数据集 ShapeNet 每个三维模型采样点生成采样点点云, 之后结合图片全局特征和局部特征预测每个采样点的 SDF 值, 最后提取 SDF 值为 0 的面作为最终生成三维模型表面.

2.1 总体架构

网络整体架构如图 1 所示, 主要包含两个模块: 相机位姿预测模块和 SDF 预测模块. 首先对输入的图像进行相机参数的预测. 之后将图片输入至 2.5D 草图模块经过编码器解码器生成 2.5D 草图, 再提取 2.5D 草图的特征作为全局特征. 接着使用之前预测到的相机参数将采样点映射到二维图像平面, 进行该采样点的局部特征提取. 最后将点特征分别与全局特征和局部特征结合, 进行降维操作后得到两路 SDF 值, 将这两路得出的结果连接生成最后预测的 SDF 值.

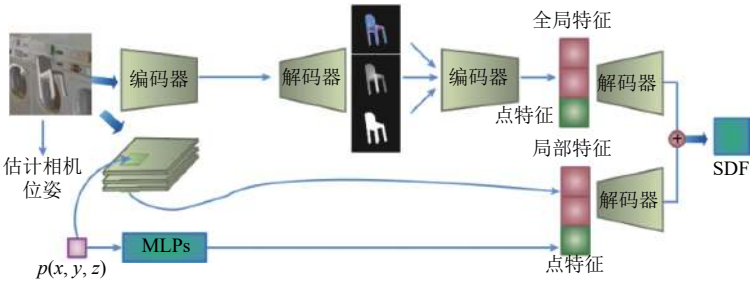

图 1 网络架构

2.2 相机位姿预测模块

使用 ShapeNet 数据集进行这个模块训练, 此模块功能为预测将默认位姿的 3D 模型投影至与输入图片相同位姿所需要的相机参数. 将 ShapeNet 数据集中各个 3D 模型默认位姿作为世界坐标下位姿, 并提取点云生成世界坐标下的点云. 在数据准备操作中会对 ShapeNet 数据集进行渲染, 会对默认位姿 3D 模型旋转、平移特定角度, 而后提取点云生成真实相机坐标下的点云. 相机位姿预测模块大致为输入一张图片至 CNN (这里采用 VGG-16) 预测位姿, 将网络预测出的位姿对世界坐标下的点云进行旋转平移变换操作生成预测相机坐标下的点云, 再与真实相机坐标下的点云损失对比优化

这一模块. 估计出来的相机参数用于局部特征提取. Zhou 等人[19] 的实验说明使用 6D 旋转表示法相对于四元数和欧拉角来说更容易使网络回归, 在此也采用该方法. 本文使用 6D 旋转表示 $b=(b_x, b_y)$ 其中 $b_x\in R^3$, $b_y\in R^3$, $b\in R^6$ 来表示物体位姿. 本模块预测出 6D 旋转表示 b 和位移量 t, 由 b 通过式 (1) 可以计算出旋转矩阵$R=(R_x,R_y,R_z)^{\mathrm{T}}\in R^{3\times3}$.

$$R_x=N(b_x),R_z=N(R_x\times b_y),R_y=R_z\times R_x \tag{1}$$

其中, $N(\cdot)$ 是标准化方程, $\times$是交叉积.

预测相机坐标下的点云和真实相机坐标下的点云之间的差距作为本模块的损失函数如式 (2).

$$L_{\mathrm{cam}}=\frac{\sum\limits_{p_w\in PC_w}\|p_{\mathrm{G}}-(Rp_w+t)\|_2^2}{\sum\limits_{p_w\in PC_w}1} \tag{2}$$

该方程本质是 MSE (mean squared error), 其中 PC_w 表示世界坐标下的点云, 而 p_w 表示世界坐标下点云中的一个点, p_G 为真实相机坐标下的点, 式 (2) 中, p_G–(Rp_w+t) 表示真实相机坐标下的点 p_G 与由世界坐标系下的点 p_w 通过旋转 R 和平移 t 变换得到的点之间的差距. 最后除总共点的个数. 相机位姿预测网络如图 2.

