李梦程, 王征, 常锦才. 无相机参数下的多视图三维点云生成方法研究[J]. 智能计算机与应用,2026,16(6):41-49. DOI:10.20169/j.issn.2095-2163.24101102

无相机参数下的多视图三维点云生成方法研究

李梦程[1,2], 王 征[1,2], 常锦才[1,2]
(1 华北理工大学 理学院, 河北 唐山 063210; 2 河北省数据科学与应用重点实验室, 河北 唐山 063210)

摘 要: 随着计算机技术的发展,二维视觉逐渐无法满足复杂对象测量、识别等要求,从二维图像生成逼真的三维内容变得至关重要。文章针对无相机参数下的多视角图像提出一种端到端的点云生成模型,该模型使用多视图池化和特征融合方法改进了 MVCNN 三维分类模型,从多视图图像中提取形状特征,并对特征进行融合形成初步点云序列,采用基于自注意力机制的 Transformer 结构对初步点云序列进行迭代优化,最终形成目标点云模型。使用 ModelNet 数据集进行预训练,在 ShapeNet-Core 数据集上进行微调并测试其 IoU 值,与 3D-R2N2、Pix2Vox 和 3D-AttriFlow 图像三维内容生成模型相比,表现出较好的点云相似性和点云生成效率,对于无相机参数下的多视图点云生成任务提供了新的解决方案和研究思路。

关键词: 无相机参数; 多视图; 点云生成; MVCNN; Transformer
中图分类号: TP391.41 **文献标志码**: A **文章编号**: 2095-2163(2026)06-0041-09

Research on multi-view 3D point cloud generation method without camera parameters

LI Mengcheng[1,2], WANG Zheng[1,2], CHANG Jincai[1,2]
(1 College of Science, North China University of Science and Technology, Tangshan 063210, Hebei, China;
2 Hebei Key Laboratory of Data Science and Applications, Tangshan 063210, Hebei, China)

Abstract: With the development of computer technology, two-dimensional vision is gradually unable to meet the requirements of complex object measurement and recognition, the generation of realistic three-dimensional content from two-dimensional images has become crucial. The paper proposes an end-to-end point cloud generation model for calibration-free multi-view images. The model improves the MVCNN 3D classification model by using multi-view pooling and feature fusion methods to extract shape features from multi-view images. The features are fused to form a preliminary point cloud sequence, and a Transformer structure based on self-attention mechanism is used to iteratively optimize the preliminary point cloud sequence, ultimately forming the target point cloud model. The model is pretrained on the ModelNet dataset, fine tuned and tested for its IoU value on the ShapeNet-Core dataset. Compared with 3D-R2N2, Pix2Vox, and 3D-AttriFlow image generation 3D models, the model exhibits better point cloud similarity and generation efficiency, providing a new solution and research approach for calibration-free multi-view point cloud generation tasks.

Key words: calibration free; multi-view; point cloud generation; MVCNN; Transformer

0 引 言

随着计算机技术的发展,二维视觉逐渐无法满足对复杂对象测量、识别等要求,而二维图像为三维真实世界的表达,蕴含三维形状信息。如何利用二维图像数据生成逼真的三维内容,已成为计算机视觉领域的关键问题。

图像生成三维内容即三维重建,深度学习利用数据训练建立起对先验知识的学习能力,将三维重建转变为图像编码与三维解码问题,实现对事物的三维重建。基于深度学习的三维重建模型主要包括基于体素表达的八叉树结构重建模型[1],基于网格表达的隐式函数重建模型[2]和基于点云表达的扩散三维生成模型[3]。上述模型虽有诸多优秀表现,但多视图图像的形状融合和特征匹配依赖于相机参数,单视图输入缺乏形状信息不能应对三维模型的

基金项目: 国家自然科学基金(52371026)。
作者简介: 李梦程(2000—),男,硕士研究生。主要研究方向:三维重建; 王 征(2001—),男,硕士研究生。主要研究方向:三维重建。
通信作者: 常锦才(1973—),男,博士,教授,硕士生导师。主要研究方向:数值逼近与计算几何等。Email:jincai@ncst.edu.cn。
收稿日期: 2024-10-11

复杂性,从而导致三维内容生成模型结构复杂、模型映射效果欠佳、生成性能较差等问题。

针对上述问题,本文提出了一种端到端的无相机参数下多视图三维点云生成方法。该方法对于输入的多视图图像分别进行特征提取,对视图特征进行视图最大池化,融合池化特征并转换为点云格式,通过多层自注意力编码器和解码器生成目标点云。本文旨在解决缺乏相机参数的多视图图像重建问题,利用神经网络融合多视图特征,生成高质量的点云表达。本文研究工作如下:

(1)改进了 MVCNN 三维分类模型。使用 AlexNet 预训练模型作为多视图特征提取网络;利用最大视图池化对特征图进行降维与特征增强,将特征融合转化为点云形式,以此形成三维模型的整体视觉特征。

(2)构建了点云 Transformer 模型。将点云类比为文本序列的表达,实现点云的自注意力机制,实现点云特征编码器和解码器,形成点云与图像的形状映射。

(3)模型支持多张图像输入,无需复杂的图像预处理或点云预处理,多视图融合时不需要预估视图的相机参数,实现端到端的三维模型生成。

本文工作致力于提高点云生成的准确性和鲁棒性,为图像三维重建提供可靠的解决方案。在 ModelNet 数据集和 ShapeNet 数据集上的实验结果验证了整体算法的先进性和普适性。

