基于 DDPG 的三维重建模糊概率点推理

李 雷 [1] 徐 浩 [1] 吴素萍 [1]

摘 要 单视图物体三维重建是一个长期存在的具有挑战性的问题. 为了解决具有复杂拓扑结构的物体以及一些高保真度的表面细节信息仍然难以准确进行恢复的问题, 本文提出了一种基于深度强化学习算法深度确定性策略梯度 (Deep deterministic policy gradient, DDPG) 的方法对三维重建中模糊概率点进行再推理, 实现了具有高保真和丰富细节的单视图三维重建. 本文的方法是端到端的, 包括以下四个部分: 拟合物体三维形状的动态分支代偿网络的学习过程, 聚合模糊概率点周围点的邻域路由机制, 注意力机制引导的信息聚合和基于深度强化学习算法的模糊概率调整. 本文在公开的大规模三维形状数据集上进行了大量的实验证明了本文方法的正确性和有效性. 本文提出的方法结合了强化学习和深度学习, 聚合了模糊概率点周围的局部信息和图像全局信息, 从而有效地提升了模型对复杂拓扑结构和高保真度的细节信息的重建能力.

关键词 三维重建, 强化学习, 深度学习, 注意力机制, 信息聚合

引用格式 李雷, 徐浩, 吴素萍. 基于 DDPG 的三维重建模糊概率点推理. 自动化学报, 2022, **48**(4): 1105−1118

DOI 10.16383/j.aas.c200543

Fuzzy Probability Points Reasoning for 3D Reconstruction Via Deep Deterministic Policy Gradient

LI Lei[1] XU Hao[1] WU Su-Ping[1]

Abstract 3D object reconstruction from a single-view image is a long-standing challenging problem. In order to address the difficulty of accurately predicting the objects of complex topologies and some high-fidelity surface details, we propose a new method based on DDPG (Deep deterministic policy gradient) to reason the fuzzy probability points in 3D reconstruction and achieve high-quality detail-rich reconstruction result of single-view image. Our method is end-to-end and includes four parts: the dynamic branch compensation network learning process to fit the 3D shape of objects, the neighborhood routing mechanism to aggregate the points around the fuzzy probability points, the attention guidance mechanism to aggregate the information, and the deep reinforcement learning algorithm to perform probabilistic reasoning. Extensive experiments on a large-scale public 3D shape dataset demonstrate the validity and efficiency of our method. Our method combines reinforcement learning and deep learning, aggregates local information around the fuzzy probability points and global information of the image, and effectively improves the model′s ability to reconstruct complex topologies and high-fidelity details.

Key words 3D reconstruction, reinforcement learning, deep learning, attention mechanism, information aggregation

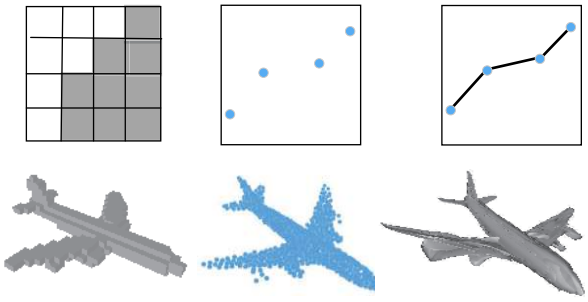
Citation Li Lei, Xu Hao, Wu Su-Ping. Fuzzy probability points reasoning for 3D reconstruction via deep deterministic policy gradient. *Acta Automatica Sinica*, 2022, **48**(4): 1105−1118

单视图三维重建是图像理解和计算机视觉的一个基本问题, 并在机器人、自动驾驶、虚拟现实和增强现实中有着广泛的应用[1−2]. 近年来, 基于深度学习的单视图三维重建得到了广泛的应用. 相比于传统的三维重建方法, 学习模型能够更好地对输入信息进行编码以防止输入信息的歧义.现有基于深度学习的三维重建分为多视图和单视图重建[3−6], 前者先利用深度网络提取到的特征信息进行立体匹配并预测深度图, 再利用深度图融合技术构建三维模型. 后者则通过使用神经网络强大的特征捕获能力从输入图像中捕获特征信息, 之后结合从海量训练数据中学习到的形状先验知识信息进行三维重建. 具体来说, 基于深度学习的单视图三维重建根据三维形状输出表示形式可以分为以下三种:

1) 基于体素的表示形式, 如图 1(a) 所示, 现有工作[7] 使用编码网络捕获输入的物体图片的形状属性信息 (物体拓扑结构以及几何、轮廓、纹理等信

收稿日期 2020-07-13 录用日期 2021-01-15
Manuscript received July 13, 2020; accepted January 15, 2021
国家自然科学基金 (62062056, 61662059) 资助
Supported by National Natural Science Foundation of China (62062056, 61662059)
本文责任编委 吴毅红
Recommended by Associate Editor WU Yi-Hong
1. 宁夏大学信息工程学院 银川 750021
1. School of Information Engineering, Ningxia University, Yinchuan 750021

息) 并将这些低层级信息编码为不同尺度下的高层级表示形式, 之后使用解码网络将三维几何外形表示为三维体素块上的二值概率分布 $S=\{(P_1, \cdots, P_{n\times n\times n})\}$, 最后通过计算网络预测的二值概率分布和真实二值概率分布之间的交叉熵来约束网络学习, 即利用网络学习二维图像到三维体素块上二值概率分布的映射关系来表达三维几何外形.



(a) 体素 (a) Voxel (b) 点云 (b) Point cloud (c) 网格 (c) Mesh

图 1 基于深度学习的单视图三维重建中三种表示形状

Fig. 1 Three representation shapes for single-view 3D reconstruction based on deep learning

2) 基于点云的表示形式, 如图 1(b) 所示, 现有工作[8] 使用编码网络捕获输入的物体图片的形状属性信息,之后使用解码网络将三维几何外形表示为无序点云 $S=\{(x_i, y_i, z_i)_{i=1}^{N}\}$ 最后通过计算预测的点云三维坐标和真实点云三维坐标之间的倒角距离等指标来约束网络学习, 即利用网络学习二维图像到无序三维点集 S 的映射关系来表示物体三维形状.

3) 基于网格的表示形式, 如图 1(c) 所示, 现有工作[9] 首先使用卷积神经网络提取输入的物体图片的特征信息, 之后使用图卷积网络[10] 结合提取特征和初始化的网格模板对初始化模板进行网格变形生成目标三维模型, 最后通过计算预测网格的信息 (点坐标、边长等) 和真实网格信息之间的误差来约束网络学习, 即利用网络学习二维图像到三维网格的映射关系来表示物体三维形状.

在网络学习过程中, 现有方法都使用反向传播[11] 算法通过监督信息来约束编解码网络进行学习, 即通过反向传播使神经网络拟合一个复杂的映射函数. 本质上,基于深度学习的单视图三维重建方法使用合适的神经网络 N 来实现从输入图像 I 到输出 Y 的连续映射函数逼近, 即对任意 $\varepsilon > 0$, $x \in I$; $|N(x) - Y| < \varepsilon$.

大部分基于深度学习的单视图三维重建工作都使用基于卷积神经网络的编解码器架构[12], 即三维重建任务通常采用 2D 卷积神经网络对二维输入图像进行编码, 再根据任务需要的表示形式, 使用不同的解码器生成不同的表示形式. 例如, 如果使用体素[13] 作为最终表示, 则使用 3D 反卷积神经网络作为解码器.

根据重建后的三维形状输出表示形式,一些工作[14−16] 基于网格进行三维形状重建. 因为这些方法只能通过使用同类形状模板进行变形, 所以上述方法只能重建出具有简单拓扑的物体, 并且容易出现网格自交叉. 总的来说, 由于没有明确和可靠的方法生成有效的网格, 所以基于网格的三维重建工作面临着巨大的挑战. 一些工作基于体素[4−5, 7] 和点云[8] 来进行三维形状重建, 但由于占用内存过高只能处理小批量数据和采用低分辨率来表示. 为了解决上述问题, Mescheder 等[17] 提出了由连续函数定义一个 3D 空间, 并通过神经网络拟合的函数来描述这样的隐式形状, 并使用 2D 图像 X 和位置 $P \in \mathbf{R}^3$ 来推断对应位置 P 的占用情况. 即使用神经网络拟合映射函数 $\mathbf{R}^3 \times X \rightarrow [0, 1]$. 该方法有效地减少了训练时占用的内存和训练时间, 但由于物体三维形状是由分类器或回归模型的权值来表示, 所以这些方法忽略了一些低级的形状信息. 总的来说, 现有的单视图三维重建方法存在以下挑战性问题: 1) 难以准确地重建具有复杂拓扑结构的物体三维形状. 2) 难以准确地重建局部细节特征从而生成高保真输出. 3) 先前的工作都是在合成数据上进行训练, 但在真实数据上进行测试时, 就会出现领域自适应问题. 因此, 一些复杂拓扑结构的连接处和局部细节的位置点占用概率往往难以准确的预测, 本文称这些难以准确预测的点为模糊概率点.

