

融合卷积 CR-FFD 与偏置 Transformer 胶囊网络的 单视图三维物体重建

连远锋^{1,2)}, 赵泽欣¹⁾

¹⁾ (中国石油大学(北京)计算机科学与技术系 北京 102249)

²⁾ (石油数据挖掘北京市重点实验室 北京 102249)

(lianyuanfeng@cup.edu.cn)

摘 要: 针对复杂拓扑结构物体单视图三维重建过程中二维图像与三维形状之间难以准确映射的问题, 提出一种融合卷积 Catmull-Rom 样条自由形变(CR-FFD)与偏置 Transformer 胶囊网络的单视图三维重建方法. 首先通过 Catmull-Rom 样条基函数对点云模型控制点进行插值, 保持点云模型形变局部拓扑结构的一致性; 然后提出卷积神经网络最小二乘求逆解法, 通过非线性参数映射加速求解过程; 最后设计偏置注意力 Transformer 胶囊网络增强局部特征表达能力, 捕获点云形状的细粒度特征. 实验结果表明, 在 ShapeNet 数据集上, 所提方法的 EMD 指标平均值为 3.84, CD 指标平均值为 3.71; 在 Pix3D 数据集上, EMD 指标平均值为 5.51, CD 指标平均值为 5.39; 与已有的单视图点云三维重建方法相比, 该方法有效地提升单视图的三维重建结果, 能够从不同角度保持重建的一致性.

关键词: Catmull-Rom 样条; 自由形变; 三维重建; 胶囊网络; 偏置注意力

中图分类号: TP391.41

DOI: 10.3724/SP.J.1089.2023-00459

Single-View 3D Object Reconstruction Based on CR-FFD and Offset Transformer Capsule Network

Lian Yuanfeng^{1,2)} and Zhao Zexin¹⁾

¹⁾ (Department of Computer Science and Technology, China University of Petroleum, Beijing 102249)

²⁾ (Beijing Key Laboratory of Petroleum Data Mining, Beijing 102249)

Abstract: Aiming at the difficulty of accurate mapping between 2D images and 3D shapes during single-view 3D reconstruction of complex topological objects, a novel single-view 3D reconstruction method combining convolution Catmull-Rom spline free-form deformation (CR-FFD) with offset transformer capsule network is proposed. Firstly, the control points of the 3D point cloud model are interpolated by the basis functions of Catmull-Rom spline to maintain the consistency of the local topological structure during the deformation process. Then, the least square method based on convolutional neural network is proposed to accelerate the calculation process through nonlinear parameter mapping. Finally, an offset attention Transformer capsule network is designed to enhance the local feature expression ability and capture the fine-grained features of point cloud shape. Experiments on ShapeNet dataset show that the average values of EMD and CD are 3.84 and 3.71, respectively. The average values of EMD and CD on Pix3D dataset are 5.51 and 5.39, respectively. Compared with the existing single-view point cloud 3D reconstruction methods, the proposed method can effectively improve the quality of single-view 3D reconstruction and maintain the consistency of reconstruction from different angles.

收稿日期: 2023-07-20; 修回日期: 2024-05-15. 基金项目: 国家自然科学基金(61972353); 中国石油天然气集团有限公司-中国石油大学(北京)战略合作科技专项(ZLZX2020-05). 连远锋(1977—), 男, 博士, 教授, 硕士生导师, 主要研究方向为图像处理、虚拟现实、深度学习; 赵泽欣(2000—), 女, 硕士研究生, 主要研究方向为三维重建.

Key words: Catmull-Rom splines; free-form deformation; 3D reconstruction; capsule network; offset attention

近年来,随着深度学习和图像处理技术的飞速发展,单视图三维模型重建技术得到了广泛关注.由于单视图存在自遮挡问题,无法充分地表达图像的上下文信息,因此基于点云的单视图三维物体重建极具挑战性.传统的面向点云的单视图三维重建算法^[1-3]根据图像表现特征恢复物体空间结构,对图像的质量要求较高,并且存在时间复杂度高、重建结果不精确等问题.与传统方法相比,基于深度学习的方法能够快速提取输入图像的细节特征并推理空间几何结构,实现高效的三维重建. Yang 等^[4]提出 3D-RecGAN 结构,将编码器-解码器结构与生成对抗网络结合进行单视图重建,能够有效地恢复遮挡区域,但对输入图像的质量要求较高; Kurenkov 等^[5]提出 DeformNet 结构,通过自由形变(free-form deformation, FFD)技术预测控制点坐标的偏移量进行单视图重建,但仍存在计算复杂度较高的局限性; van Houtte 等^[6]提出一种基于深度学习的半监督网络,通过基于 B 样条的 FFD 约束形变模型,能够有效地提升三维模型的重建精度:上述方法虽然能够有效地恢复三维模型的拓扑结构,但仍存在形变曲面不通过所有控制点导致重建结果不准确的局限性.与传统的胶囊网络相比,基于 Transformer 的胶囊网络能够更充分地利用输入图像特征提升三维模型的重建精度; Saha 等^[7]基于 Transformer 的双向编码器表示(bidirectional encoder representations from Transformers, BERT)提出 BERT-caps,通过 BERT 和胶囊层生成高质量文本特征,具有较强的鲁棒性,但存在效率较低的问题; Chen 等^[8]提出基于 Shape-Transformer 的胶囊路由算法,将 Transformer 编码的局部信息聚合成包含上下文信息的胶囊向量,能够较好地捕获复杂模型的空间结构信息:上述方法能够较好地表达几何模型的复杂空间特征,但仍存在对图像局部信息利用得不充分导致重建结果不精确的局限性.

针对以上问题,本文提出了一种融合卷积 Catmull-Rom 样条 FFD(Catmull-Rom splines based FFD, CR-FFD)模型与偏置 Transformer 胶囊网络的单视图三维物体重建方法.首先 CR-FFD 模型通过 Catmull-Rom 样条基函数对点云模型控制点进行插值,以保持形变过程的拓扑结构一致性;然后提出基于卷积神经网络(convolutional neural network, CNN)的最小二乘求逆解法加速 CR-FFD 的求解过

程;为了增强网络的局部特征表达能力,将偏置注意力(offset-attention, OA)与基于 Transformer 的胶囊网络结合,提出偏置 Transformer 胶囊网络,通过 OA 捕获点云形状的细粒度特征,以保持二维图像与三维模型的空间结构一致性.