1 相关工作

1.1 多视图的三维重建

多视图三维重建方法通过结合多张图像的信息来重建三维模型,主要采用图像特征融合与形状特征融合两种方式。网络结构主要采用编码器-解码器架构来提升三维重建的性能。

1.1.1 图像特征融合

基于图像特征融合的方法有:Choy 等学者[4]将图像依次输入到 3D-LSTM 中,控制输入门和遗忘门来选择性更新特征表示,使用 3D-CNN 将特征解码为体素模型。Xie 等学者[5]构建以 VGG 网络为基础的编码器对不同视图分别进行特征提取,使用 3D-CNN 解码生成粗糙 3D 模型,通过上下文注意力模块进行特征融合,自适应地选择不同粗三维体中的部件进行重建。Lin 等学者[6]使用二维 Transformer 作为编码器、三维 Transformer 作为解码器,同时进行特征提取和视图融合,提高了输入图像较多情况下的体素三维重建方法的性能。Yuan 等学者[7]对单视图或多视图进行重建,使用 CNN 提取主视图特征及辅助图像特征,根据物体部分可见性调整特征权重,将具有权重的辅助视图特征添加到主视图特征中,融合多视图特征从而生成更丰富的三维网格模型。王佳欣等学者[8]在复杂光场环境下对图像特征配准融合进行三维重建。在输入图像较多时,模型同时使用 3D-CNN 和注意力机制会消耗较大的计算资源。

1.1.2 图像形状特征融合

基于图像形状特征融合的方法有:Yao 等学者[9]提出一种端到端的深度学习架构 MVSNet,通过可微单应性变换将多视角图像转换到相机截锥体上构建三维代价体,经过 3D-CNN 生成深度图。Spezialetti 等学者[10]对多视图图像进行相对姿态估计,根据姿势和轮廓构建体素占用网格,对体素占用网格逐渐细化得到体素模型。Murez 等学者[11]使用相机参数对图像特征投影并累积成体素网格,使用 3D-CNN 对三维特征细化并预测回归截断符号距离函数 TSDF。相较于特征融合,图像形状特征融合方式可以取得更好的模型表现,但是处理多张图像进行多视图立体匹配时需要使用相机参数。目前,Wang 等学者[12]采用点阵图的三维表达形式解决相机参数约束问题,进行多视图三维立体重建,采用 ViT 编码器提取图像特征,通过自注意力机制关联视图内部特征,交叉注意力机制关联视图间特征,最终通过全连接网络生成点阵图。Wang 等学者[13]提出一种处理无姿态图像的三维重建模型,输入无位姿的稀疏图像,使用 DINO ViT 进行标记化处理,结合初始化的三平面 NeRF 与图像特征连接输入 Transformer,最终得到三平面 NeRF 表达和粗略点云模型,为后续模型的体积渲染和姿态预测提供了基础。

1.2 基于点云表达的三维重建

点云是利用三维坐标、颜色等信息表示物体表面的点的集合,点云为三维重建网络提供了更好的表示形式。基于点云形式重建得到的三维形状更加平滑,相较于网格和体素模型运算所占用的内存更少,模型生成效率更快。

Fan 等学者[14]在图像编码后使用全连接和反卷积作为解码器,使用倒角距离(CD)和搬土距离(EMD)作为损失函数指标,重建点云形式的三维模型。Jiang 等学者[15]联合生成对抗损失和多视图一致损失,使用 GAN 网络重建点云模型。Nguyen 等

学者[16]提出一种单张图像重建点云模型，将二维图像中提取的物体形状特征与噪声点云特征混合，输入图神经网络生成目标点云。Wen 等学者[17]则结合隐式特征与显式语义属性，通过属性流处理输入图像和初始球点云，提取几何代码和语言特征，利用变形流将点云变形为目标形状。Wang 等学者[18]提出一种基于扩散模型和三平面表示的三维虚拟人像生成方法，提取多视角人像的三平面特征，训练多个扩散模型以学习低分辨率三平面特征到高分辨率的映射，自动生成神经辐射场表示的三维数字人像。Nichol 等学者[3]提出 Point-E 模型，输入文本信息生成多视图图像，经过点云扩散模型生成点云。点云生成模型往往以单视图输入为基础，而单视图输入缺乏形状信息不能应对三维模型的复杂性，重建的点云模型结构未臻完善，分辨率较低。

基于上述问题和挑战，本文以多视图图像为输入，提取多视图的形状特征，设计基于自注意力机制的编码器和解码器网络，实现图像到点云的快速生成。

2　模型设计

本文提出 MVCAT（Multi-View Convolutional neural networks And Transformer）模型，针对无相机参数图像进行点云生成。模型采用改进的多视图特征提取网络提取形状特征并进行特征融合，将特征视为点云序列问题，序列长度为点云集合的点的数量，特征为点云的坐标特征，采用自定义点云 Transformer 模型进行特征编码和解码生成目标点云。本文训练一个神经网络 G 作为多视图点云 S 的生成器：

$$S = G(V;\Theta) \tag{1}$$

其中，集合 $V = \{v_i \mid i = 1,2,3,\cdots\}$ 表示多视图集合，Θ 表示网络参数。MVCAT 模型流程结构如图 1 所示。

2.1　多视图特征提取模块

为了更好地提取多视图图像形状特征，克服无相机参数的多视图特征融合问题，本文改进了 MVCNN 网络结构来提取多视图形状特征，模型最初由 Su 等学者[19]提出，以 3D 模型投影图像作为输入，提取多视图特征用于 3D 模型分类，不仅提高了计算效率，而且在分类性能上优于直接使用 3D 模型的方法。