为了解决上述的挑战性问题, 本文通过深度强化学习算法 DDPG[18] 来训练智能, 并不断地调整这些模糊概率点的占用概率并使其跳出概率模糊区间 $P \in [0.4, 0.6]$. 具体来说, 受到 Li 等[19]的启发, 本文首先通过动态分支代偿网络生成了更多样化的特征表示并得到预测结果, 之后通过预测结果找到模糊概率点后聚合模糊概率点周边的局部信息和全局图像信息, 再通过 DDPG 训练的智能体调整这些模糊概率点, 使其达到到最佳的占用概率. 本文给出了本文方法在真实图像上进行三维重建的结果, 如图 2 所示. 本文的主要贡献如下:

1) 本文使用动态分支代偿网络来使得模型从输入图像中捕捉到更多样化的特征信息以提高模型的泛化能力.

2) 本文考虑到了局部信息对位置点占用概率预测的影响并使用了注意力机制引导的信息聚合机制聚合了局部信息和全局图像信息.

3) 本文使用深度强化学习算法 DDPG 训练的

(a) 真实图片
(a) Real image

(b) DISN
(b) DISN

(c) 本文结果
(c) Our results

图 2　本文方法和 DISN 方法在真实图像上的单视图重建结果

Fig. 2　Single image reconstruction using a DISN, and our method on real images

智能体对模糊概率点的占用概率进行了再推理.

4) 大量定量、定性和消融实验证明了本文的方法在公开的大规模三维物体数据集 ShapeNet[20] 上的评估相比最先进的方法都有相应的提升.

1　相关工作综述

早期的单视图三维物体重建是通过 shape-from-shading[21−22] 重建物体三维形状.在早期方法下, 纹理和散焦信息提供了更多有意义的重建信息. 具体来说, 这些工作从输入图像中捕获多条线索信息 (例如: 纹理、散焦等) 和物体的几何结构信息来推理物体可见表面的深度信息.

近年来, 随着生成对抗网络[23],可变分自编码器[24] 在图像生成方面取得显著成果, Wu 等[4] 将生成对抗网络从图像领域扩展到体素, 并训练了 3D 生成对抗网络从潜在向量生成三维体素. Kar 等[25]将相机参数和输入图像编码为 3D 体素代表, 之后应用 3D 反卷积从多个视图重建 3D 场景. 先前的工作都是在合成数据上进行训练, 但在真实数据上进行测试时, 存在领域自适应问题. 为了解决上述问题, Wu 等[7] 使用全监督的方式通过学习输入图像到 2.5D 草图的映射, 再通过训练一个三维形状估计器得到最终的三维形状. 但由于过高的内存占用, 重建的三维形状通常被限制在 32^3 分辨率的体素块内. 为了解决内存限制的问题, Tatarchenko 等[26] 对输出空间进行了分层分区, 以提高计算和存储效率, 这有助于预测更高分辨率的三维形状. Wang 等[9] 使用图卷积网络[10] 使椭球面模板逐渐形变成为目标对象, 但结果往往是受限于球形拓扑. Wang 等[27] 通过变形初始化的源网格来重建三维形状. Fan 等[8] 引入了点云作为表示形式来表示物体 3D 形状. 然而, 基于点云的表示形式需要许多复杂的后处理步骤[28−30] 来生成三维网格. Choy 等[5] 借鉴了长短期记忆网络和门控循环单元 (Gated recurrent unit, GRU) 思想构建了循环网络结构来重建三维物体. Groueix 等[14] 则使用小块面片来拼接三维物体表面形状, 但是拼接三维形状的小块面片之间很容易出现重叠和自交. Chen 和 Zhang[31]在深度网络中使用符号距离函数来完成三维形状生成的任务.虽然他们的方法在生成任务上取得了良好的效果, 但在单视图重建任务中无法实现高保真恢复三维物体的细粒度细节.

最近, 一些工作[31−32] 将物体三维形状表面隐式地表示为深度神经网络分类器的连续决策边界. 换句话说, 这些工作通过学习一个分类器来预测一个点是在封闭的形状边界内还是在边界外, 并使用这个分类器作为三维形状的表示形式. 但是,由于三维形状是由分类器或回归模型的权值来表示的, 所以这些方法往往忽略了一些低级的形状信息.

2　方法

2.1　概述

本文的目标是使用 2D 图像 X 和位置 $P \in \mathbf{R}^3$ 来推断对应位置 P 的占用情况, 即使用神经网络拟合映射函数 $\mathbf{R}^3 \times X \rightarrow [0, 1]$. 对于一个封闭的形状 S, 该二分类神经网络等价于对每个位置 P 给出一个 0 到 1 之间的占用概率来决定 P 点是否在封闭形状内, 如式 (1) 所示:

$$S(P) = \begin{cases} 0, & \text{if } P \notin \text{shape} \\ 1, & \text{if } P \in \text{shape} \end{cases} \tag{1}$$

但三维物体存在着大量复杂的拓扑结构和表面细节, 这些位置的占用情况往往难以准确的预测, 本文称这些点为模糊概率点. 因此, 本文结合模糊概率点的邻域特征和全局特征并使用深度强化学习算法对模糊概率点的占用概率进行再推理. 图 3 显示了本文方法的整个流程. 具体来说, 本文首先使用编解码器将 2D 图片信息和下采样点特征解码为向量 $\boldsymbol{V}$ 并找到模糊概率点. 其次, 邻域路由机制将搜索模糊概率点周围的邻域点并组合为邻域点阵块. 然后, 特征聚合模块提取邻域点阵块和对应的图片信息为模糊概率点集的局部特征, 并与全局图像特征进行聚合. 最后, DDPG 模块将聚合后的特征作为初始状态后进行动作选取并输出调整后的占用概率.

2.2　动态分支代偿网络模块

本小节分别介绍了动态分支代偿网络模块的整

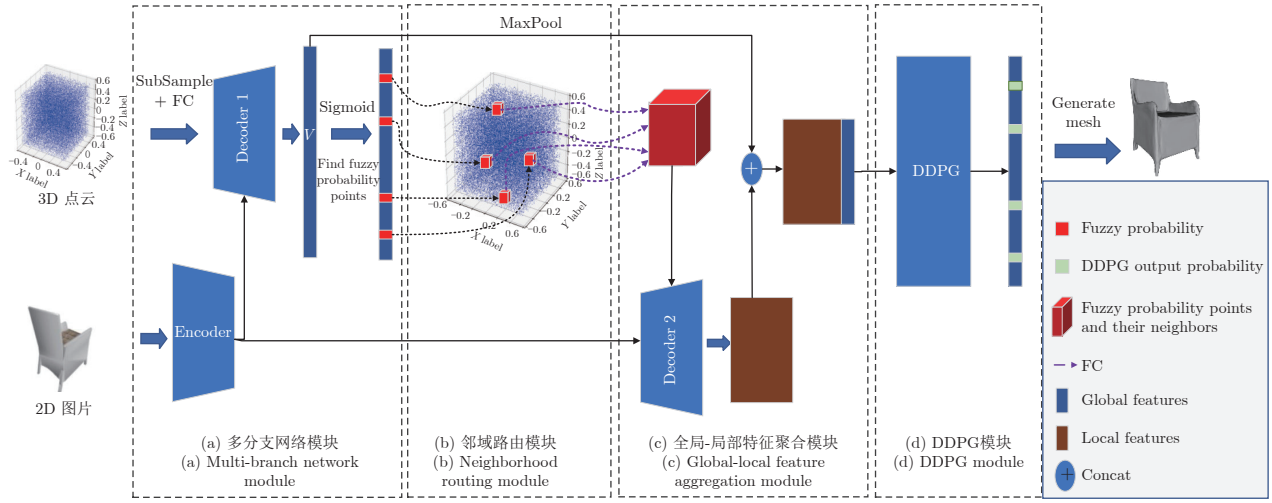


图 3 MNGD 框架的整体流程图

Fig. 3 The workflow of the proposed MNGD framework

体流程、动态分支代偿网络以及分支网络的优化和代偿过程.