1 相关工作

1.1 FFD

FFD 通过计算控制点坐标的改变量实现三维模型形变,具有广泛的适应性,目前已在虚拟现实、图像处理等领域得到广泛应用. Jung 等^[9]通过 FFD 估计控制点偏差来恢复三维人脸网格,有效地提升三维模型的重建精度,但仍存在因固定拓扑结构而三维模型的重建细节不准确的局限性.针对传统 FFD 方法的固定拓扑结构问题, Zhang 等^[10]引入局部细化的 T 样条定义变形空间,使得复杂模型的局部拓扑结构更加平滑,但存在灵活性差和实时性低等问题; Pontes 等^[11]根据图像类型在模型库中检索相似模型,通过 FFD 和模型的线性组合进行三维模型重建,进一步增强对三维模型的表达能力,但仍存在因输入图像背景复杂、未关注不同模型间的对应关系等因素,而重建结果不准确的局限性.为此, Lamousin 等^[12]提出基于非均匀有理 B 样条的 FFD(NURBS-based FFD, NFFD),将非均匀控制点矩阵与加权控制点相结合,提升模型局部形变性能,但需要大量的控制点来控制形变,导致较高的计算复杂度; Tao 等^[13]将改进的粒子群优化算法与 NFFD 方法结合,通过迭代得到三维模型的最佳控制点坐标,可有效地提升模型形变精度,但该网络的灵活性不够理想.上述方法在部分数据集上取得较好的重建结果,但均需人为地设置控制点,且控制形变的曲面不通过所有控制点,在一定程度上降低模型形变精度.因此,本文提出通过 CR-FFD 对点云模型控制点进行插值,能够有效地保持点云模型形变局部拓扑结构的一致性.

1.2 胶囊网络

胶囊网络采用动态路由算法替代池化,能够在一定程度上减少空间信息的丢失,目前已在目标检测、语义分割等方向得到广泛应用. Zhao 等^[14]将无监督的自动编码器与胶囊网络结合处理稀疏点云,实现三维点云模型重建任务,但该网络的计算复杂度较高.为了降低网络模型的计算量, Chen

等^[15]提出 U 型胶囊网络, 通过基于 k 邻域(k -nearest neighbor, k -NN)的三维卷积动态路由算法提取空间特征, 有效地提高模型的表达能力, 但其在真实复杂场景中的泛化能力欠佳; Chen 等^[16]提出基于注意力的双路卷积胶囊网络, 在对特征间的位置关系进行建模的同时能较好地保留局部信息, 具有较好的泛化能力和鲁棒性, 但对输入数据的质量要求较高; Yu 等^[17]通过无监督的层次胶囊网络自适应地构建人脸层次结构, 能够较好地表达图像与三维物体间的映射关系, 但由于人脸层次结构由一组可视化模板表示, 其重建精度受模板数量限制; Liu 等^[18]基于胶囊注意力的生成对抗网络由粗到细地进行单视图三维重建, 有效地提升三维模型的重建精度, 但对于遮挡区域的重建效果不够理想: 上述方法能够较好地提取空间特征, 但对

于点云模型局部特征的表达能力较差. 因此, 本文提出偏置 Transformer 胶囊网络, 将基于 OA 的 Transformer 编码器嵌入胶囊网络捕获点云形状的细粒度特征, 增强三维重建模型的局部特征表达能力.

2 本文方法

本文方法的结构如图 1 所示, 包括控制点生成网络、卷积 CR-FFD 模型和偏置 Transformer 胶囊网络. 首先将单幅图像输入到控制点生成网络获得三维模型的控制点; 然后利用卷积 CR-FFD 模型对模板模型进行调整; 最后利用偏置 Transformer 胶囊网络对局部点云进行调整, 得到最终的三维点云重建结果.

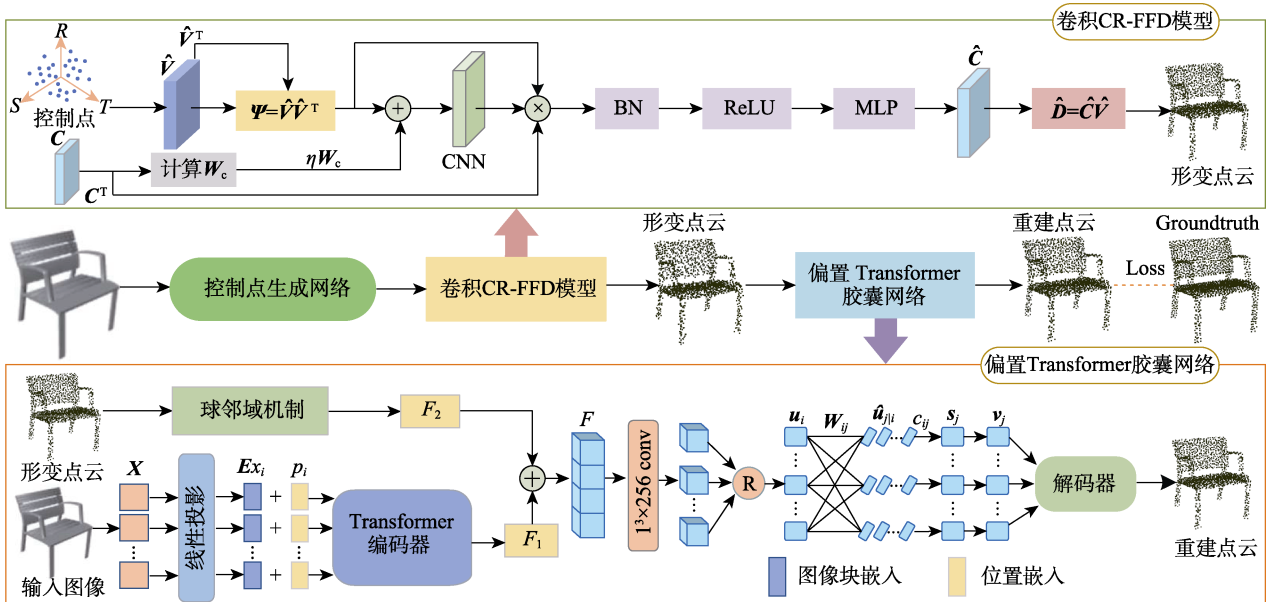


图 1 本文方法结构

2.1 卷积 CR-FFD 模型

受点集生成网络(point set generation network, PSGN)^[19]的启发, 控制点生成网络采用如图 2 所示编码器-预测器结构.

首先从模型数据库选择与图像类别相同的模型点云; 然后将图像及对应的三维模型输入控制点生成网络^[5], 将图像特征 F_I 与三维模型特征 F_M 进行级联得到特征 F_{IM} , 通过 OA 和三维卷积对特征进行增强; 最后利用反卷积和全连接层对级联特征进行解码, 得到控制点的坐标偏移量, 并与初始控制点坐标叠加, 生成最终的控制点坐标.

CR-FFD 模型通过操纵控制点位置使形变体

更接近参照物, 其形变过程如下: 首先将单幅图像输入控制点生成网络, 获得三维模型的控制点坐标; 然后通过 Catmull-Rom 样条基函数对控制点进行插值以约束形变模型; 最后利用基于 CNN 的最小二乘法加速求解过程.