多视图特征提取模块主要目的是快速处理并融合多视图图像特征，转化特征格式为点云格式。因此采用 AlexNet 预训练模型，通过少量的卷积、激活、池化操作提取二维图像的初始形状特征，与 ResNet、DINO 相比，AlexNet 结构更为简单，具有较少的参数量和计算复杂度，在处理多张图像输入时表现出更高的计算效率。AlexNet 网络包含 5 层卷积层，使用 ReLU 激活函数，并在第 2、3、5 层卷积后应用最大池化降低特征参数，数据变化过程如图 2 所示。

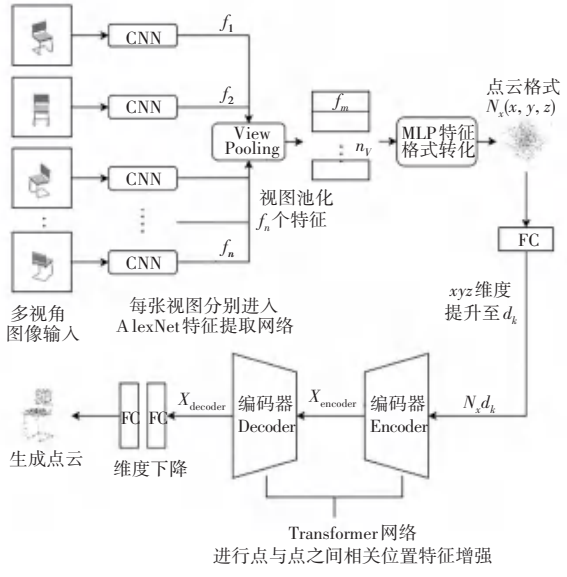


图 1　网络流程结构图

Fig. 1　Diagram of network process structure

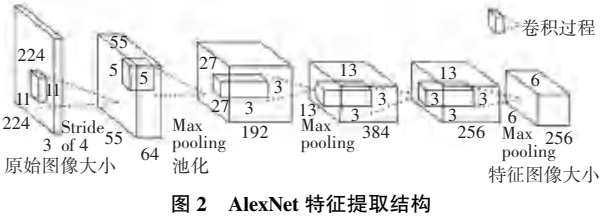


图 2　AlexNet 特征提取结构

Fig. 2　AlexNet feature extraction structure

模型以多视图图像 $V \in \mathbb{R}^{n\times c\times h\times w}$ 作为输入，为了得到统一的多视角物体形状特征，将每个视角 v_i 分别经过相同 CNN 网络得到大小为 $c_o \times h_o \times w_o$ 的特征 f_i，得到所有视图特征 $f_i(i = 1,2,\cdots)$，CNN 网络计算公式如下：

$$f_i = \sum_{h=1}^{K}\sum_{w=1}^{K} v_i(p(x,h,w)) \cdot \omega(h,w) \tag{2}$$

其中，x 表示图像像素位置；(h,w) 表示位置 x 的邻域位置；$\omega(h,w)$ 表示对应位置的权重函数；K 表示卷积核大小；$p(x,h,w)$ 表示采样函数，定义为：

$$p(x,h,w) = x + p(h,w) \tag{3}$$

视图特征的简单拼接或迭代输入[4-7]需要较大的模型参数，影响模型性能，MVCAT 模型主要学习

如何聚合多视图特征,将所有视图的信息聚合成为单一、紧凑的 3D 形状描述特征,降低模型参数。根据池化层在 CNN 网络中的计算方式和作用类比到视图特征之间,对视图特征之间使用视图最大池化操作,保留特征较为明显的特征视图。对于大小为 $c_o \times h_o \times w_o$ 的视图特征 f_i 展平成大小为 d 的一维特征向量 f_a, 对于视图特征 f_a, f_b 保留每个特征位置的前 n_v 个最大值, d 列的最大值向量构成新视图特征 f_m ,最终 n 个视图特征仅保留前 n_v 个视图特征。研究推得的公式为:

$$f_m[i] = \max(f_a[i], f_b[i], \cdots), i = 1,2,\cdots,d \quad (4)$$

其中, $f_m(m = 1,2,3,\cdots,n_v)$ 表示池化后的特征, max 表示视图特征最大值函数。最大池化在卷积中对通道图的局部区域进行特征增强并缩小特征大小,而视图池化则在不同通道的对应位置进行特征增强,保持特征大小不变。由于视图的不同通道使用不同的卷积核,同一位置的特征值存在差异。视图池化不仅增强了局部特征,还提升了全局形状特征。视图池化方式如图 3 所示。

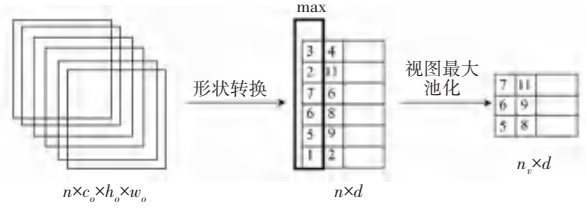


图 3 视图池化方式

Fig. 3 View pooling method

将视图池化得到的 n_v 个增强特征直接进行特征向量拼接得到视图融合特征向量,将视图融合特征向量经过线性变化得到点云初始格式 $N \times (x,y,z)$, 最后经过 MLP 多层感知机网络增强初始点云特征。对此可以表示为:

$$N \times (x,y,z) = \mathrm{MLP}(f_m \oplus f_{m+1}) \quad (5)$$

其中, $m = \{1,2,\cdots,n_v\}$, (x,y,z) 为点云中每一个点的坐标。

2.2 点云生成模块

由于点云数据的不规则性和无序性,传统的卷积神经网络无法直接处理点云。清华团队[20]提出 PCT 模型,将 Transformer[21]推广到点云上,显著提升了三维点云的分类和分割性能。本文的点云生成模块将采用 Transformer 模型的编码器提取深层次图像三维特征,解码器细化整合深层特征表达,最终生成相应的三维点云,模块结构如图 4 所示。