2.2.1 动态分支代偿网络模块整体流程

本文使用动态分支代偿网络编码输入的 2D 图像与下采样后的位置点送入解码器中对各自位置的占用情况进行预测得到向量 $\boldsymbol{V}$, 再从向量 $\boldsymbol{V}$ 中寻找模糊概率点. 首先将向量 $\boldsymbol{V}$ 转换到 0 到 1 之间, 再取得概率分布在 $P \in [0.4, 0.6]$ 的点作为模糊概率点集. 如式 (2) 所示, 本文对模糊概率点集进行初始化.

$$\text{fuzzy}(p)=\begin{cases} 0, & p<0.4 \ \vee \ p>0.6 \\ 1, & 0.4 \le p \le 0.6 \end{cases}$$
$$\text{s.t.} \quad p \in \text{Sigmoid}(\boldsymbol{V}) \tag{2}$$

上式中, fuzzy(.) 表示对应的点是否为模糊概率点, 0 代表该点不是模糊概率点, 1 代表该点是模糊概率点并将其加入模糊概率点集.

2.2.2 动态分支代偿网络

如图 4 所示, 本文通过在神经网络的中间层中添加边分支, 使得神经网络能够沿着每条边分支产生更多样化的特征表示. 之后, 本文通过注意力机制来动态混合多分支输出预测概率, 从而得到更精确地预测占用概率, 如式 (3) 所示:

$$P=\sum_{i=1}^{3} w_i \times p_i$$
$$\text{s.t.} \ \ i \in M[1,3], \sum_{i=1}^{3} w_i = 1 \tag{3}$$

其中, p_i; $i \in M[1, 3]$ 代表每条分支的预测的占用概率, 并且根据当前处理的样本, w_i 代表针对该样本每一条分支的权重值.

2.2.3 动态分支网络的优化和代偿过程

在优化动态分支代偿网络时, 本文不仅直接收集每个分支的分类损失来优化网络, 而且关注每个分支在其各自路径中生成的不同特征. 即当边分支或主分支学习并生成知识时, 分支网络之间可以通过相互的公共路径实现实时的知识交互和补偿, 如式 (4) 所示:

$$L_B = \frac{1}{|B|}\sum_{i=1}^{|B|}\sum_{j=1}^{K} L_{\text{M}_{\text{ce}}}(f_\theta; I_\Theta(p_{ij}, x_i), o_{ij}) + \frac{1}{|B|}\sum_{n=1}^{N}\sum_{i=1}^{|B|}\sum_{j=1}^{K} L_{\text{S}_{\text{ce}}}(f_{\Theta_n}; I_\theta(p_{ij}, x_i), o_{ij}) \tag{4}$$

其中, L_B 代表一个小批次 B 条数据的训练损失, $L_{\text{M}_{\text{ce}}}(.,.)$ 代表主分支的交叉熵分类损失, $L_{\text{S}_{\text{ce}}}(.,.)$ 代表边分支的交叉熵分类损失. $i = 1, \cdots, B$ 代表一个小批次第 i 条数据, $n = 1, \cdots, N$, N 代表第 n 条边分支. f_θ 代表主分支的网络参数, f_{Θ_n} 代表第 n 条边分支的网络参数. I_Θ 代表主分支生成的知识补偿给边分支, I_θ 代表边分支生成的知识补偿给主分支. p_{ij} 代表输入的第 i 条数据的第 j 个位置点, 其中 $p_{ij} \in \mathbf{R}^3$, $j = 1, \cdots, K$. $o_{ij} \in \{0, 1\}$ 代表 p_{ij} 是否在封闭形状 S 内的真实标签.

2.3 邻域路由模块

在本小节中, 本文通过邻域路由聚合模糊概率点周围位置点形成邻域点阵.因为模糊位置点周围信息可以帮助模型对模糊位置点的占用概率进行更

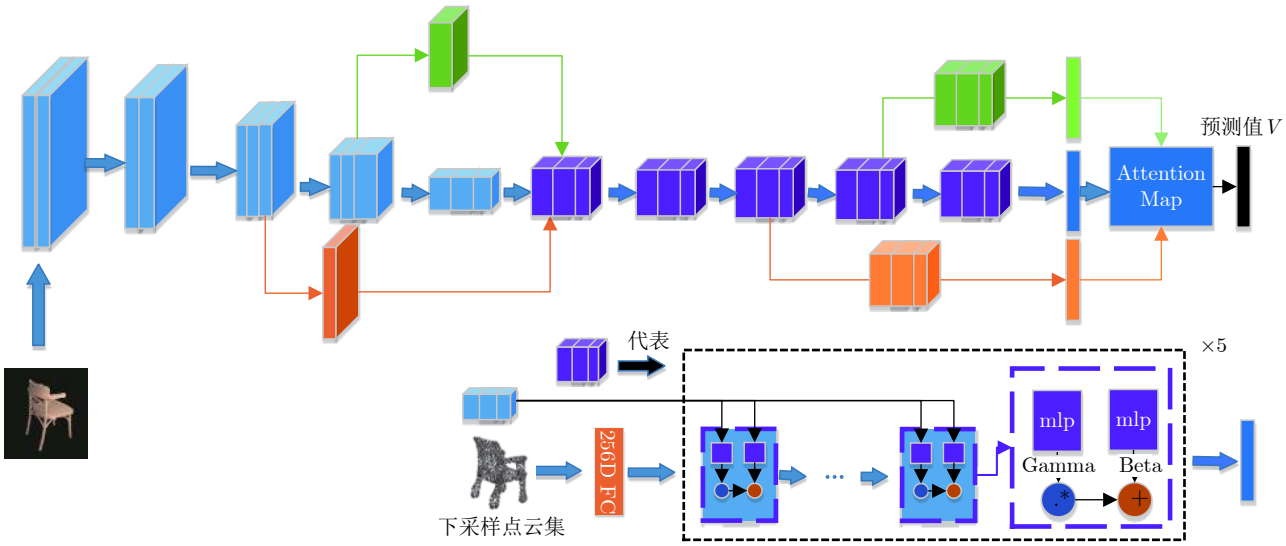


图 4　动态分支代偿网络框架图

Fig. 4　The framework of the dynamic branch compensation network

好的再推理, 所以本文通过邻域路由来寻找模糊位置点周围的点并组合为邻域点阵, 如图 3(b) 和图 5 所示. 在邻域路由中, 如果路由点个数过多则会导致训练速度较慢, 如果路由点个数过少则会导致网络捕捉的局部特征不稳定. 本文均衡了不同路由点个数在训练时间和实验效果上的表现选择了 $N = 64$ 作为路由点个数.

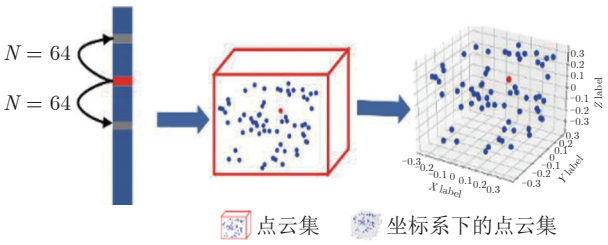


图 5　邻域路由过程

Fig. 5　The whole process of neighbor routing

2.4　注意力引导的特征聚合模块

以往的工作仅仅通过训练多层感知机抽取图像全局信息来对位置点的占用概率进行推理, 它往往会忽视局部的细节信息, 故模糊概率点再推理需要结合全局和局部信息. 如图 3(c) 所示, 本文借鉴通道注意力机制来聚合全局和局部信息. 本文的目的是将邻域点阵和对应图像结合提取出对应邻域点阵含有的局部特征, 之后再聚合局部特征和图像全局特征形成下一层网络的输入. 然而不同的样本有着不同的特征, 一些模糊概率点需要更多的全局的语义信息, 相反另一些模糊概率点则需要更多的局部细节特征, 所以本文在聚合特征时加入了通道注意力机制[33]. 如图 6 所示, 模糊概率点集特征被分割为 N 个模糊概率点特征并使用图示过程对不同的通道 c_i 乘以对应的权重 w_i 后形成新的单模糊点特征, 最后将新的 N 个模糊概率点再次组成模糊概率点集特征.