建立 R , S 和 T 方向上的局部坐标系, 初始控制点在各个方向上均匀分布, 设在 3 个方向上分别有 O , P 和 Q 个控制点, 已知坐标系原点为 H , 则初始控制点 V_{ijk} 的坐标满足

$$V_{ijk} = H + \left(\frac{i}{O}\right)R + \left(\frac{j}{P}\right)S + \left(\frac{k}{Q}\right)T.$$

设 $\hat{V}_{ijk}$ 为最终的控制点坐标, 形变后点云模型

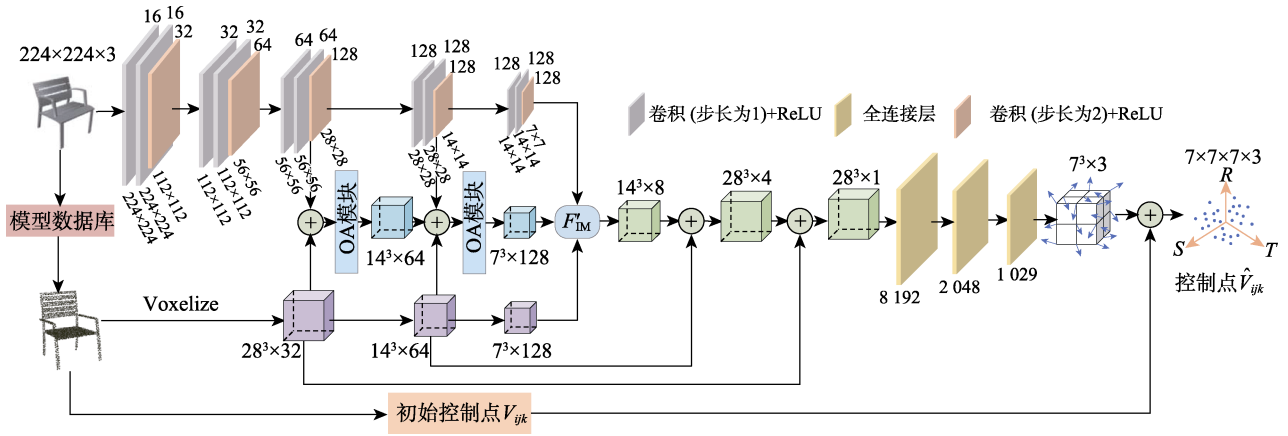


图 2 控制点生成网络结构

的顶点坐标定义为

$$X(r, s, t) = \sum_{i=0}^O \sum_{j=0}^P \sum_{k=0}^Q \hat{V}_{ijk} \Phi_{i,d_1}(r) \Phi_{j,d_2}(s) \Phi_{k,d_3}(t) \quad (1)$$

其中, (r, s, t) 表示模型顶点的初始位置, $\Phi_{i,d_1}(r)$, $\Phi_{j,d_2}(s)$ 和 $\Phi_{k,d_3}(t)$ 表示 Catmull-Rom 样条基函数。以 R 方向为例, 其表达式为

$$\Phi_{i,d_1}(r) = \sum_{i+e=d_1} B_{i,2n+1}(r) L_{h,d_1}(r).$$

其中, $i \in [-n, 1]$, $e \in [0, h]$, $B_{i,2n+1}(r)$ 表示 B 样条基函数, $L_{h,d_1}(r)$ 表示拉格朗日插值基函数。

为了优化形变点云坐标, 将其建模为最小二

乘问题。令 $\mathbf{C} = \sum_{i=0}^O \sum_{j=0}^P \sum_{k=0}^Q \Phi_{i,d_1}(r) \Phi_{j,d_2}(s) \Phi_{k,d_3}(t)$, 将

最终的控制点坐标 $\hat{V}_{ijk}$ 用矩阵 $\hat{\mathbf{V}} \in \mathbb{R}^{n \times 3}$ 表示, 形变后点云模型的顶点坐标 $X(r, s, t)$ 用矩阵 $\mathbf{D} \in \mathbb{R}^{m \times 3}$ 表示。则式(1)可以转换为 $\mathbf{D} = \mathbf{C}\hat{\mathbf{V}}$ 。

为了降低重建点云与真实点云间的误差, 采用最小二乘法对 $\mathbf{D}^T = \hat{\mathbf{V}}^T \mathbf{C}^T$ 进行处理, 并引入参照物的先验信息, 以进一步提高重建点云质量。其目标函数表达式为

$$\mathbf{R} = \|\mathbf{D}^T - \hat{\mathbf{V}}^T \mathbf{C}^T\|_2^2 + \eta \|\mathbf{W}_c \mathbf{C}^T\|_2^2 \quad (2)$$

其中, $\mathbf{W}_c$ 表示对参数 $\mathbf{C}^T$ 的 L_1 正则化约束, 即 $W_{ii} = 1/(\|\mathbf{C}_i^T\| + \varepsilon^2)$, $\varepsilon \rightarrow 0$ 。由式(2)可得, $\mathbf{D}^T = \hat{\mathbf{V}}^T \mathbf{C}^T$ 的最小二乘解为

$$\hat{\mathbf{C}}^T = (\hat{\mathbf{V}} \hat{\mathbf{V}}^T + \eta \mathbf{W}_c)^{-1} \hat{\mathbf{V}} \mathbf{D}^T \quad (3)$$

为了进一步简化计算, 对 $\mathbf{D}^T = \hat{\mathbf{V}}^T \mathbf{C}^T$ 进行转换, 可得 $\hat{\mathbf{V}} \mathbf{D}^T = \Psi \mathbf{C}^T$, 其中, $\Psi = \hat{\mathbf{V}} \hat{\mathbf{V}}^T$ 。将其代入式(3), 可得优化后的系数矩阵为 $\hat{\mathbf{C}}^T = (\Psi + \eta \mathbf{W}_c)^{-1} \Psi \mathbf{C}^T$ 。

为了加快 $\hat{\mathbf{C}}^T$ 的求解速度, 通过 CNN 的参数映射实现求逆运算。则系数矩阵 $\hat{\mathbf{C}}$ 可以表示为

$$\hat{\mathbf{C}} = \text{MLP}(\text{ReLU}(\text{BN}(\hat{\mathbf{C}}^T)))$$

其中, $\text{MLP}(\cdot)$ 表示多层感知器(multilayer perceptron, MLP); $\text{BN}(\cdot)$ 表示批量归一化(batch normalization, BN), $\text{ReLU}(\cdot)$ 表示线性整流函数(rectified linear unit, ReLU)。进一步, 可以通过 $\hat{\mathbf{D}} = \hat{\mathbf{C}} \hat{\mathbf{V}}$ 求得优化后的点云矩阵 $\hat{\mathbf{D}}$ 。