2.2.1 点云特征编码器

点云特征编码器的自注意力机制能够建立输入序列中每个位置之间的关联。在点云生成任务中,自注意力机制使得每个点的特征能够融合周围邻居点的特征,从而增强模型的三维生成能力和形状特征表达能力。

点云特征编码器将点云集合 S' 视为文本长序列,点云中每一点 s_i 则表示文本的一个词组,点的三维坐标 (x,y,z) 代表点的特征初始维度、即词组的 Embedding。得到点云与序列的对应关系后,首先要提升点的特征维度为 d_k, 将位置信息转换为特征信息编码。

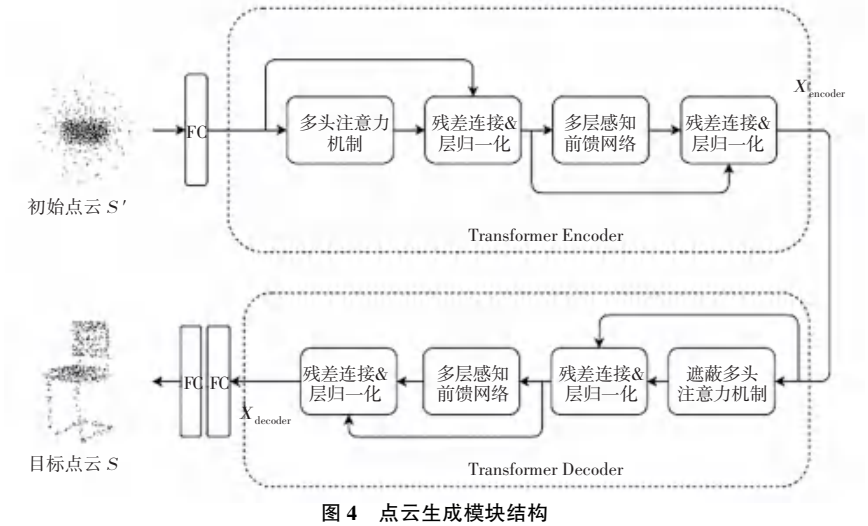


图 4 点云生成模块结构

Fig. 4 Structure of point cloud generation module

点云自注意力机制的输入为矩阵 $\boldsymbol{X}$, 表示特征升维后的点云集合 S',$\boldsymbol{X}$ 通过线性映射变换得到查询向量 ($\boldsymbol{Q}$)、键向量($\boldsymbol{K}$) 和值向量($\boldsymbol{V}$), 在点云自注意力特征形成中, $\boldsymbol{Q}$ 表示当前位置的特征表示,

用于与其他位置进行比较，计算当前位置与其他位置的关联程度；$\boldsymbol{K}$ 用于与查询向量进行比较，确定哪个查询与哪些位置更相关；$\boldsymbol{V}$ 用于根据查询和键的关联程度来计算最终的加权和。$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 通过放缩点积注意力计算相似度，公式为：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (6)$$

多头注意力机制将输入 $\boldsymbol{X}$ 进行多次投影得到 h 个注意力特征，将多个注意力头的输出 $\boldsymbol{Z}_i$ 拼接和映射，形成最终的注意力特征输出 $\boldsymbol{Z}$，公式为：

$$\boldsymbol{Z} = (\boldsymbol{Z}_0 \oplus \boldsymbol{Z}_1 \oplus \cdots \oplus \boldsymbol{Z}_h)\cdot \boldsymbol{W}_o \quad (7)$$

其中，$\boldsymbol{W}_o$ 表示可学习参数的矩阵。每个注意力头能够关注点云数据中的不同部位的形状特征，从而提升模型对输入点云的多方面建模能力。

点云特征编码器采用残差连接，残差结构用于防止梯度爆炸，每一个残差块由多头注意力机制和多层感知机网络两部分组成，数据计算前都经过层归一化（Layer Normalization，LN）处理，LN 在每个样本内部进行归一化，保留样本内部自身特征的相对大小，残差连接计算过程为：

$$\boldsymbol{Z} = \boldsymbol{X} + \text{Attention}(\text{LN}(\boldsymbol{X})) \quad (8)$$

$$\boldsymbol{X}_{\text{encoder}} = \boldsymbol{Z} + \text{MLP}(\text{LN}(\boldsymbol{Z})) \quad (9)$$

最终编码器的网络设计结构为 6 层点云残差网络块，点云位置特征升维至 $d_k = 128$，编码器输出特征向量矩阵 $\boldsymbol{X}_{\text{encoder}}$。

2.2.2　点云特征解码器

由于点云生成模块的输入输出均为点云格式，从点云生成另一种形状的点云属于点云序列建模，解码器需要处理序列之间的依赖关系，因此解码器自注意计算需要加入值为 $-\infty$ 的上三角矩阵做一个未来遮蔽 Mask，经过 Softmax 归一化使 $-\infty$ 的位置权重变为 0，即在位置 i 时，不能看到位置 $i+1$ 及以后的位置，确保模型在点云生成过程中仅依赖已生成的内容，而非未来的内容。遮蔽注意力计算公式为：

$$\text{Attention} = \text{Softmax}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}} + \text{Mask}\right)\boldsymbol{V} \quad (10)$$