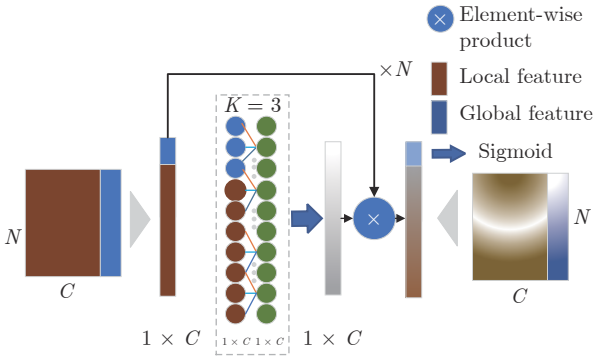


图 6　聚合特征时的注意力机制

Fig. 6　Attention mechanism when features are aggregated

2.5　DDPG 模块

在这个模块中, 智能体通过分析模糊概率点附近的局部信息和全局信息后从量化后的连续动作空间选取动作来改变模糊概率点的占用概率, 再根据奖励进行优化, 最终学习到可以跳出模糊概率范围的策略, 进而调整模糊概率点的占用概率使其跳出模糊概率区间. 本文给定一个由上一个模块输出的模糊概率点集特征 F, 并通过 DDPG 模块来调整模糊概率点的占用概率. 首先, 模糊概率点集特征被分割为 N 个单模糊概率点特征$\{F_0, F_1, \cdots, F_N\}$, 之

后引出根据模糊概率点特征如何获得该模糊概率点最佳占用概率的问题, 最后该问题被定义为一个马尔科夫决策过程, 由状态、行为、奖励、在过程中采取行动的状态改变、学习过程和交互环境组成. 智能体通过当前输入的状态信息, 输出相应的最优动作, 从环境中获得最大的奖励作为一个流程. 本文使用一个深度强化学习算法 DDPG[18] 来训练智能体. 本文定义的整个马尔科夫决策过程和训练过程如下:

状态: 本文将状态定义为当前模糊概率点的特征信息 F_i, 初始状态 S_0 为当前模糊概率点第一次进入 DDPG 模块的模糊概率点特征 F_i. 随着每一步迭代, 在第 i_{th} 迭代后状态是 S_{i-1}, 它累积了以前所有迭代的更新.

动作空间: 本文将在连续空间 $T \in [0, 1]$ 内选取一个动作 A_i 来调整模糊概率点特征信息用来获得最准确的占用概率, 所以第 i_{th} 迭代的操作是为了获得更准确的占用概率对应的模糊概率点特征信息. 其次, 本文量化了连续空间 $T \in [-1, 1]$ 作为动作选取的连续空间, 使模糊概率点可以跳出模糊概率区间 $P \in [0.4, 0.6]$.

奖励函数: 奖励函数通常用于评估智能体执行动作的结果. 在本文中, 奖励函数被设计为根据模糊概率点特征所得到的占用概率与模糊点对应的真实标签的交叉熵损失, 所以这个奖励函数可以评估该模糊概率点对应的重建误差. 奖励函数形式如下:

$$R_i(S_i, A_i) = \begin{cases} -(1-y_i)\log_2(1-P_i), & \text{if } y_i = 0 \\ -y_i \log_2(P_i), & \text{if } y_i = 1 \end{cases}$$
$$\text{s.t.} \quad P_i = \text{Sigmoid}(\text{AvgPool}(S_i + A_i)) \tag{5}$$

上式中, $y_i \in \{0, 1\}$ 代表周围点是否在封闭形状内的真实标签, P_i 是模糊概率点的占用概率, S_i 代表当前状态, A_i 代表基于当前状态给出的动作.

学习过程: 学习阶段的目标是更新评价和动作网络的参数, 初始化的评价网络 $Q(S, A|\theta^Q)$ 参数为 θ^Q, 初始化的动作网络 $u(S, A|\theta^u)$ 参数为 θ^u, 目标评价网络 Q' 和目标动作网络 u'.

在动作网络预测出动作后评价网络会给出 $Q(S, A)$ 值作为当前状态 S 执行动作 A 后的评价, 为了能使评价网络精确给出对应的 Q 值, 最小化式 (6) 来更新评价网络参数:

$$L = \frac{1}{N}\sum_i \left[R_i + \gamma Q'(S_{i+1}|\theta^{u'})|\theta^{Q'}) - Q(S_i, A_i|\theta^Q)\right]^2 \tag{6}$$

上式中, N 代表从记忆库中选择学习的样本个数, R_i 代表奖励值, γ 代表折扣参数.在动作网络中为了能使动作网络能够获得更大 Q 值, 使用式 (7) 来更新动作网络的参数:

$$\nabla_{\theta^u} u \mid S_i \approx \frac{1}{N}\sum_i \nabla_A Q\left(S, A \mid \theta^Q\right)\Big|_{S=S_i,\, A=u(S_i)} \nabla_{\theta^u} u\left(S \mid \theta^u\right)\Big|_{S_i} \tag{7}$$

上式中, ∇ 代表梯度, 接下来采用式 (8)(9) 更新目标网络参数:

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'} \tag{8}$$

$$\theta^{u'} \leftarrow \tau\theta^u + (1-\tau)\theta^{u'} \tag{9}$$

上式中 τ 代表软更新的参数.

交互环境: 环境指的是智能体执行动作时所处的场景. 本文将环境设置为当智能体对模糊概率点占用概率进行再推理后得到该模糊概率点新的占用概率. 当新的占用概率处在模糊概率区间 $P \in [0.4, 0.6]$ 时, 环境将该模糊概率点新的占用概率替换之前的占用概率后作为下一个状态. 当新的占用概率跳出了模糊概率区间 $P \in [0.4, 0.6]$ 或者达到了最大调整步数时, 环境将该模糊概率点新的占用概率替换之前的占用概率并给出相应的奖励值.

3 网格生成过程

本节首先可视化了卷积神经网络抽取特征的过程和特征激活图. 其次, 本文详细说明了整个预测三维形状网格和真实三维形状网格的生成过程.

3.1 卷积可视化

如图 7(a) 所示, 本文可视化了每层卷积网络处理提取前层输出后得到的结果.从整个可视化后的卷积过程可以看出: 1) 浅层网络主要是对低级特征的提取, 例如边缘特征和纹理特征等. 2) 深层网络则主要是对高级特征信息进行抽取, 例如高级语义信息等. 如图 7(b) 所示, 本文使用 Grad-CAM[34] 通过抓取梯度计算量和特征图相乘来计算热力图, 并在原图中进行特征激活可视化.

3.2 网格生成过程

如图 7((e) ~ (j)) 所示, 本文使用多分辨率等值面提取法 (Multiresolution isosurface extraction, MISE)[17] 将本文框架输出的结果提取等值面, 最后使用游动立方体算法 (Marching cubes algorithm)[35] 对每个体素以三角面片来逼近其内部的等值面 (Isosurface) 来生成目标网格.

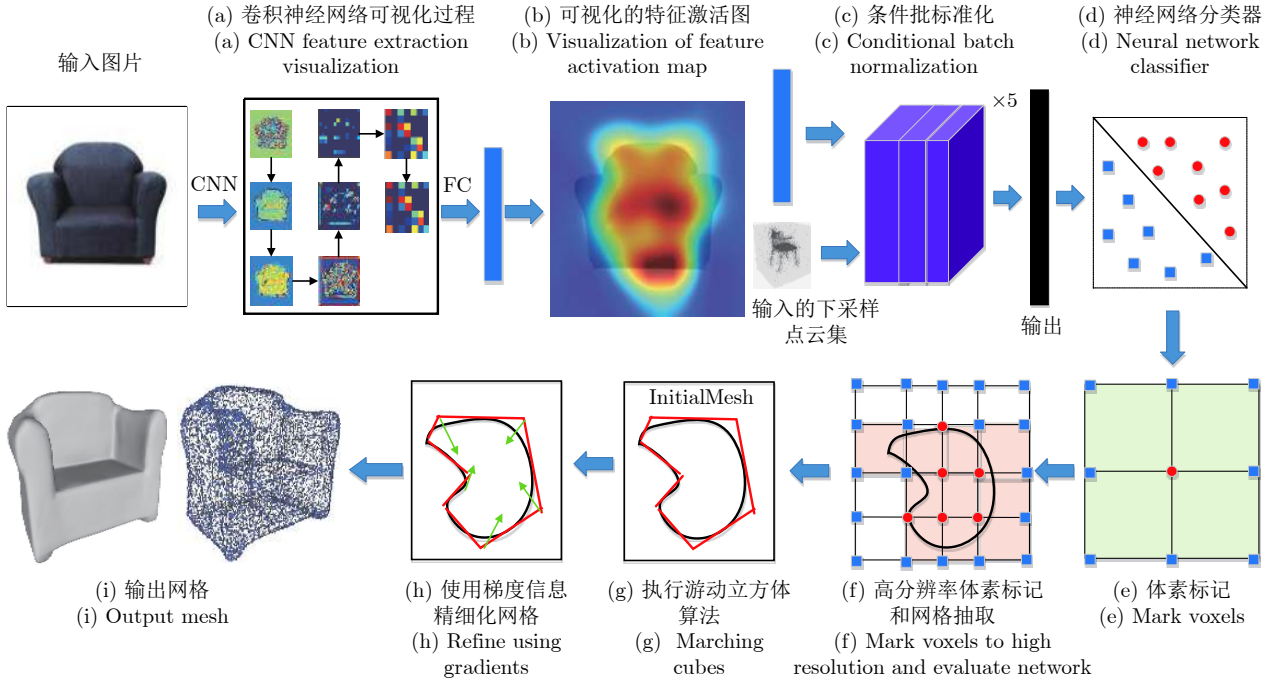


图 7　卷积可视化与网格生成过程

Fig. 7　Convolution visualization and mesh generation process

如图 7(d) 所示, 本文首先将本文框架输出的结果 $f_\theta(p, x)$ 中大于阈值 ε 的点标记为占用点 (圆形点), 剩下的标记为未占用点 (方形点). 其中, f 为网络拟合的映射函数, θ 为网络的参数, p 为输入的下采样点云, x 为输入的 2D 图像.