2.2 Catmull-Rom 样条基函数

以 R 方向为例, 设 Catmull-Rom 样条的一般形式为 $\mathbf{F}(r) = \sum_i \mathbf{W}_i(r) \mathbf{P}_i(r)$ 。其中, $\mathbf{W}_i(r)$ 选取 B 样条基函数 $B_{i,m}(r)$; $\mathbf{P}_i(r)$ 表示拉格朗日插值函数。 B 样条基函数的表达式为

$$B_{i,0}(r) = \begin{cases} 1, & r_i \leq r \leq r_{i+1} \\ 0, & \text{otherwise} \end{cases}$$

$$B_{i,m}(r) = \frac{r - r_i}{r_{i+m} - r_i} B_{i,m-1}(r) + \frac{r_{i+m+1} - r}{r_{i+m+1} - r_{i+1}} B_{i+1,m-1}(r).$$

其中, $m = 2n + 1$, n 表示样条曲线满足 C^n 连续。插值于顶点 $\mathbf{V}_i, \mathbf{V}_{i+1}, \dots, \mathbf{V}_{i+h}$ 的拉格朗日插值函数表达式

为 $\mathbf{P}_i(r) = \sum_{e=0}^h L_{h,i+e}(r) \mathbf{V}_{i+e}(r)$ 。其中, $L_{h,i+e}(r) =$

$$\prod_{a=0, a \neq e}^h \frac{r - r_{i+a}}{r_{i+e} - r_{i+a}}, \quad e \in [0, h], \quad h \text{ 表示 } \mathbf{P}_i(r) \text{ 的插值窗口宽度。}$$

$\mathbf{F}(r)$ 的第 q 段曲线 $f_q(r)$ 可以写成

$$f_q(r) = \sum_{i=2-D}^1 \mathbf{W}_{q+i}(r) \mathbf{P}_{q+i}(r), \quad r \in [q, q+1] \quad (4)$$

其中, $D = n + 2$ 。将调配函数 $\mathbf{W}_i(r)$ 和插值函数 $\mathbf{P}_i(r)$ 代入式(4), 得到

$$f_q(r) = \sum_{d_1=-n}^{h+1} \left(\sum_{i=e=d_1} B_{q+i,2n+1}(r) L_{h,q+d_1}(r) \right) \mathbf{V}_{q+d_1}(r) \quad (5)$$

因此, 根据式(4)(5), $2n+1$ 次 Catmull-Rom 样条可以描述为 $F(r) = \sum_{d_1=-n}^{h+1} \Phi_{i,d_1}(r) V_{i,d_1}(r)$. 其中, $\Phi_{i,d_1}(r) = \sum_{i+e=d_1} B_{i,2n+1}(r) L_{h,d_1}(r)$ 表示 Catmull-Rom 样条基函数.

2.3 偏置 Transformer 胶囊网络

如图 1 所示, 特征提取分为 2 路, 首先采用基于 OA 的 Transformer 编码器^[20-21]提取输入图像特征 F_1 ; 然后通过球邻域模型^[22-23]对形变点云进行处理, 获得特征 F_2 ; 级联特征 F_1 与 F_2 获得特征 F . 为了增强特征集成, 采用胶囊层对特征 F 进行处

理, 通过解码器将胶囊层输出的特征向量映射为点云特征, 获得三维重建预测点云.

编码器的注意力层由 OA 和通道注意力组成, 分别负责提取输入图像的局部特征和全局特征. 如图 3 所示, 首先将图像 $X \in \mathbb{R}^{H \times W \times C}$ 分割成 N 个图像块 $X = [x_1, x_2, \dots, x_N] \in \mathbb{R}^{N \times P^2 \times C}$, 其大小为 $P \times P$, 并将图像块线性投影为一维嵌入序列; 为了保留图像空间信息, 为每个图像块编码位置嵌入信息. 将位置嵌入信息与图像块嵌入信息相加, 得到最终的输入序列 $P_{\text{embed}} = [Ex_1 + p_1, Ex_2 + p_2, \dots, Ex_N + p_N]$, 其中, E 表示线性投影矩阵.

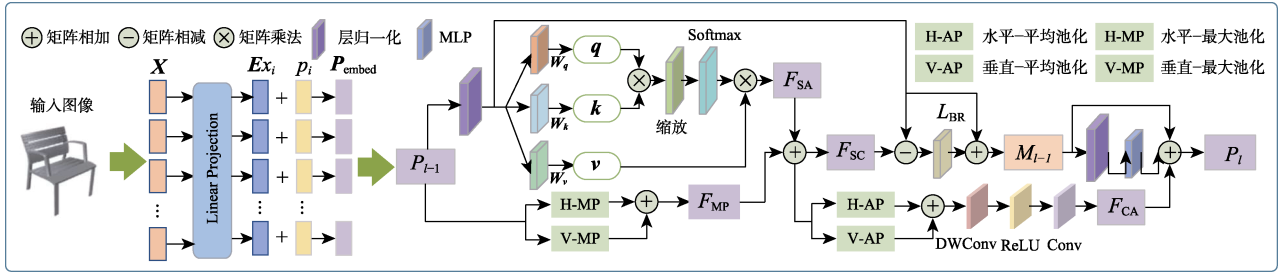


图 3 基于 OA 的 Transformer 编码器模块

为了增强局部特征表达能力, 以捕获点云形状的细粒度特征, 采用 OA 对编码器第 l 层的输入特征 P_{l-1} 进行处理. 将通过最大池化(max pooling, MP)操作后的特征 F_{MP} 与自注意力特征 F_{SA} 相加, 计算其与输入特征 P_{l-1} 的偏移量, 并与输入特征进行残差运算, 得到 P_l 的中间特征表示为

$$M_{l-1} = L_{BR}(\text{LN}(P_{l-1}) - (F_{SA} + F_{MP})) + \text{LN}(P_{l-1}) \quad (6)$$

其中, L_{BR} 表示线性层、BN 层和 ReLU 层, F_{SA} 通过值向量 v 与其对应的注意力权重的乘积获得

$$F_{SA} = \text{Softmax}\left(\frac{q \times k^T}{\sqrt{d}}\right) \cdot v \quad (7)$$

其中, q, k 和 v 由输入序列 P_{l-1} 通过线性投影获得; d 表示向量 k 的维度. 进一步, 通过水平、垂直方向的平均池化(average pooling, AP)操作提取特征间的相互关系, 得到通道注意力特征 F_{CA} . 根据式(6)(7), 将 F_{CA} 与特征 M_{l-1} 拼接, 得到编码器第 l 层的输出特征为 $P_l = \text{MLP}(\text{LN}(M_{l-1})) + F_{CA} + M_{l-1}$.