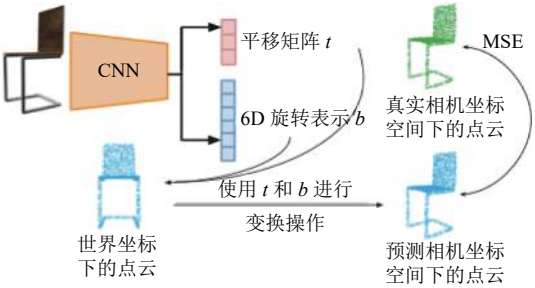

图 2 相机位姿预测网络

2.3 SDF 预测模块

SDF 是一个将 3D 模型采样的点 $p=(x, y, z)$ 映射到实数 $s=\mathrm{SDF}(p)$ 的连续函数, 其中 s 的正负代表在物体表面外部还是内部, s 的绝对值表示这个点到物体表面距离. 再使用 Marching cubes 方法[20] 提取 0 等值面作为 3D 物体的形状. SDF 预测模块分为全局特征提取和局部特征提取两个过程.

2.3.1 全局特征提取

在全局特征提取时使用了 2.5D 草图模块如图 3 所示, 以 2D RGB 图片作为输入预测它的 2.5D 草图: 表面法线、深度和轮廓, 其主要目的是从输入图像中提取出固有的物体属性, 与直接从 RGB 中提取特征相比, 提取出的特征更丰富, 实验表明使用这种中间表示方式可以使网络更容易处理具有复杂背景的图片, 并且因从图像中得到更丰富的信息, 网络的泛化能力也得到了提升. 在 2.5D 草图模块使用编码器为 ResNet18 (deep residual network) 将 256×256 RGB 图像编码成大小为 8×8 的 512 特征图, 然后通过解码器 (解码器包含 4 组 5×5 全卷积层和 ReLU 层然后是 4 组 1×1 全卷积层和 ReLU 层) 输出相应的深度图、表面法线、轮廓图像, 分辨率都为 256×256. 将得到的 2.5D 草图输入到同样是 ResNet18 编码器中得到全局特征. 用 Blender 软件渲染出的 2.5D 草图作为地面真实计算损失函数, 训练过程中我们使用 MSE 损失优化 2.5D 表示模块.

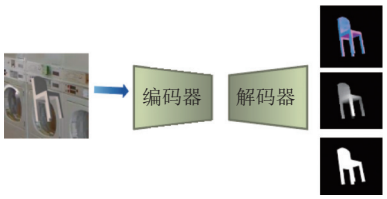

图 3 2.5D 草图预测模块

2.3.2 局部特征提取

如图 4 所示, 在局部特征提取模块使用编码器 VGG-16 (visual geometry group). 通过相机位姿预测模块得到的相机参数将 3D 的点 p 投影到 2D 图像上得到点 q, 在 VGG-16 中 5 个特征子图中找到 q 点对应的位置取下, 维度分别为 64、128、256、512、512, 而后进行连接得到维度为 1 472 的局部特征 (因为特征图尺寸不一样, 所以先采用双线性插值再取下与 q 点对应模块).

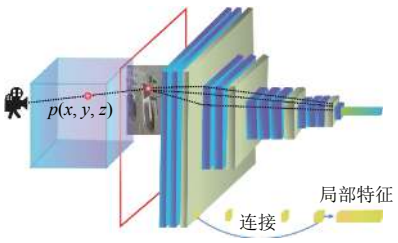

图 4 局部特征提取

2.3.3 预测 SDF 值

如图 5 所示, 当经过全局特征提取和局部特征提取之后得到全局特征维度为 1 024 和局部特征维度为 1 472, 采样点维度初始为 3 经过 MLP (multilayer

perceptron) 生成维度为 512 的点特征后分别与全局特征和局部特征结合. 最后通过图 5 中所示一系列操作分别得到两个维度为 1 的特征, 将得到的两个特征结合则是网络预测的 SDF 值.