之后, 如图 7(e) 所示, 本文将至少两个相邻位置具有不同占用状态的点所在的体素块标记为占用体素, 这些相交于网格之间的标记体素块的个数记为网格分辨率.如图 7(f) 所示, 本文将占用体素进行精细化分解以得到更高的分辨率. 如图 7(g) 和 7(h) 所示, 本文使用游动立方体算法提取近似等值面并得到初始网格.

$$\{p \in \mathbf{R}^3 | f_\theta(p, x) = \varepsilon\} \tag{10}$$

如果在当前分辨率下的初始网格包含网格内外各连通部分的点, 则算法结果收敛于正确的网格. 为了更好地使算法收敛于正确的网格, 网格分辨率通常设置为 32^3.

如图 7(i) 所示, 本文首先使用快速二次网格简化算法 (Fast-quadric-mesh-simp-lification algorithm)[36] 来精细化得到的初始网格. 之后, 本文再使用一阶和二阶梯度来进一步精细化网格. 为了达到这个目标, 本文从输出网格的每个面取随机采样点 k 个点, 并最小化式 (11):

$$\sum_{k=1}^{K} (f_\theta(p, x) - \varepsilon)^2 + \lambda \left\| \frac{\nabla_p f_\theta(p_k, x)}{\|\nabla_p f_\theta(p_k, x)\|} - \boldsymbol{n}(p_k) \right\|^2 \tag{11}$$

上式中 $\boldsymbol{n}(p_k)$ 代表网格 p_k 的法向量. 本文设置 $\lambda = 0.01$ 并通过双重反向传播[37] 方法来有效地使用二阶梯度信息.

4　实验

本节首先介绍了实验设置以及训练和测试的实现细节. 其次, 展示了本文的方法与目前最先进的方法在 ShapeNet[20] 数据集上定量实验的结果. 之后, 本文展示了本文的方法在合成数据集 ShapeNet 上和在线产品数据集 (Online products dataset)[38] 上的定性结果. 最后, 本文使用消融实验对模型中的各个模块的作用进行了验证并展示了本文方法在合成数据 ShapeNet 中全部类别上的定性结果.

4.1　数据集

数据集: ShapeNet[20] 是一个包含大量三维物体模型的合成数据集. 本文使用 ShapeNet 核心数据集, 它包括了 55 个常见的对象类别共 55 300 个 3D 模型、13 个主要类别和一个官方分割的训练、测试数据集. 本文的模型在包含全部类别的数据集上进行训练, 并报告本文模型在包含全部类别的测试集上的测试结果.

真实图片数据: Online products dataset[38] 是一个包含真实世界图片的数据集. 本文的模型并没有在该真实图片数据集上进行训练, 所以本文使用该真实图片数据集来验证训练模型的泛化能力、真

实世界物体图片的三维重建能力和域可转移能力.

本文在实验中分别进行了定量评估和定性评估, 对于定量评估, 本文使用交并比 (Intersection over union, IoU)、倒角距离 (Chamfer distance, CD) 和法线一致性 (Normal consistency, NC) 作为评估指标. 为了测量 IoU 等评估指标,本文使用 Stutz 等[39] 的代码去生成水密网格并确定位置点是否位于网格的内部或外部.

4.2 评价标准

4.2.1 交并比

本文使用基于体素化网格的交并比 (IoU), 体素化网格的交并比是两个体素化网格交集的体积和并集的体积的商.

$$\text{IoU}(G,\ R) = \frac{|G \cap R|}{|G \cup R|} \tag{12}$$

如式 (12) 所示, G 和 R 代表体素化网格.

4.2.2 倒角距离

倒角距离被定义为真实点云形状和重建点云形状之间完整性和准确性的度量. 准确性为重建后各点与真实标签各点之间的平均距离, 完整性则为真实标签各点与重建后各点的平均距离.

$$\begin{aligned}\text{CD}(G,\ R) = &\frac{1}{|R|}\sum_{r\in R}\min_{g\in G}||r-g||_2 + \\ &\frac{1}{|G|}\sum_{g\in G}\min_{r\in R}||g-r||_2\end{aligned} \tag{13}$$

如式 (13) 所示, G 和 R 分别代表真实点云形状和重建点云形状.

4.2.3 法线一致性

法线一致性定义为重建网格中法线和真实网格中相对应最近的法线之间点积的绝对值的平均值.

$$\begin{aligned}\text{NC}(G,\ R) = &\frac{1}{|R|}\sum_{r\in R,\ g\in G}|r\cdot g| + \\ &\frac{1}{|G|}\sum_{r\in R,\ g\in G}|g\cdot r|\end{aligned} \tag{14}$$

如式 (14) 所示, G 和 R 分别代表重建网格和真实网格中的法线.

4.3 实现细节

本文所有的网络结构都是使用 Python3.6 和 Pytorch1.0 实现的. 本文使用在 ImageNet 数据集上预训练的 ResNet-18 来初始化编码器的参数, 并使用基于条件批归一化 (Conditional batch normalization, CBN)[40] 的 5 个 ResNet[41] 块作为解码器. 本文在一块 CUDA 9.0、cudnn 7 的 GeForce RTX 2080 Ti 上来训练动态分支代偿网络, 如图 3(a) 所示, 其中 48 个样本作为一个批次, 并使用初始学习率为 1×10^{-4} 的优化算法 Adam[42]. 特征聚合模块和 DDPG 模块中, 本文使用基于 CBN 的解码器和容量为 C 的记忆库. 强化学习过程中动作网络每次在连续空间中随机选择一个动作并且将元组 (S_i, A_i, S_{i+1}, R, $done$) 添加到记忆库. 当记忆库的数据量达到 C−1 时, 网络通过随机从记忆库中选取一个批次 N 条经验来进行学习. 本文设置批次样本个数 N 为 100, 容量 C 设置为 400 000, 折扣参数 γ 设置为 0.99, 最大调整步数 S 为 3, 软更新参数 ε 设置为 0.005.

4.4 定量实验

在这一节中, 本文将在单视图重建方面使用本文的方法和其他最先进的方法 3D-R2N2[5]、Pix2Mesh[9]、AtlasNet[14] 和 ONet[17] 进行定量的比较. 本文的方法仅使用单张图片输入来实现三维重建. 如表 1 所示, 本文的方法在 IoU 上的评价结果均有显著的提升并优于其他最先进的方法, 由于 AtlasNet 不能产生水密网格, 所以本文无法评估该方法的 IoU 值. 如表 2 所示,对于法线一致性 (NC), 本文的方法在 NC 上的评价结果均有显著的提升并优于其他最先进方法的结果. 如式 (13) 所示, AtlasNet 通过输入信息直接回归三维点云坐标值以计算倒角距离进行训练. 本文通过使用连续函数定义一个 3D 空间,并通过神经网络拟合的函数来描述这样的隐式形状, 不是直接回归重建物体中点云的三维坐标值, 即本文利用神经网络拟合映射函数来隐式的描述 3D 形状. 所以, 本文方法不能像 PSGN 和 AtlasNet 一样训练倒角距离. 在评估过程中, 本文通过网格生成步骤 (该步骤不可微分)来生成网格并与真实标签生成的真实网格中随机抽样 100k 个点计算倒角距离. 如表 3 所示, 本文在训练过程中没有像 PSGN 和 AtlasNet 一样训练倒角距离, 但本文在定量实验中也取得了较好的结果.