输入优化后的形变点云 $\hat{D} \in \mathbb{R}^{m \times 3}$, 通过球邻域查询构建局部区域, 并聚合局部区域的特征得到局部采样; 然后通过 3 个大小分别为(512, 1 024, 2 048)的单层感知机(single-layer perceptron, SLP), 在不同球邻域半径内提取样本浅层特征, 得到特

征 F_2 ; 级联编码器的输出特征 F_1 与球邻域模型的输出特征 F_2 获得特征 F ; 通过胶囊网络的动态路由, 从特征 F 中选择有效的特征. 其过程为

$$\begin{cases} s_j = \sum_i c_{ij} W_{ij} u_i \\ v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \cdot \frac{s_j}{\|s_j\|} \end{cases}$$

其中, u_i 表示主胶囊层的输出向量; W_{ij} 表示权重矩阵; c_{ij} 表示 s_j 与 $W_{ij} u_i$ 的相似度. 通过动态路由算法迭代更新, 过程为

$$\begin{cases} b_{ij}^r = b_{ij}^{r-1} + (W_{ij} u_i) \cdot v_j \\ c_{ij} = \frac{\exp(b_{ij}^r)}{\sum_k \exp(b_{ik}^r)} \end{cases}$$

其中, r 表示迭代次数; $\sum_j c_{ij} = 1$; b_{ij}^r 表示 c_{ij} 的对数先验概率, 初值为 0.

解码器将胶囊网络的输出向量映射成点云特征, 获得最终的预测点云坐标. 解码器由 3 个全连接层组成, 其中, 前 2 个采用 ReLU 激活函数, 最后一个采用 Sigmoid 激活函数.

2.4 损失函数

为了更好地评估三维重建的精度, 端到端训

练本文的网络模型, 本文采用倒角距离(chamfer distance, CD)与推土机距离(earth mover's distance, EMD)对重建结果进行评价, 这 2 个指标值越低, 表示重建效果越准确.

2.4.1 CD

CD 用于度量重建点云形状与真实点云形状间的完整性和准确性, 定义为

$$L_{CD} = \sum_{x \in R_1} \min_{y \in R_{GT}} \|x - y\|_2^2 + \sum_{y \in R_{GT}} \min_{x \in R_1} \|y - x\|_2^2.$$

其中, R_1 表示重建点云; R_{GT} 表示 Ground Truth 点云.

2.4.2 EMD

EMD 距离定义为集合 R_1 中的元素 q 到集合 R_{GT} 中所有元素距离的最小和, 定义为

$$L_{EMD}(R_1, R_{GT}) = \min_{\phi: R_1 \rightarrow R_{GT}} \sum_{q \in R_1} \|q - \phi(q)\|_2.$$

其中, ϕ 表示双射关系.

2.4.3 归一化

总损失函数为上述各项损失函数之和, 公式为

$$L_{sum} = \lambda_1 L_{CD} + \lambda_2 L_{EMD}.$$

其中, λ_1 和 λ_2 分别表示 CD 和 EMD 损失函数的系数. 本文中设置 $\lambda_1=0.45$, $\lambda_2=0.55$.

3 实验结果与讨论

3.1 实验设置

本文模型基于 Pytorch 深度学习框架搭建, 其

中, BatchSize 设置为 32; 为了加快收敛速度, 引入 Adam 优化器^[24]. 采用 ShapeNet 数据集^[25]进行训练, 输入图像大小为 224×224 像素, 模型输出顶点数为 2 048 的重建点云. 模型迭代 50 个 epoch 达到收敛.

3.2 数据集

本文方法采用 ShapeNet 数据集进行训练, 包括 Airplane, Bench, Car 等 13 个类别的三维模型. 对 ShapeNet 数据集进行采样, 获得顶点数量为 2 048 的 Ground Truth 点云; 选取不同的视角渲染模型并获得二维图像, 作为网络模型的输入, 并按照 4:1 的比例划分训练集和测试集. 选取 Pix3D 数据集^[26]对本文方法进行测试, 验证真实场景下的三维重建效果.

3.3 定量分析

选取交并比(intersection over union, IoU)对本文方法进行定量评估, 以证明其有效性. IoU 的计算公式为

$$IoU = \frac{|A \cap B|}{|A \cup B|}.$$

其中, A 表示重建模型; B 表示目标模型; IoU 的取值范围为 0~1. 该指标用以评价网络重建模型与目标模型之间的拟合程度, IoU 值越高, 表示重建效果越准确. 将本文方法与 Pixel2Mesh^[27], Pix2Vox^[28], Pix2Vox++^[29], AttSets^[30]和 3D-R2N2^[31]这 5 种方法进行对比, 结果如表 1 所示.

表 1 6 种方法在 ShapeNet 数据集上的 IoU 值对比

类别	Pixel2Mesh ^[27]	Pix2Vox ^[28]	Pix2Vox++ ^[29]	AttSets ^[30]	3D-R2N2 ^[31]			本文
					1view	3views	5views	
Airplane	0.508	0.684	0.674	0.594	0.513	0.549	0.561	0.728
Bench	0.379	0.616	0.608	0.552	0.421	0.502	0.527	0.683
Cabinet	0.732	0.792	0.799	0.783	0.716	0.763	0.772	0.742
Car	0.670	0.854	0.858	0.844	0.798	0.829	0.836	0.679
Chair	0.484	0.567	0.581	0.559	0.466	0.533	0.550	0.681
Lamp	0.399	0.443	0.457	0.445	0.381	0.415	0.421	0.618
Monitor	0.582	0.537	0.548	0.565	0.468	0.545	0.565	0.683
Phone	0.762	0.776	0.809	0.743	0.661	0.732	0.754	0.712
Rifle	0.468	0.615	0.617	0.601	0.544	0.593	0.600	0.695
Sofa	0.622	0.709	0.725	0.703	0.628	0.690	0.706	0.687
Speaker	0.672	0.714	0.721	0.721	0.662	0.708	0.717	0.713
Table	0.536	0.601	0.620	0.590	0.513	0.564	0.580	0.645
Vessel	0.471	0.594	0.603	0.601	0.513	0.596	0.610	0.637
平均 IoU	0.552	0.661	0.670	0.642	0.560	0.617	0.631	0.684

注: 粗体表示最优值.

从表1可以看出, 本文方法在8个类别上取得最佳表现; Pix2Vox++^[29]在5个类别中 IoU 较高, AttSets^[30]在 Speaker 类别取得最好表现; 在平均 IoU 上, 本文方法比 Pix2Vox++^[29]提升 2.09%, 比 AttSets^[30]提升 6.54%. 实验结果表明, 本文方法具有良好的三维重建效果.