解码器的残差块在编码器的基础上修改多头自注意力计算过程，添加遮蔽机制，其余部分保持不变，解码器的网络结构设计为 6 层点云残差网络块，输出解码特征向量 $\boldsymbol{X}_{\text{decoder}}$，最后经过 2 层全连接网络，转化为三维点云格式，最终得到目标点云 S。

2.3　损失函数

本文 MVCAT 模型迭代更新训练过程中采用拟合形状的倒角距离损失函数和点云生成排斥损失的加权和作为模型综合损失函数。

在点云处理中，CD 倒角距离常用于点云生成、重建和配准任务，以衡量生成点云与目标点云之间差异。CD 倒角距离通过计算生成点云中每个点到目标点云最近邻点的平均距离，以及目标点云中每个点到生成点云最近邻点的平均距离之和，来优化生成模型，使生成点云尽可能接近目标点云，从而提高生成质量和准确性。CD 损失通常作为优化目标函数的一部分，指导模型在训练过程中改善输出。

本文定义点云集合 $S_1, S_2 \subseteq \mathbb{R}^3$ 的倒角距离为：

$$L_{\text{CD}} = \sum_{x\in S_1}\min_{y\in S_2}\| x-y \|_2^2 + \sum_{y\in S_2}\min_{x\in S_1}\| x-y \|_2^2 \quad (11)$$

其中，第 1 项表示 S_1 中的任意一点 x 到 S_2 的最小距离之和，第 2 项则表示 S_2 中任意一点 y 到 S_1 的最小距离之和。如果该距离较大，则说明 2 组点云区别较大；如果距离较小，说明重建效果较好。

由于数据处理过程中对点云进行归一化处理，将每个点映射到 $[-1,1]$ 的立方体空间中。这种映射导致点与点之间数值差距较小，生成点容易相互接近、导致点堆叠现象，从而显著降低了生成点云的均匀性和特征表现。因此，根据 Yu 等学者[22]的点云上采样模型 PU-NET 的排斥损失公式，实现点云排斥损失，优化点云的生成质量。点云排斥损失计算公式为：

$$L_{\text{rep}} = \sum_{i=0}^{S}\sum_{j\in K(i)}\eta(\| x_j - x_i \|)\omega(\| x_j - x_i \|) \quad (12)$$

其中，S 表示生成点云集合；$K(i)$ 表示 x_i 的 K 近邻点的索引集合；排斥函数 $\eta(r)$ 和高斯权重函数 $\omega(r)$ 的定义公式分别如下：

$$\eta(r) = -r \quad (13)$$

$$\omega(r) = e^{-r^2/h^2} \quad (14)$$

随着距离 r 增加，函数值迅速趋近于 0，以此给每一个生成点进行加权。

因此，联合损失函数表示为：

$$L(Y,S) = L_{\text{CD}} + \partial L_{\text{rep}} \quad (15)$$

其中，∂ 表示点云排斥损失函数的权重。

3　实验与结果分析

3.1　数据集

本文采用 ModelNet40 数据集[23]进行模型训练验证，采用 ShapeNet-Core 数据集[24]进行性能评估与对比实验，训练过程和评估过程一致。

ModelNet40 包含 40 类模型，共 12 311 个三维模型文件[25]。使用每个视角间隔 30°的虚拟相机，

利用 Blender 将 off 模型渲染出 12 张不同视角下的图像,作为多视角输入,如图 5 所示。根据文献[3,14,16]等相关工作表明,生成点云模型常取 1 024 或 4 096 个点进行训练,同时对三维模型进行采样观察可知,1 024 点已经可以清晰描绘三维形状,因此使用 CloudCompare 将 off 模型采样 1 024 个点生成 ply 点云文件作为 Ground Truth。



图 5 多视角渲染图(部分)

Fig. 5 Multi perspective rendering (partial)

3.2 数据预处理

在模型训练过程中,点云生成仅关注渲染图像中的形状信息,将图像数据转换为单通道灰度图,并将不同分辨率的多视角图像统一调整为 224×224 的尺寸。针对部分模型采样点云数量少于 1 024 的问题,使用 KD-Tree 算法计算每个点的 $k = 5$ 个近邻点进行点云插值补全。为解决不同三维模型在大小和尺度上的不一致,本文使用点云尺度归一化算法,将点云坐标映射到[-1,1]的立方体空间中。

3.3 实验结果

实验设备配置为 Linux 系统,显卡为 NVIDIA A100。训练过程中点云个数 $N = 1\,024$,视图池化保存视图特征数 $n_v = 5$。数据集采用 7∶3 的训练集、验证集划分比例,训练集数据个数为 8 617,测试集数据个数为 3 694,每批训练数据 Batch Size = 32,为了加快模型收敛速度采用 ExponentialLR 学习率衰减函数,初始学习率为 0.001,gamma 为 0.98,训练迭代轮数 Epochs 为 500,训练过程中使用 Adam 优化器来更新参数,模型训练损失过程如图 6 所示。