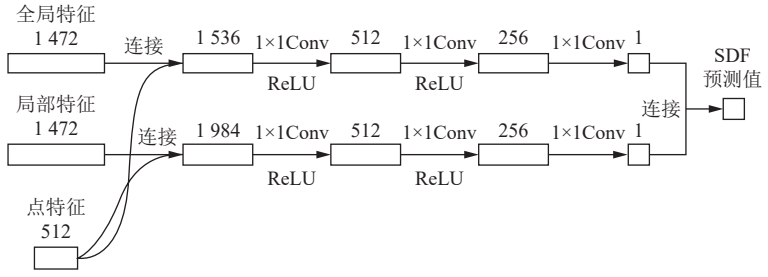

图 5 通过图像特征预测 SDF

2.4 损失函数

用原始数据集 ShapeNet 中的三维模型先提取出点云, 而后测量出点云中各个点的 SDF 值作为地面真实 SDF 值, 而我们的网络 $f(\cdot)$ 是估计使用采样策略采样点的 SDF 值, 使用以下损失函数优化预测 SDF 值模块. SDF 损失函数为式 (3).

$$\begin{cases} LSDF = \sum_{p} m\left|f(I,p) - SDF^{I}(p)\right| \\ m = \begin{cases} m_1, SDF^{I}(p) < \sigma \\ m_2, \text{其他情况下} \end{cases} \end{cases} \tag{3}$$

其中, $f(I, p)$ 表示以图像 I 和三维点 p 作为输入输入到本网络中, 而 $SDF^{I}(p)$ 则表示地面真实 SDF 值, m_1 和 m_2, 这两个值代表不同的权重, σ 为一个阈值.

2.5 表面重建

为了生成隐式平面, 首先定义一个分辨率为 256^3 稠密 3D 网格将采样点点云放入其中并为网格中的每个点预测 SDF 值. 得到了稠密网格中每个点的 SDF 值之后使用 Marching cubes 在 SDF 值为 0 的等值面上生成对应的平面.

3 实验

传统实验都是采用 ShapeNet 数据集中所有 13 种种类按照官方推荐的训练集数据集分割进行实验. 本文为了检测泛化能力, 验证网络是否可以重建出训练集中未出现的物体种类, 采用数据集中最大的 3 种种类 (长凳、汽车、桌子) 作为训练集而其余十种作为训练集. 重建结果与最先进的方法进行对比, 结果显示本文可以重构出训练集中未出现过的种类, 同时还解决了重构中细节恢复的问题.

3.1 数据集和数据准备

使用 ShapeNet 作为数据集, 在实验中我们使用 3 种种类作为训练集, 而使用另外 10 种种类作为测试集. 实验中地面真实深度图、法线贴图等 2.5D 草图我们使用 Cycles 光线追踪引擎在 Blender 中实现了自定义数据生成管道, 以 2D 图片为渲染对象, 生成的 2.5D 草图作为地面真实值与网络生成的 2.5D 草图计算损失. 为了生成实验中的复杂背景 2D 图片, 用不同的 20 个随机图像作背景渲染每个 3D 模型而后生成具有复杂背景的 2D 图片. 为了增加 2D 图像数据量, 本文通过旋转、平移特定角度生成不同角度下的 3D 模型再添加背景图生成 2D 图像. 每个三维模型旋转、平移 8 个特定角度, 因此每个模型生成 8 张具有复杂背景的 2D 图像. 实验中使用的点云也是由数据集 ShapeNet 得到的, 把数据集默认位姿提取得到的点云作为世界坐标下的点云, 而由旋转、平移不同角度后得到的三维模型中提取得到的点云作为真实相机坐标空间下的点云, 其旋转平移的相机参数则作为真实相机参数.

3.2 实现细节

要生成三维模型因此更关注物体表面周边的点, 所以在训练过程中采样策略为在高斯分布 $N(0, 0.1)$ 下采样 2 048 个点. 把采样点云装入生成的 256^3 稠密 3D 网格中结合图片全局特征和局部特征逐个点预测对应的 SDF 值, 再与真实 SDF 值做对比.