进一步, 选取 CD 和 EMD 对本文方法进行定量评估, 将评价指标缩放 100 倍, 并与 PSGN^[19],

3D-LMNet^[32], Pix2point^[33]和 3D-SSRecNet^[34]这 4 种方法进行对比, 结果如表 2 所示. 可以看出, 在 CD 评价指标上, 本文方法在 13 个类别上均取得更高的重建精度, 明显优于其他方法; 在 EMD 评价指标上, 本文方法在大部分数据集上的重建结果均优于其他方法, 3D-SSRecNet^[34]方法在 5 个类别取得最好表现; 但在平均重建精度上, 本文方法比 3D-SSRecNet^[34]提升 2.29%.

表 2 5 种方法在 ShapeNet 数据集上的 CD 和 EMD 值对比

类别	CD					EMD				
	PSGN ^[19]	3D-LMNet ^[32]	Pix2point ^[33]	3D-SSRecNet ^[34]	本文	PSGN ^[19]	3D-LMNet ^[32]	Pix2point ^[33]	3D-SSRecNet ^[34]	本文
Airplane	3.74	3.34	3.29	2.38	2.27	6.38	4.77	3.82	2.72	2.58
Bench	4.63	4.55	4.59	3.51	3.21	5.88	4.99	4.31	3.15	3.32
Cabinet	6.98	6.09	6.07	4.77	4.38	6.04	6.35	4.94	3.91	4.17
Car	5.20	4.55	4.39	3.56	3.25	4.87	4.10	3.61	2.84	2.81
Chair	6.39	6.41	6.48	4.35	3.34	9.63	8.02	6.45	4.17	3.48
Lamp	6.33	7.10	6.58	4.99	4.76	16.17	15.80	8.45	6.19	6.04
Monitor	6.15	6.40	6.39	4.72	4.42	7.59	7.13	5.94	4.49	4.05
Phone	4.56	4.63	4.27	3.52	3.39	5.07	5.43	3.77	3.05	3.29
Rifle	2.91	2.75	2.89	2.45	2.24	8.48	6.08	4.25	3.48	3.61
Sofa	6.98	5.85	5.85	4.44	4.08	7.42	5.65	5.03	3.71	3.67
Speaker	8.75	8.10	8.39	5.94	5.51	8.70	9.15	7.37	5.23	4.92
Table	6.00	6.05	6.26	4.35	3.96	8.40	7.82	6.05	4.16	4.24
Vessel	4.38	4.37	4.55	3.72	3.43	6.18	5.68	4.89	3.96	3.79
平均	5.62	5.40	5.38	4.05	3.71	7.75	7.00	5.30	3.93	3.84

注: 粗体表示最优值.

3.4 定性分析

表 3 所示为在 Sofa 数据集上本文方法的多视角重建结果. 可以看出, 在不同视角下, 重建模型的局部细节均得到较好的保留, 拓扑结构与 Ground Truth 模型十分接近, 展现出较强的鲁棒性. 为了更好地展示本文方法的重建能力, 使用 PSGN^[19]与 Pixel2Mesh^[27]方法在 Lamp 和 Rifle 数据集上分别完成单视图重建, 并与本文方法的重建结果进行对比, 结果如表 4 所示. 选取 3 种基于不同卷积神经网络的 Pixel2Mesh 模型进行单视图三维重建, 验证本文方法的泛化能力. 可以看出, 本文方法在不同数据集的重建精度明显高于其他方法, 能够更准确地表达模型的拓扑结构.

进一步, 在 Airplane 和 Bench 数据集上, 与 PSGN^[19], VGG-Pixel2Mesh^[27]和 DISN^[35]这 3 种方法进行对比, 结果如表 5 所示. 可以看出, 本文方法在 Airplane 数据集上的重建结果与 VGG-Pixel2Mesh^[27]

相近, 优于其他 2 种方法, 均取得较好的重建效果, 但本文方法重建结果的噪声更少, 点云分布更加均匀, 更加贴近真实形状; 在有孔洞的 Bench 模型上, 本文方法的重建结果明显优于其他方法, 证明该方法具有较好的鲁棒性.

3.5 真实场景三维重建

为了验证本文方法在真实场景的重建能力, 在 Pix3D 数据集上, 对本文方法的 CD 和 EMD 评价指标进行定量评估, 并与 PSGN^[19], 3D-LMNet^[32]和 3D-SSRecNet^[34]这 3 种方法进行对比, 结果如表 6 所示. 可以看出, 本文方法在 Chair, Sofa 和 Table 这 3 个类别上均取得更高的重建精度, 优于其他方法, 表明该方法具有较强的鲁棒性. 表 7 所示为部分真实场景图像的三维重建结果. 可以看出, 本文方法对于双层结构的 Table 和 Chair 等复杂结构的重建细节更加准确, 能够有效地保持点云模型形变局部拓扑结构的一致性.

表 3 本文方法多视角三维重建结果

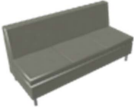
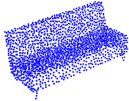
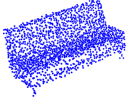
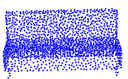
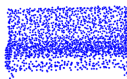
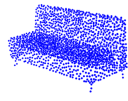
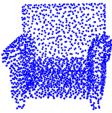
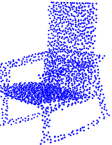
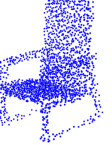
原图	视角 1		视角 2		视角 2	
	Ground Truth	本文	Ground Truth	本文	Ground Truth	本文
						
						
						
						

表 4 3 种方法三维重建结果对比

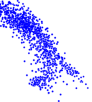
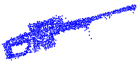
原图	Ground Truth	本文	PSGN ^[19]	VGG-Pixel2Mesh ^[27]	Stereo-Pixel2Mesh ^[27]	ResNet-Pixel2Mesh ^[27]
						
						
						
						
						
						
						
						