4.5 ShapeNet 数据集上的定性结果

本文与其他方法比较的定性结果, 如图 8 所示. 通过图 8 可以看到所有的方法都对物体的基本的几何特征进行了准确的提取. 本文发现 3D-R2N2[5] 在重建拓扑结构复杂的物体上出现了较为明显的空洞. PSGN[8] 可以产生高保真的输出, 但是缺乏连接性. 因此, PSGN 需要额外的有损后处理步骤来生成最终的网格. Pix2Mesh[9] 同样也在拓扑结构较为复杂的物体上出现了局部变形和空洞. AtlasNet[14]

表 1 本文的方法在 ShapeNet 数据集上与最先进方法的交并比 (IoU) 的定量比较

Table 1 The quantitative comparison of our method with the state-of-the-art methods for IoU on ShapeNet dataset

类别\方法	3D-R2N2	Pix2Mesh	AtlasNet	ONet	Our
Airplane	0.426	0.420	—	0.571	**0.592**
Bench	0.373	0.323	—	0.485	**0.503**
cabinet	0.667	0.664	—	0.733	**0.757**
Car	0.661	0.552	—	0.737	**0.755**
Chair	0.439	0.396	—	0.501	**0.542**
Display	0.440	0.490	—	0.471	**0.548**
Lamp	0.281	0.323	—	0.371	**0.409**
Loudspeaker	0.611	0.599	—	0.647	**0.672**
Rifle	0.375	0.402	—	0.474	**0.500**
Sofa	0.626	0.613	—	0.680	**0.701**
Table	0.420	0.395	—	0.506	**0.547**
Telephone	0.611	0.661	—	0.720	**0.763**
Vessel	0.482	0.397	—	0.530	**0.569**
Mean	0.493	0.480	—	0.571	**0.605**

表 2 本文的方法在 ShapeNet 数据集上与最先进方法法线一致性 (NC) 的定量比较

Table 2 The quantitative comparison of our method with the state-of-the-art methods for NC on ShapeNet dataset

类别\方法	3D-R2N2	Pix2Mesh	AtlasNet	ONet	Our
Airplane	0.629	0.759	0.836	0.840	**0.847**
Bench	0.678	0.732	0.779	0.813	**0.818**
Cabinet	0.782	0.834	0.850	0.879	**0.887**
Car	0.714	0.756	0.836	0.852	**0.855**
Chair	0.663	0.746	0.791	0.823	**0.835**
Display	0.720	0.830	0.858	0.854	**0.871**
Lamp	0.560	0.666	0.694	0.731	**0.751**
Loudspeaker	0.711	0.782	0.825	0.832	**0.845**
Rifle	0.670	0.718	0.725	0.766	**0.781**
Sofa	0.731	0.820	0.840	0.863	**0.872**
Table	0.732	0.784	0.832	0.858	**0.864**
Telephone	0.817	0.907	0.923	0.935	**0.938**
Vessel	0.629	0.699	0.756	0.794	**0.801**
Mean	0.695	0.772	0.811	0.834	**0.844**

已经可以重建出良好的表面, 但容易产生小面片之间的自交和重叠. ONet[17] 则很好地捕获了拓扑结构复杂的物体的特征并且生成了更加平滑的表面, 但缺失了部分局部细节. 本文的方法能够捕获复杂拓扑结构和生成高保真的三维形状输出. 另外, 从台灯和飞机的重建结果可以看出, 本文的方法有效地恢复了物体复杂拓扑结构的连接处和局部细节.

4.6 真实图片数据集上的定性结果

为了检验本文的方法对真实数据的泛化能力, 本文将本文方法应用于 Online Products dataset[36] 用于定性评价. 本文的模型并没有在 Online Products dataset 上进行训练.

如图 9 所示, 本文展示了本文方法在 Online Products dataset 定性结果. 本文在真实图片数据集中选择一些有代表性的图像来显示定性结果. 通过结果可以看出, 虽然本文的模型只是在合成数据上进行训练, 但本文的模型对真实图片数据也有很好的泛化能力.

4.7 模型与最先进方法的比较

本文分别从真实图片泛化能力和模型鲁棒性以及重建复杂拓扑结构和细节表达能力方面进一步进行分析.

表 3 本文的方法在 ShapeNet 数据集上与最先进方法倒角距离 (CD) 的定量比较

Table 3 The quantitative comparison of our method with the state-of-the-art methods for CD on ShapeNet dataset

类别\方法	3D-R2N2	Pix2Mesh	AtlasNet	ONet	Our
Airplane	0.227	0.187	**0.104**	0.147	0.130
Bench	0.194	0.201	**0.138**	0.155	0.149
Cabinet	0.217	0.196	0.175	0.167	**0.146**
Car	0.213	0.180	**0.141**	0.159	0.144
Chair	0.270	0.265	0.209	0.228	**0.200**
Display	0.314	0.239	**0.198**	0.278	0.220
Lamp	0.778	0.308	**0.305**	0.479	0.364
Loudspeaker	0.318	0.285	**0.245**	0.300	0.263
Rifle	0.183	0.164	**0.115**	0.141	0.130
Sofa	0.229	0.212	**0.177**	0.194	0.179
Table	0.239	0.218	0.190	0.189	**0.170**
Telephone	0.195	0.149	0.128	0.140	**0.121**
Vessel	0.238	0.212	**0.151**	0.218	0.189
Mean	0.278	0.216	**0.175**	0.215	0.185

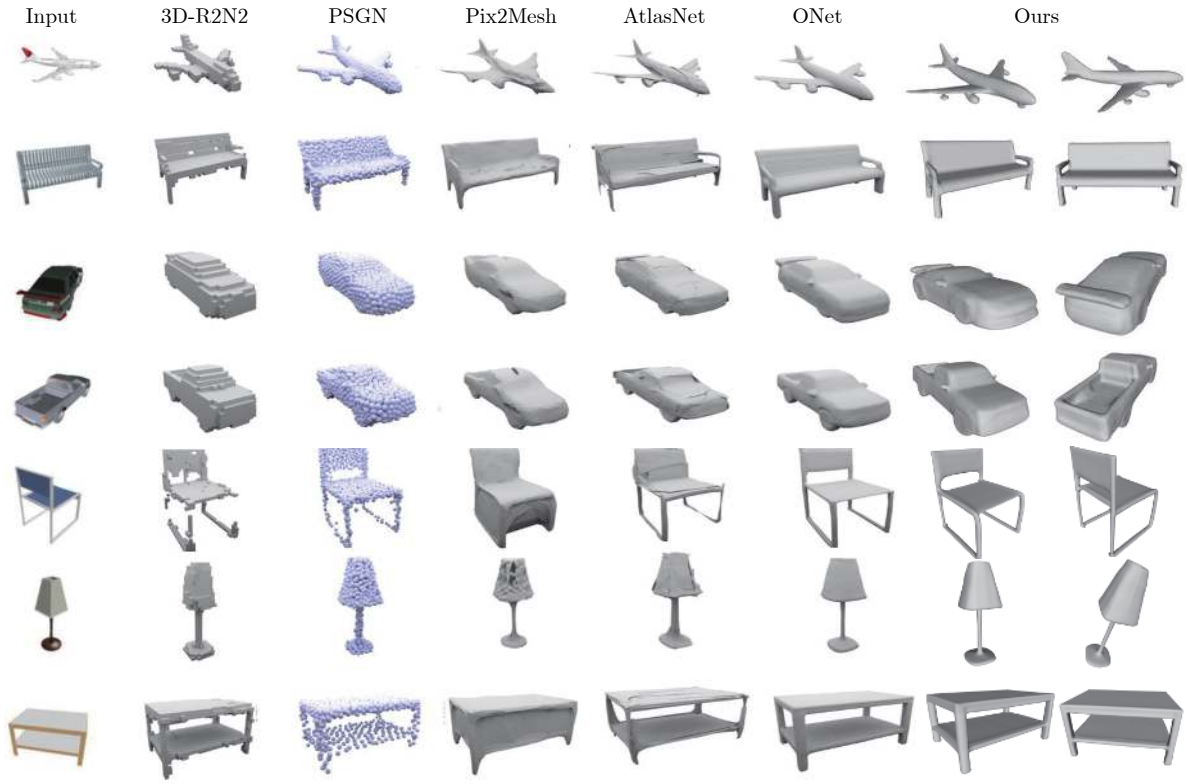


图 8 ShapeNet 数据集上的定性结果

Fig. 8 Qualitative results on the ShapeNet dataset

4.7.1 真实图片泛化能力和模型鲁棒性

为了测试本文模型对真实图片的泛化能力, 本文的模型在 Online Products dataset 进行了三维重建. 如图 2 和图 9 所示, 本文的模型只在 ShapeNet 数据集上进行了训练, 但是在真实世界的图片上还是获得了不错的效果. 通过这个实验验证了本文模型的域可转移性. 如表 1、2、3 所示, 本文的方法在 ShapeNet 数据集上所有的类别上都取得了最优的交并比和法线一致性, 并在倒角距离上也取得了不错的效果, 这证明了本文的方法在处理所有类别上

图 9　Online Products dataset 的定性结果

Fig. 9　Qualitative results on Online Products dataset

更加具有鲁棒性.