本文选取了 ModelNet40 数据集中不同类别各 10 个模型、共计 400 个模型作为可视化对象。图 7 显示了由 MVCAT 模型生成点云样本,展示了多视图图像无相机参数生成点云的能力。对于椅子、飞机、桌子、显示器、马桶等大部分的类别,生成点云与真实建模采样点云相比具有较好形状特征,丰富了日常 3D 建模的内容。对于小部分类别如人物、吉他、浴缸、键盘等具有镂空、注重局部特征的模型,模型只能学习到整体形状特征和显著的局部特征,在局部特征建模上还有较大提升空间。

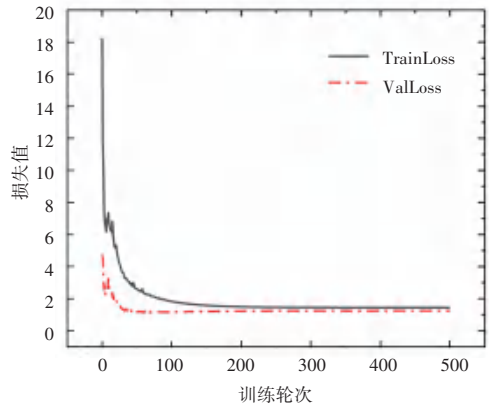


图 6 训练过程

Fig. 6 Training process

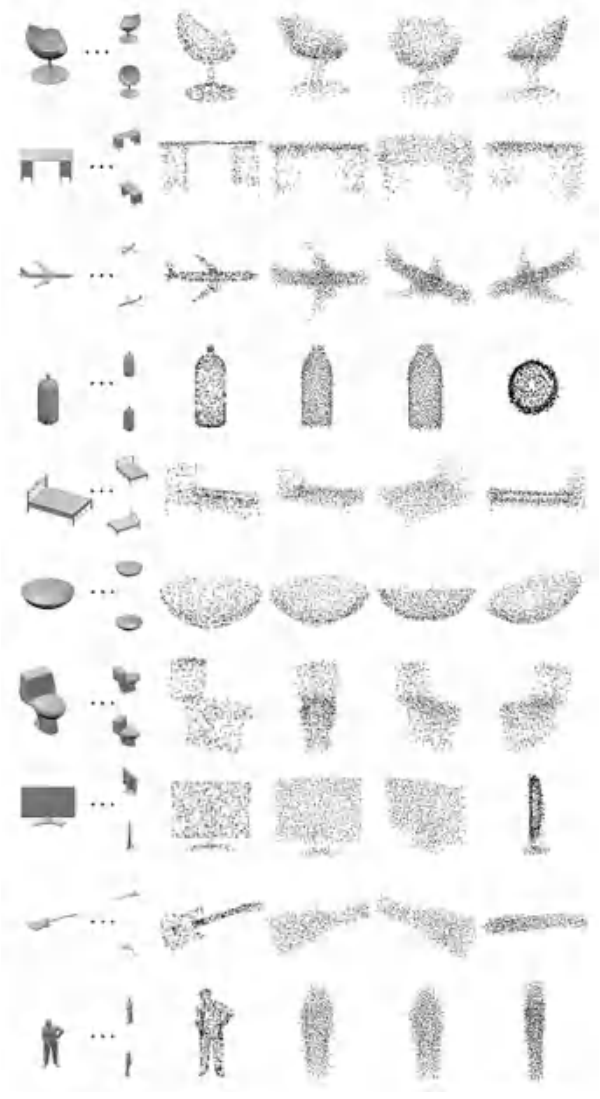


图 7 点云生成结果

Fig. 7 Results of point cloud generation

3.4　定性分析

为进一步评估本文方法的可行性与有效性，本文模型与 3D-R2N2[4]、Pix2Vox[5]、3D-AttriFlow[17]等模型进行实验结果对比。上述模型中均基于 ShapeNet 数据集，本文模型采用在 ModelNet40 上的预训练模型在 ShapeNet-Core 上重新训练，进行点云生成结果对比。对比过程中对 3D-R2N2 与 Pix2Vox 模型生成的体素网格模型进行点云采样得到点云表达结果，对比结果如图 8 所示。相较于其他模型，本文模型在整体形状和模型均匀程度上表现较为出色，在某些类别上接近或超越目前的图像三维生成方法，并且能够捕捉一些显著的细微特征，如飞机的机翼、台灯的灯杆以及椅子的腿部。因此，本文方法在处理缺乏相机参数的多视图点云生成方面具有重要的实用性。



(a) Input Image　(b) 3D-R2N2[4]　(c) Pix2Vox[5]　(d) 3D-AttriFlow[17]　(e) Ground Truth　(f) 本文模型

图 8　模型生成点云对比结果

Fig. 8　Comparison results of model generated point clouds

3.5　定量分析

本文模型进一步与 3D-R2N2[4]、Pix2Vox[5]、3D-AttriFlow[17]等模型在 ShapeNet-Core 数据集上采用相同评价指标进行了多次定量实验。评价指标包括模型复杂度相关的浮点运算数(FLOPs)、模型参数量(Params)以及模型预测时间(Time)。指标数值越大表明模型越复杂。点云生成质量评价指标相关的交并比(IoU)值越大，表示生成的点云与真实点云越相似。

表 1 展示了在相同输入图像条件下的模型复杂度评价结果。值得注意的是，本文模型相对于其他多视图三维生成模型，在浮点运算数和模型参数量上具有较小的指标值，运行时间有较大提升。