表 5 4 种方法三维重建结果对比

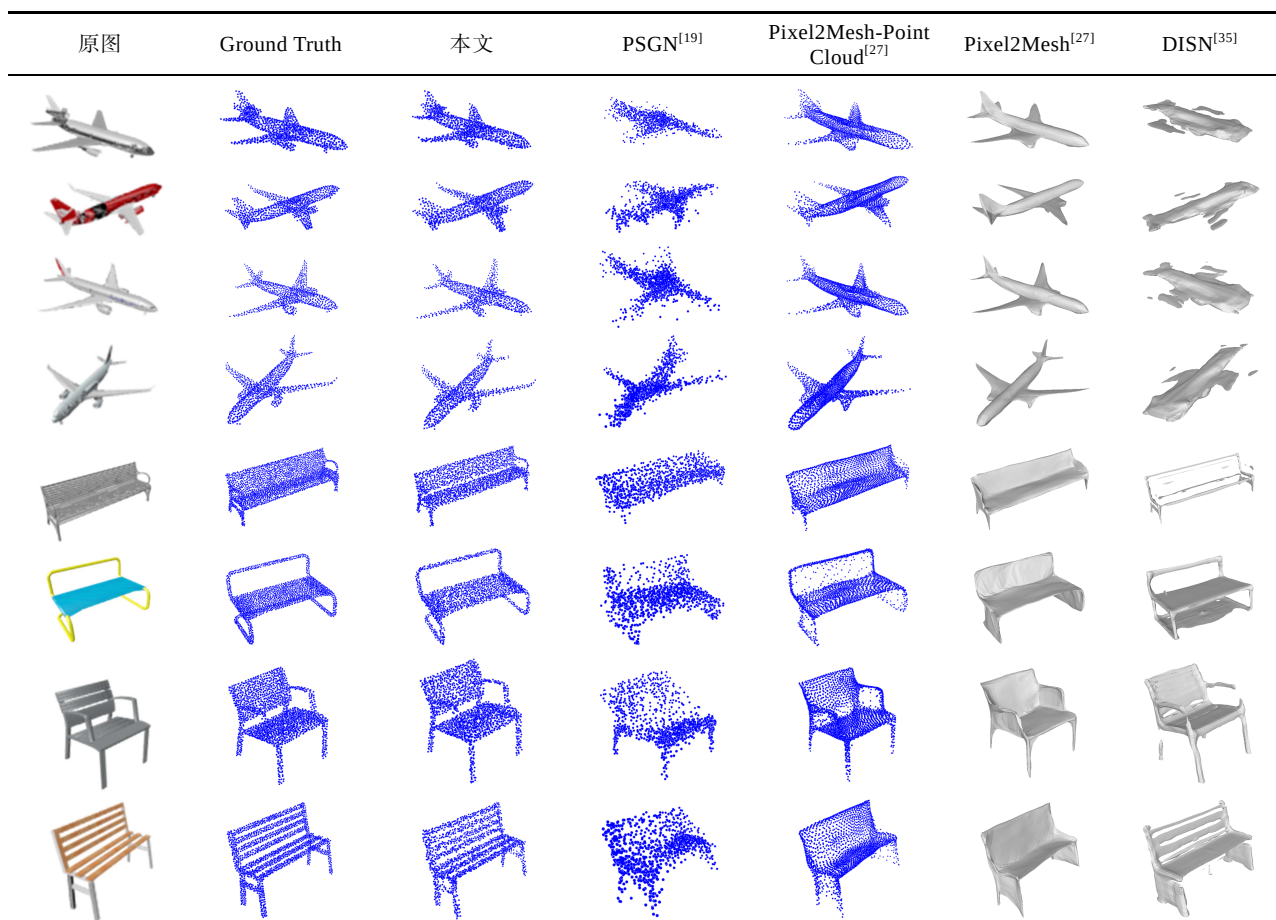


表 6 真实场景重建的 CD 和 EMD 评价指标对比

类别	CD				EMD			
	PSGN ^[19]	3D-LMNet ^[32]	3D-SSRecNet ^[34]	本文	PSGN ^[19]	3D-LMNet ^[32]	3D-SSRecNet ^[34]	本文
Chair	8.05	7.35	5.52	5.26	12.55	9.14	6.05	5.47
Sofa	8.45	8.18	6.04	4.73	9.16	7.22	4.90	4.81
Table	10.82	11.20	6.88	6.19	15.16	12.73	7.59	6.25
平均	9.11	8.91	6.15	5.39	12.29	9.70	6.18	5.51

注: 粗体表示最优值。

3.6 消融实验

为了验证本文方法中各模块设计的合理性, 将 CR-FFD 作为 baseline, 并添加其他模块分别训练模型, 选取 ShapeNet 数据集中的 Airplane, Chair 和 Speaker 模型进行测试, CD 评价指标的结果如表 8 所示。可以看出, 添加 Transformer 编码器模块后, 不同类别的 CD 评价指标均得到一定提升, 平均提升 0.34, 其中, Airplane 类别提升最大, 达到 12.04%; 添加胶囊网络模块后, 各类别重建结果的 CD 评价指标进一步提升, 平均提升 0.13。实验结果表明, 添加各模块后能够有效地提升三维模型的重建精度, 充分验证了各模块设计的合理性。

4 结 语

本文提出一种融合卷积 CR-FFD 模型与偏置 Transformer 胶囊网络的点云三维重建方法。为实现高质量的三维点云重建模型, 通过 Catmull-Rom 样条基函数对点云模型控制点进行插值, 保持点云模型形变局部拓扑结构的一致性, 并使用 CNN 最小二乘求逆解法加速求解过程。在此基础上, 将基于 OA 的 Transformer 编码器模块和球邻域模型嵌入胶囊网络, 增强局部特征表达能力以捕获点云形状的细粒度特征。实验结果表明, 本文方法有效地提升单视图的三维重建结果, 具有较强的鲁棒性。

表 7 本文方法真实场景三维重建效果

原图	Ground Truth	本文	原图	Ground Truth	本文
					
					
					
					
					
					

表 8 消融实验 CD 对比

模块组合方式	Airplane	Chair	Speaker	平均 CD
CR-FFD	2.74	3.86	5.94	4.18
CR-FFD+Transformer 编码器	2.41	3.48	5.64	3.84
CR-FFD+胶囊网络	2.36	3.42	5.58	3.79
CR-FFD+Transformer 编码器+胶囊网络	2.27	3.35	5.51	3.71

注. 粗体表示最优值.

本文方法在单视图三维重建方面取得较好表现, 但仍存在一定局限性, 如在真实场景中, 由于物体间的遮挡问题, 重建模型的局部拓扑结构表达欠佳, 存在进一步优化的空间. 未来将结合隐表达方式描述三维模型拓扑结构, 实现更高质量的重建效果; 同时, 进一步优化网络结构, 实现重建质量与模型性能之间的平衡.