4.7.2　重建复杂拓扑结构和细节表达能力

对于拓扑结构复杂的物体 (例如: 飞机、椅子和桌子), 本文结合了输入图像的全局特征和模糊概率点周围的局部特征进一步强化了复杂拓扑结构和细节表达能力. 如图 2 和图 9 所示, 本文的模型在真实物体连接处和凹凸处具有更强的表达力.

4.8　消融实验

在这个部分, 本文对本文方法进行了消融实验, 主要研究了分支代偿网络模块和 DDPG 再推理模糊概率点的占用概率模块对模型整体性能的影响. 本文将分支代偿网络模块、DDPG 再推理模糊概率点的占用概率模块和完整模型分别用 MB、DR 和 FM 表示.

如表 4 所示, 本文首先验证了 DDPG 再推理模糊概率点的占用概率模块显著提高了 IoU 指标. 其次, 本文验证了分支代偿网络模块明显提高了 IoU、NC 和 CD 等指标, 另外因为分支代偿网络的加入可以生成更多样化的特征表示以更好地定位模糊概率点和提高模型的泛化能力, 所以在加入分支代偿网络后模型效果有了显著的性能提升. 最后, 如图 10 所示, 本文展示了消融实验中不同模型的定性实验结果.

4.9　MNGD 调整模糊概率点的定量结果

在这个部分, 本文的方法不依赖 IoU 等评价指标对本文方法进行定量评估, 而是展示了本文方法对模糊概率点调整能力的定量结果, 如图 11 所示. 圆点代表一张随机图片中的模糊概率点整体调整正确个数, 叉号代表一张随机图片中整体调整错误个数, 虚线代表整体调整正确或错误的决策边界.

表 4　消融实验

Table 4　Ablation study

模型\指标	IoU	NC	CD
FM w/o DR, MB	0.593	0.840	0.194
FM w/o MB	0.599	0.839	0.194
FM	**0.605**	**0.844**	**0.185**

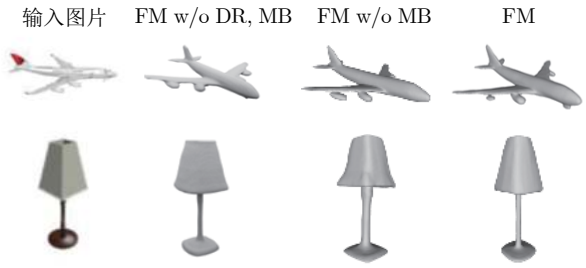


图 10　消融实验的定性结果

Fig. 10　Qualitative results of ablation study

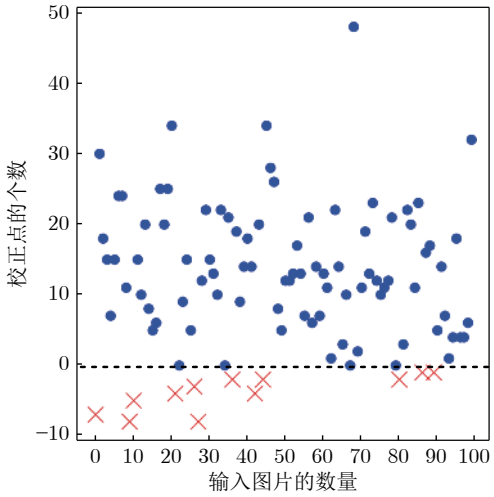


图 11　MNGD 随机调整 100 张图片中模糊概率点的结果

Fig. 11　The result of MNGD adjusting the fuzzy probability points in 100 random images

4.10　ShapeNet 中所有类别的定性实验

在这个部分, 本文展示了本文方法在训练数据集 ShapeNet 中所有类别的三维重建结果. 如图 12 所示, 本文的方法能够准确捕获 ShapeNet 数据集上所有类别物体的基本几何特征, 另外本文方法也有效的恢复了所有类别物体的复杂拓扑结构的连接

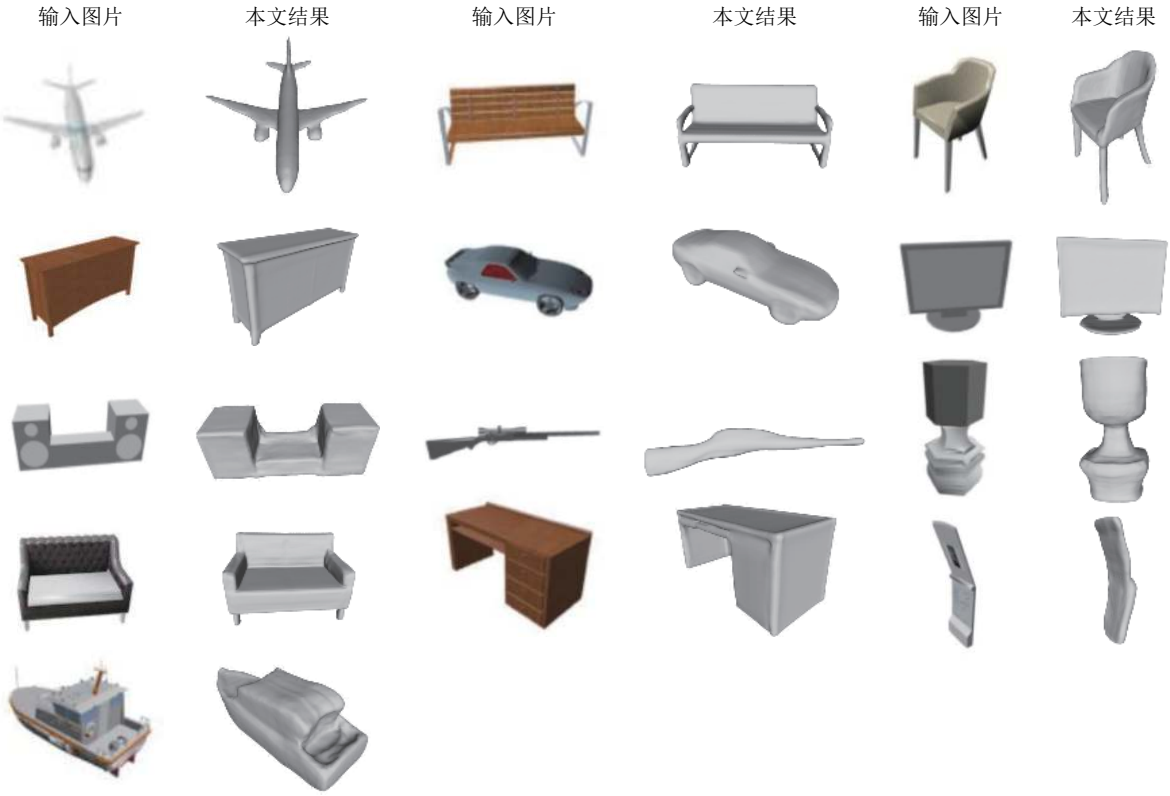


图 12 ShapeNet 上所有类别的定性结果

Fig. 12 Qualitative results on ShapeNet of all categories

处和局部细节.

4.11 进一步的工作

如图 13 所示 (实线圈表示本文方法与 ONet 的比较, 虚线圈表示本文方法与原图的比较). 相比与之前的工作, 本文的方法在大多数情况下能够生成高保真的三维模型, 但还需要在未来工作中对以下方面进行进一步优化: 1) 如图 13 所示,由于本文没有引入额外的相机参数信息来定位点到图像平面的映射使得本文模型不能非常准确地恢复细小的凹凸结构 (图 13(a) 和图 13(b)) 和薄片结构 (图 13(c)). 2) 如图 13 所示, 单视图三维重建本质上缺少多视角观测数据导致无法确定唯一的真实形状, 所以学习模型经常趋于逼近光滑表面但是细节模糊的平均形状 (图 13(a), 图 13(b) 和图 13(c)). 3) 全标注三维形状公共数据集较少,如何增广或利用无标注三维形状数据集来提高模型泛化能力也十分重要.

在未来的工作中, 我们计划引入相机参数对模糊概率点对应的特征图进行更进一步的精确定位以及使用自监督[43−44] 来有效利用无标注的三维形状数据集, 从而使深度强化学习算法训练的智能体做出更加合理的再推理.