表 1　模型复杂度评价

Table 1　Complexity evaluation of the model

Model	Input	FLOPs/G	Params/M	Time/s
3D-R2N2[4]	12×3×127×127	130.32	34.30	0.587
Pix2Vox[5]	12×3×224×224	476.76	114.24	0.825
3D-AttriFlow[17]	12×3×224×224	74.01	11.64	1.022
本文模型	12×1×224×224	18.25	28.62	0.140

表 2 展示了模型的点云生成质量评价结果。从结果来看，本文模型在 lamp 模型的生成上明显不足，但在 IoU 的值上均高于其他 2 个经典模型，在某些类别上与目前最新 3D-AttriFlow 模型相比略显不足，但程度接近。这些结果表明本文模型在无相机参数下的多视图点云生成任务中具有重要意义，为多视图点云生成领域的进一步研究提供了参考价值。

表 2　多视图点云生成质量评价

Table 2　Quality evaluation of multi-view point cloud generation

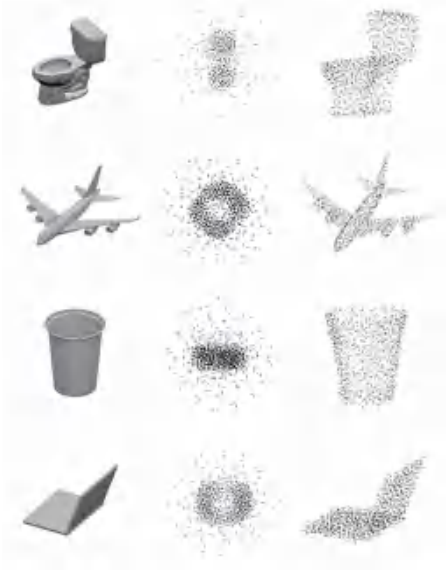
Category	3D-R2N2[4]	Pix2Vox[5]	3D-AttriFlow[17]	本文模型
airplane	0.509	0.727	0.452	0.733
car	0.798	0.877	0.515	0.703
chair	0.474	0.606	0.846	0.691
lamp	0.386	0.456	0.697	0.619
table	0.497	0.632	0.867	0.688
mean	0.533	0.659	0.675	0.687

3.6　消融实验

为探究点云生成模块的影响，仅使用多视图特征提取模块对数据进行训练。图 9 展示了对比实验结果，多视图特征提取模块对于任何输入均生成圆环点云。因此点云生成模块可以赋予模型三维化的能力，使点与点之间形成三维位置特征。

为探究视图池化操作对三维点云生成模型的影响，将视图池化过程替换成简单特征拼接，模型参数量从 28.62 M 增长为 49.62 M，浮点计算量从

18.25 G 增长为 18.41 G。图 10 展示了对比实验结果,左侧图为简单特征拼接模型结果,特征拼接模型生成效果与原模型几乎一致,但对于某些局部特征表达存在缺陷。因此,多视图池化操作可以增强特征表达,减少计算资源的消耗。



(a) Input (b) 多视图特征提取模块 (c) 本文模型

图 9 多视图特征提取模块生成结果

Fig. 9 Results of multi-view feature extraction module generation

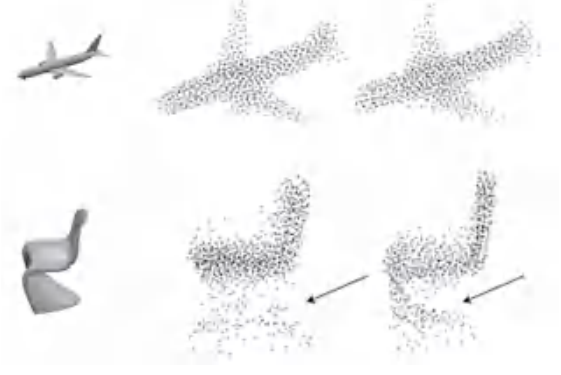


图 10 去除视图池化点云生成结果

Fig. 10 Results of removing view pooling point cloud generation

点云生成任务的常用损失函数为 CD 损失,为探究排斥损失的作用,去除排斥损失只保留 CD 损失进行点云生成。图 11 展示了对比实验结果,缺少排斥损失生成点云存在点聚集现象,导致生成点云部分特征不明显,点云分布稀疏且不均匀。因此排斥损失的引入使生成点云更均匀,并增强模型对整体形状特征的捕捉能力。

点云数据集在空间位置上表现出不同的尺度,点云的尺度会影响模型收敛速度和学习能力。图 12 展示了未归一化处理的训练过程,模型训练的损失值显著增大,模型无法学习到有效的特征,将很难收敛。因此点云尺度归一化对于增强网络的学习能力和加快模型训练速度具有重要作用。