本文分开训练相机位姿预测网络和 SDF 预测网络. 相机位姿预测网络中 CNN 使用 VGG-16, 使用式 (2) 作为这个模块的损失函数. 因 SDF 预测网络需要用到相机位姿预测网络中预测出的相机参数, 为更好训练 SDF 预测网络, 我们使用地面真实相机参数训练此网络. 在本模块中提取全局特征线路使用 2.5D 草图作为中间表达, 用 Blender 渲染出来的 2.5D 草图和网络

训练得到草图对比来调优 2.5D 草图提取模块. 最后由局部特征、全局特征和点特征共同预测的 SDF 值与地面真实 SDF 值使用式 (3) 作为损失函数, 其中 m_1=4, m_2=1, σ=0.01. 本文使用 Adam 优化器学习率为 1×10^{-4}, batch-size 为 16.

测试阶段中, 先使用相机位姿预测模块估计相机位姿, 把预测到的位姿用于 SDF 预测模块预测各个点的 SDF 值. 所有采样点 SDF 已知后采用第 2.5 节中提及表面重建的方法生成输出模型.

3.3 评价指标

对于质量评测, 使用常见的衡量指标倒角距离 chamfer distance (CD) 比较重建后的模型和地面真实模型. CD 计算公式为式 (4), 为了使用 CD 评判重建后模型和地面真实模型的相似度需对两个模型采样生成点云. S_1 和 S_2 分别代表这两个点云, 公式中第 1 项代表 S_1 点云中任意一点 x 到 S_2 最小距离之和, 第 2 项则表示 S_2 中任意一点 y 到 S_1 最小距离之和. 所以 CD 值越大说明两组点云区别越大, 反之 CD 值越小两组点云区别越小, 重建效果越好.

$$d_{CD}(S_1,S_2)=\sum_{x\in S_1}\min_{y\in S_2}\|x-y\|_2^2+\sum_{y\in S_2}\min_{x\in S_1}\|x-y\|_2^2 \quad (4)$$

3.4 实验对比结果分析

通过渲染复杂背景图后的图片作为输入, 由 2.5D 草图估计网络得到对应的 2.5D 草图如图 6. 之后再进行 SDF 值预测和局部特征提取等操作得到的三维模型与 GenRe[21] 和 OccNet[12] 重建出来的结果进行对比, GenRe 是研究基于泛化至训练集中未出现的物体种类算法中经典的模型, 而 OccNet 是近期使用隐式曲面作为表达方式算法中有代表性的模型, 我们挑选出在训练集中未出现过的 4 种种类进行对比如图 7, 对比可以很明显的看出本文的方法不仅可以恢复训练集中未见过的物体种类还可以很好的处理模型中细节的恢复, 如椅子和长椅靠背孔洞处的恢复以及飞机机翼和手枪手柄处的孔洞恢复更贴近地面真实模型. 表 1 则为部分未见过种类 CD 值对比, CD 值越低表示重建效果越好, 可以看到本文方法 CD 值在 6 种种类和平均值对比中都是最低的, 表明本文重建效果优于其他两种.

4 总结与展望

本文提出的网络使用 SDF 隐式曲面来表示三维物体, 生成后的模型相较于之前单视图三维重建方法有更为清晰的表面. 引入局部特征提取的模块使得最后生成的模型能够捕获细粒度的细节生成高质量 3D 模型. 我们还进行了泛化能力的实验, 通过 2.5D 模块和选择以视觉为中心的坐标系来测试泛化能力. 定性和定量的实验可以验证出我们的方法在重建模型的质量上和泛化能力方面都优于现有的方法.