参考文献(References):

[1] Li Tiejun, Xue Luming, Liu Jinyue, *et al.* Research on 3D re-

construction technology of extended depth of field based on auxiliary camera[J]. *Infrared and Laser Engineering*, 2023, 52(4): Article No.20220647(in Chinese)
(李铁军, 薛路明, 刘今越, 等. 基于辅助相机的景深拓展三维重建技术研究[J]. *红外与激光工程*, 2023, 52(4): Article No.20220647)
[2] Liu Yuxuan, Zhang Li, Ai Haibin, *et al.* Progress and prospect of 3D reconstruction based on light field cameras[J]. *Acta Electronica Sinica*, 2022, 50(7): 1774-1792(in Chinese)
(刘玉轩, 张力, 艾海滨, 等. 光场相机三维重建研究进展与展望[J]. *电子学报*, 2022, 50(7): 1774-1792)
[3] Cai Zhenjiao, Zhang Sulan, Li Xiaoming, *et al.* Enhanced motion consistency and guided diffusion feature matching for 3D reconstruction[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2022, 34(2): 273-282(in Chinese)
(蔡振蛟, 张素兰, 李晓明, 等. 面向三维重建的增强运动一致性引导扩散特征匹配[J]. *计算机辅助设计与图形学学报*, 2022, 34(2): 273-282)
[4] Yang B, Rosa S, Markham A, *et al.* Dense 3D object reconstruction from a single depth view[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41(12): 2820-2834

- [5] Kurenkov A, Ji J W, Garg A, *et al.* DeformNet: free-form deformation network for 3D shape reconstruction from a single image[C] //Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2018: 858-866
- [6] van Houtte J, Gao X R, Sijbers J, *et al.* 2D/3D registration with a statistical deformation model prior using deep learning[C] //Proceedings of the IEEE EMBS International Conference on Biomedical and Health Informatics. Los Alamitos: IEEE Computer Society Press, 2021: 1-4
- [7] Saha T, Jayashree S R, Saha S, *et al.* BERT-caps: a transformer-based capsule network for tweet act classification[J]. IEEE Transactions on Computational Social Systems, 2020, 7(5): 1168-1179
- [8] Chen Y, Zhao J Y, Qiu Q L. A transformer-based capsule network for 3D part-whole relationship learning[J]. Entropy, 2022, 24(5): 678
- [9] Jung H, Oh M S, Lee S W. Learning free-form deformation for 3D face reconstruction from in-the-wild images[C] //Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics. Los Alamitos: IEEE Computer Society Press, 2021: 2737-2742
- [10] Zhang Y Z, Zheng J M, Cai Y Y. Proxy-driven free-form deformation by topology-adjustable control lattice[J]. Computers & Graphics, 2020, 89: 167-177
- [11] Pontes J K, Kong C, Sridharan S, *et al.* Image2Mesh: a learning framework for single image 3D reconstruction[C] //Proceedings of the 14th Asian Conference on Computer Vision. Heidelberg: Springer, 2019: 365-381
- [12] Lamoussin H J, Waggenspack N N. NURBS-based free-form deformations[J]. IEEE Computer Graphics and Applications, 1994, 14(6): 59-65
- [13] Tao J, Sun G, Si J Z, *et al.* A robust design for a winglet based on NURBS-FFD method and PSO algorithm[J]. Aerospace Science and Technology, 2017, 70: 568-577
- [14] Zhao Y H, Birdal T, Deng H W, *et al.* 3D point capsule networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 1009-1018
- [15] Chen Y S, Guo B, Shen Y, *et al.* Capsule boundary network with 3D convolutional dynamic routing for temporal action detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(5): 2962-2975
- [16] Chen B, Liu T T, He C, *et al.* Fault diagnosis for limited annotation signals and strong noise based on interpretable attention mechanism[J]. IEEE Sensors Journal, 2022, 22(12): 11865-11880
- [17] Yu C, Zhu X Y, Zhang X M, *et al.* HP-capsule: unsupervised face part discovery by hierarchical parsing capsule network[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 4022-4031
- [18] Liu C X, Kong D H, Wang S F, *et al.* A spatial relationship preserving adversarial network for 3D reconstruction from a single depth view[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2022, 18(4): Article No.110
- [19] Fan H Q, Su H, Guibas L. A point set generation network for 3D object reconstruction from a single image[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 2463-2471
- [20] Zheng S X, Lu J C, Zhao H S, *et al.* Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 6877-6886
- [21] Strudel R, Garcia R, Laptev I, *et al.* Segmenter: Transformer for semantic segmentation[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 7242-7252
- [22] Wu Mingrui, Zhang Bo. A neural network algorithm for large scale pattern recognition problems[J]. Journal of Software, 2001, 12(6): 851-855(in Chinese)
(吴鸣锐, 张钊. 一种用于大规模模式识别问题的神经网络算法[J]. 软件学报, 2001, 12(6): 851-855)
- [23] Zhang Ling, Zhang Bo. A geometrical representation of M-P neural model and its applications[J]. Journal of Software, 1998, 9(5): 334-338(in Chinese)
(张铃, 张钊. M-P 神经元模型的几何意义及其应用[J]. 软件学报, 1998, 9(5): 334-338)
- [24] Kingma D P, Ba J. Adam: a method for stochastic optimization[OL]. [2023-07-20]. <https://arxiv.org/abs/1412.6980>
- [25] Chang A X, Funkhouser T, Guibas L, *et al.* ShapeNet: an information-rich 3D model repository[OL]. [2023-07-20]. <https://arxiv.org/abs/1512.03012>
- [26] Sun X Y, Wu J J, Zhang X M, *et al.* Pix3D: dataset and methods for single-image 3D shape modeling[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 2974-2983
- [27] Wang N Y, Zhang Y D, Li Z W, *et al.* Pixel2Mesh: generating 3D mesh models from single RGB images[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 52-67
- [28] Xie H Z, Yao H X, Sun X S, *et al.* Pix2Vox: context-aware 3D reconstruction from single and multi-view images[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 2690-2698
- [29] Xie H Z, Yao H X, Zhang S P, *et al.* Pix2Vox++: multi-scale context-aware 3D object reconstruction from single and multiple images[J]. International Journal of Computer Vision, 2020, 128(12): 2919-2935
- [30] Yang B, Wang S, Markham A, *et al.* Robust attentional aggregation of deep feature sets for multi-view 3D reconstruction[J].

- International Journal of Computer Vision, 2020, 128(1): 53-73
- [31] Choy C B, Xu D, Gwak J Y, *et al.* 3D-R2N2: a unified approach for single and multi-view 3D object reconstruction[C] //Proceedings of the 14th European Conference on Computer Vision. Heidelberg: Springer, 2016: 628-644
- [32] Mandikal P, Navaneet K L, Agarwal M, *et al.* 3D-LMNet: latent embedding matching for accurate and diverse 3D point cloud reconstruction from a single image[OL]. [2023-07-20]. <https://arxiv.org/abs/1807.07796>
- [33] Afifi A J, Magnusson J, Soomro T A, *et al.* Pixel2point: 3D object reconstruction from a single image using CNN and initial sphere[J]. IEEE Access, 2021, 9: 110-121
- [34] Li B, Zhu S A, Lu Y. A single stage and single view 3D point cloud reconstruction network based on DetNet[J]. Sensors, 2022, 22(21): 8235
- [35] Wang W Y, Xu Q G, Ceylan D, *et al.* DISN: deep implicit surface network for high-quality single-view 3D reconstruction[C] //Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2019: Article No.45