图 11 去除排斥损失点云生成结果

Fig. 11 Results of removing exclusion loss point cloud generation

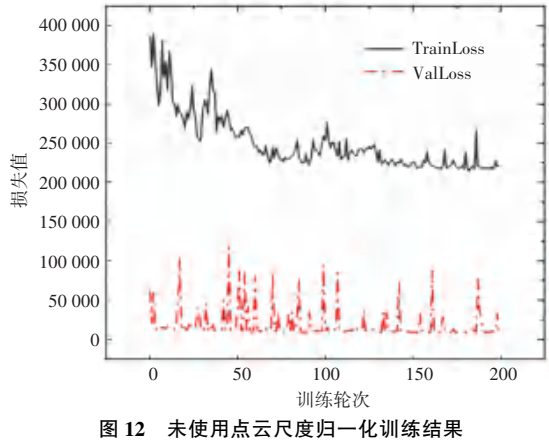


图 12 未使用点云尺度归一化训练结果

Fig. 12 Results of unused point cloud scale normalization training

4 结束语

本文针对无相机参数下的多视图点云生成问题,改进 MVCNN 三维分类模型并结合 Transformer 模型,提出 MVCAT 多视角多特征点云生成模型。该模型仅使用简单的数据处理方法就可以得到高质量的点云模型,在 ModelNet 数据集和 ShapeNet 数据集上的定性分析和定量分析表明,本文生成点云形状接近真实 CAD 模型采样点云,与经典的图像三维内容生成模型 3D-R2N2[4]、Pix2Vox[5]、3D-AttriFlow[17]等模型相比,整体形状上表现更为出色,并且能够捕捉一些显著的细微特征。在 IoU 的评价指标中都较为领先,同时在模型参数量和运行时间上提升了模型生成效率。本文提出的 MVCAT 模型成功实现了在无相机参数下选取不同视角的图像中恢复三维形状,为多视图三维重建提供了新的解决方案和研究思路。

参考文献

[1] HÄNE C, TULSIANI S, MALIK J. Hierarchical Surface Prediction for 3D Object Reconstruction[C]//Proceedings of 2017 International Conference on 3D Vision (3DV). Piscataway, NJ: IEEE, 2017: 412-

420.
[2] BAUTISTA M, TALBOTT W, ZHAI S, et al. On the generalization of learning - based 3D reconstruction [C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway, NJ: IEEE, 2021: 2180-2189.
[3] NICHOL A, JUN H, DHARIWAL P, et al. Point-E: A system for generating 3D point clouds from complex prompts[J]. arXiv preprint arXiv,2212.08751, 2022.
[4] CHOY C, XU D, GWAK J, 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 (ECCV 2016). Cham: Springer, 2016: 628-644.
[5] XIE Haozhe, YAO Hongxun, SUN Xiaoshuai, et al. Pix2Vox: Context-aware 3D reconstruction from single and multi-view images [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2019: 2690-2698.
[6] LIN K, WANG Lijuan, LIU Zicheng. End-to-end human pose and mesh reconstruction with Transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 1954-1963.
[7] YUAN Y, TANG J, ZOU Z. Vanet: A view attention guided network for 3d reconstruction from single and multi-view images [C]//Proceedings of 2021 IEEE International Conference on Multimedia and Expo (ICME). Piscataway, NJ: IEEE, 2021: 1-6.
[8] 王佳欣, 窦小磊. 复杂光场环境下多重聚焦深度图像三维重建[J]. 智能计算机与应用, 2023, 13(8): 140-143.
[9] YAO Yao, LUO Zixin, LI Shiwei, et al. MVSNet: Depth inference for unstructured multi-view stereo[C]//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer, 2018: 767-783.
[10] SPEZIALETTI R, TAN D J, TONIONI A, et al. A divide et impera approach for 3D shape reconstruction from multiple views [C]//Proceedings of 2020 International Conference on 3D Vision (3DV). Piscataway, NJ: IEEE, 2020: 160-170.
[11] MUREZ Z, VAN AS T, BARTOLOZZI J, et al. Atlas: End-to-end 3D scene reconstruction from posed images[C]// Proceedings of the 16th European Conference on Computer Vision (ECCV 2020). Cham: Springer, 2020: 414-431.
[12] WANG Shuzhe, LEROY V, CABON Y, et al. DUSt3R: Geometric 3D vision made easy[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 20697-20709.
[13] WANG Peng, TAN Hao, BI Sai, et al. PF-LRM: Pose-free large reconstruction model for joint pose and shape prediction [J]. arXiv preprint arXiv,2311.12024, 2023.
[14] FAN Haoqiang, SU Hao, GUIBAS L J. 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. Piscataway, NJ: IEEE, 2017: 605-613.
[15] JIANG Li, SHI Shaoshuai, QI Xiaojuan, et al. GAL: Geometric adversarial loss for single-view 3D-object reconstruction[C]// Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer, 2018: 802-816.
[16] NGUYEN A D, CHOI S, KIM W, et al. GraphX-convolution for point cloud deformation in 2D-to-3D conversion [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2019: 8628-8637.
[17] WEN Xin, ZHOU Junsheng, LIU Yushen, et al. 3D shape reconstruction from 2D images with disentangled attribute flow [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 3803-3813.
[18] WANG Tengfei, ZHANG Bo, ZHANG Ting, et al. RODIN: A generative model for sculpting 3D digital avatars using diffusion [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 4563-4573.
[19] SU Hang, MAJI S, KALOGERAKIS E, et al. Multi-view convolutional neural networks for 3D shape recognition [C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2015: 945-953.
[20] GUO Menghao, CAI Junxiong, LIU Zhengning, et al. PCT: Point Cloud Transformer [J]. Computational Visual Media, 2021, 7: 187-199.
[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.
[22] YU Lequan, LI Xianzhi, FU C, et al. PU-Net: Point cloud upsampling network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 2790-2799.
[23] WU Zhirong, SONG Shuran, KHOSLA A, et al. 3D ShapeNets: A deep representation for volumetric shapes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1912-1920.
[24] CHANG A, FUNKHOUSER T, GUIBAS L, et al. ShapeNet: An information-rich 3D model repository [J]. arXiv preprint arXiv,1512.03012, 2015.
[25] 杨航, 陈瑞, 安仕鹏, 等. 深度学习背景下的图像三维重建技术进展综述[J]. 中国图象图形学报, 2023, 28(8): 2396-2409.