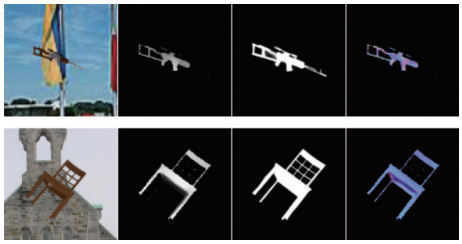

图 6 输入图片和 2.5D

图 7 三维模型重构对比

表 1 6 种训练集未出现的种类 CD 比较

种类	OccNet	GenRe	Ours
沙发	0.15	0.13	0.08
枪	0.15	0.17	0.13
台灯	0.23	0.14	0.09
床	0.30	0.20	0.17
飞机	0.22	0.16	0.11
凳子	0.25	0.15	0.09
平均	0.21	0.16	0.11

参考文献

1 Chang AX, Funkhouser T, Guibas L, *et al*. ShapeNet: An

information-rich 3D model repository. arXiv: 1512.03012, 2015.

2 Wu JJ, Wang YF, Xue TF, *et al*. MarrNet: 3D shape reconstruction via 2.5d sketches. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 540–550.

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

4 Mandikal P, Navaneet KL, Agarwal M, *et al*. 3D-LMNet: Latent embedding matching for accurate and diverse 3D point cloud reconstruction from a single image. British Machine Vision Conference 2018. Newcastle: BMVA Press, 2018.

5 Zhang Y, Liu Z, Liu TP, *et al*. RealPoint3D: An efficient generation network for 3D object reconstruction from a single image. IEEE Access, 2019, 7: 57539–57549. [doi: 10.1109/ACCESS.2019.2914150]

6 Choy CB, Xu DF, Gwak J, *et al*. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 628–644.

7 Girdhar R, Fouhey DF, Rodriguez M, *et al*. Learning a predictable and generative vector representation for objects. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 484–499.

8 Dai A, Qi CR, Nießner M. Shape completion using 3D-encoder-predictor CNNs and shape synthesis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 6545–6554.

9 Pontes JK, Kong C, Sridharan S, *et al*. Image2Mesh: A learning framework for single image 3D reconstruction. 14th Asian Conference on Computer Vision. Perth: Springer, 2019. 365–381.

10 Groueix T, Fisher M, Kim VG, *et al*. A papier-Mache approach to learning 3D surface generation. Proceedings of the IEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018. 216–224.

11 Jack D, Pontes JK, Sridharan S, *et al*. Learning free-form deformations for 3D object reconstruction. 14th Asian Conference on Computer Vision. Perth: Springer, 2019. 317–333.

12 Park JJ, Florence P, Straub J, *et al*. DeepSDF: Learning continuous signed distance functions for shape representation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 165–174.

13 Wang JR, Fang ZY. GSIR: Generalizable 3D shape interpretation and reconstruction. 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 498–514.

14 Mescheder L, Oechsle M, Niemeyer M, *et al*. Occupancy networks: Learning 3D reconstruction in function space. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 4455–4465.

15 Chen ZQ, Zhang H. Learning implicit fields for generative shape modeling. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5932–5941.

16 Wang WY, Xu QG, Ceylan D, *et al*. DISN: Deep implicit surface network for high-quality single-view 3D reconstruction. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 45.

17 Thai A, Stojanov S, Upadhya V, *et al*. 3D reconstruction of novel object shapes from single image. 2021 International Conference on 3D Vision (3DV). London: IEEE, 2021. 85–95.

18 Kleineberg M, Fey M, Weichert F. Adversarial generation of continuous implicit shape representations. 41st Annual Conference of the European Association for Computer Graphics. Norrköping: Eurographics Association, 2020. 41–44.

19 Zhou Y, Barnes C, Lu JW, *et al*. On the continuity of rotation representations in neural networks. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5738–5746.

20 Lorensen WE, Cline HE. Marching cubes: A high resolution 3D surface construction algorithm. ACM SIGGRAPH Computer Graphics, 1987, 21(4): 163–169. [doi: 10.1145/37402.37422]

21 Zhang XM, Zhang ZT, Zhang CK, *et al*. Learning to reconstruct shapes from unseen classes. Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS). Montréal: Curran Associates Inc., 2018. 2263–2274.

(校对责编: 孙君艳)