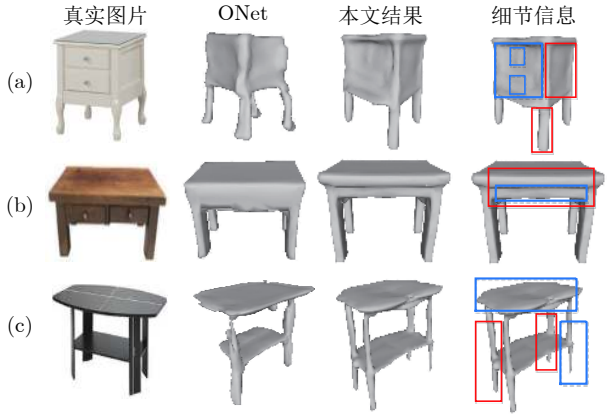


图 13 单视图三维重建中具有挑战性案例

Fig. 13 Challenging cases in single-view 3D reconstruction

5 结论

本文提出一个新的方法来实现对三维重建中模糊概率点进行再推理从而实现具有高保真和丰富细节的单视图三维重建. 本文的方法首先通过动态分支代偿网络从输入信息中捕捉到更多样化的特征信息以提高模型的泛化能力. 其次, 本文通过注意力

引导的聚合机制聚合了模糊概率点的局部信息和全局图像信息, 之后通过深度强化学习算法 DDPG 训练的智能体对模糊概率点进行再推理并给出相应的结果. 最后, 本文通过大量的定性和定量实验表明, 本文的方法在一定程度上解决了具有复杂拓扑的物体以及一些表面细节信息难以准确重建的问题.

References

1 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
(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)

2 Zheng Tai-Xiong, Huang Shuai, Li Yong-Fu, Feng Ming-Chi. Key techniques for vision based 3D reconstruction: A review. *Acta Automatica Sinica*, 2020, **46**(4): 631−652
(郑太雄, 黄帅, 李永福, 冯明驰. 基于视觉的三维重建关键技术研究综述. 自动化学报, 2020, **46**(4): 631−652)

3 Xue Jun-Shi, Yi Hui, Wu Zhi-Huan, Chen Xiang-Ning. A hybrid multi-view 3D reconstruction method based on scene graph partition. *Acta Automatica Sinica*, 2020, **46**(4): 782−795
(薛俊诗, 易辉, 吴止锾, 陈向宁. 一种基于场景图分割的混合式多视图三维重建方法. 自动化学报, 2020, **46**(4): 782−795)

4 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 International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates, Inc., 2016. 82−90

5 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

6 Yao Y, Luo Z X, Li S W, Fang T, Quan L. MVSNet: Depth inference for unstructured multi-view stereo. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 785−801

7 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. Long Beach, USA: Curran Associates, Inc., 2017. 540−550

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

9 Wang N Y, Zhang Y D, Li Z W, Fu Y W, Liu W, Jiang Y G. Pixel2Mesh: Generating 3D mesh models from single RGB images. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 55−71

10 Scarselli F, Gori M, Tsoi A C, Hagenbuchner M, Monfardini G. The graph neural network model. *IEEE Transactions on Neural Networks*, 2009, **20**(1): 61−80

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

12 Roth S, Richter S R. Matryoshka networks: Predicting 3D geometry via nested shape layers. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 1936−1944

13 Wu J J, Zhang C K, Zhang X M, Zhang Z T, Freeman W T, Tenenbaum J B. Learning shape priors for single-view 3D completion and reconstruction. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 673−691

14 Groueix T, Fisher M, Kim V G, Russell B C, Aubry M. A Papier-Mache approach to learning 3D surface generation. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 216−224

15 Kanazawa A, Black M J, Jacobs D W, Malik J. End-to-end recovery of human shape and pose. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 7122−7131

16 Kong C, Lin C H, Lucey S. Using locally corresponding CAD models for dense 3D reconstructions from a single image. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 5603−5611

17 Mescheder L, Oechsle M, Niemeyer M, Nowozin S, Geiger A. Occupancy networks: Learning 3D reconstruction in function space. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 4455−4465

18 Lillicrap T P, Hunt J J, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning [Online], available: https://arxiv.org/abs/1509.02971, July 5, 2019

19 Li D, Chen Q F. Dynamic hierarchical mimicking towards consistent optimization objectives. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 7639−7648

20 Chang A X, Funkhouser T, Guibas L, et al.Shapenet:An information-rich 3d model repository [Online], available: https://arxiv.org/abs/1512.03012, December 9, 2015

21 Durou J D, Falcone M, Sagona M. Numerical methods for shape-from-shading: A new survey with benchmarks. *Computer Vision and Image Understanding*, 2008, **109**(1): 22−43

22 Zhang R, Tsai P S, Cryer J E, Shah M. Shape-from-shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1999, **21**(8): 690−706

23 Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014. 2672−2680

24 Kingma D P, Welling M. Auto-encoding variational bayes [Online], available: https://arxiv.org/abs/1312.6114, May 1, 2014

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

26 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

27 Wang W Y, Ceylan D, Mech R, Neumann U. 3DN: 3D deformation network. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 1038−1046

28 Bernardini F, Mittleman J, Rushmeier H, Silva C, Taubin G. The ball-pivoting algorithm for surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 1999, **5**(4): 349−359

29 Kazhdan M, Hoppe H. Screened poisson surface reconstruction. *ACM Transactions on Graphics*, 2013, **32**(3): 29

30 Calakli F, Taubin G. SSD: Smooth signed distance surface reconstruction. *Computer Graphics Forum*, 2011, **30**(7): 1993−2002

31 Chen Z Q, Zhang H. Learning implicit fields for generative shape modeling. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 5932−5941

32 Wang W Y, Xu Q G, Ceylan D, Mech R, Neumann U. DISN:

Deep implicit surface network for high-quality single-view 3D reconstruction. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc., 2019. Article No. 45

33 Wang Q L, Wu B G, Zhu P F, Li P H, Zuo W M, Hu Q H. ECA-Net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 11531−11539

34 Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 618−626

35 Garland M, Heckbert P S. Simplifying surfaces with color and texture using quadric error metrics. In: Proceedings of the 1998 Visualization′ 98 (Cat. No.98CB362-76). Research Triangle Park, USA: IEEE, 1998. 263−269

36 Lorensen W E, Cline H E. Marching cubes: A high resolution 3D surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, 1987, **21**(4): 163−169

37 Drucker H, Le Cun Y. Improving generalization performance using double backpropagation. *IEEE Transactions on Neural Networks*, 1992, **3**(6): 991−997

38 Oh Song H, 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 (CVPR). Las Vegas, USA: IEEE, 2016. 4004−4012

39 Stutz D, Geiger A. Learning 3D shape completion from laser scan data with weak supervision. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 1955−1964

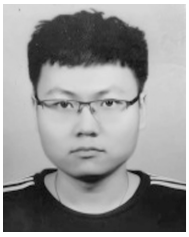
40 de Vries H, Strub F, Mary J, Larochelle H, Pietquin O, Courville A C. Modulating early visual processing by language. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates, Inc., 2017. 6597−6607

41 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 770−778

42 Kingma D P, Ba J. Adam: A method for stochastic optimization [Online], available: https: //arxiv. org/abs/1412. 6980, January 30, 2017

43 Zhu C C, Liu H, Yu Z H, Sun, X H. Towards Omni-supervised face alignment for large scale unlabeled videos. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020. 13090−13097

44 Zhu C C, Li X Q, Li J D, Ding G T, Tong W Q. Spatial-temporal knowledge integration: Robust self-supervised facial landmark tracking. In: Proceedings of the 28th ACM International Conference on Multimedia. Lisboa, Portugal: ACM, 2020. 4135−4143



李 雷 宁夏大学信息工程学院硕士研究生. 主要研究方向为三维物体重建, 人脸重建以及关键点对齐, 图像处理和计算机视觉与模式识别.
E-mail: lliicnxu@163.com
(**LI Lei** Master student at the School of Information Engineering, Ningxia University. His research interest covers 3D object reconstruction, face reconstruction and landmark alignment, image processing, computer vision and pattern recognition.)

徐 浩 宁夏大学信息工程学院硕士研究生. 主要研究方向为计算机视觉和三维人体姿态估计.
E-mail: hao_xu321@163.com
(**XU Hao** Master student at the School of Information Engineering, Ningxia University. His research interest covers computer vision and 3D human pose estimation.)

吴素萍 宁夏大学信息工程学院教授. 主要研究方向为三维重建, 计算机视觉, 模式识别, 并行分布处理与大数据. 本文通信作者.
E-mail: pswuu@nxu.edu.cn
(**WU Su-Ping** Professor at the School of Information Engineerring, Ningxia University. Her research interest covers 3D reconstruction, computer vision, pattern recognition, parallel distributed processing and big data. Corresponding author of this paper.)