计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(5):76−84 [doi: 10.15888/j.cnki.csa.009500]

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

# 杂乱场景中多尺度注意力特征融合抓取检测网络①

徐　衍[1,2], 林云汉[1,2,3], 闵华松[3]

[1](武汉科技大学 计算机科学与技术学院, 武汉 430081)
[2](武汉科技大学 智能信息处理与实时工业系统湖北省重点实验室, 武汉 430081)
[3](武汉科技大学 机器人与智能系统研究院, 武汉 430081)
通信作者: 林云汉, E-mail: yhlin@wust.edu.cn

**摘　要**: GSNet 使用抓取度区分杂乱场景的可抓取区域, 显著地提高了杂乱场景中机器人抓取位姿检测准确性, 但是 GSNet 仅使用一个固定大小的圆柱体来确定抓取位姿参数, 而忽略了不同大小尺度的特征对抓取位姿估计的影响. 针对这一问题, 本文提出了一个多尺度圆柱体注意力特征融合模块 (Ms-CAFF), 包含注意力融合模块和门控单元两个核心模块, 替代了 GSNet 中原始的特征提取方法, 使用注意力机制有效地融合 4 个不同大小圆柱体空间内部的几何特征, 从而增强了网络对不同尺度几何特征的感知能力. 在大规模杂乱场景抓取位姿检测数据集 GraspNet-1Billion 的实验结果表明, 在引入模块后将网络生成抓取位姿的精度最多提高了 10.30% 和 6.65%. 同时本文将网络应用于实际实验, 验证了方法在真实场景当中的有效性.

**关键词**: 点云; 机器人抓取位姿检测; 多尺度特征融合; 杂乱场景; 注意力机制

引用格式: 徐衍,林云汉,闵华松.杂乱场景中多尺度注意力特征融合抓取检测网络.计算机系统应用,2024,33(5):76–84. http://www.c-s-a.org.cn/1003-3254/9500.html

## Grasping Detection Network of Multi-scale Attention Feature Fusion in Cluttered Scenes

XU Yan[1,2], LIN Yun-Han[1,2,3], MIN Hua-Song[3]

[1](School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430081, China)
[2](Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan 430081, China)
[3](Institute of Robotics and Intelligent Systems (IRIS), Wuhan University of Science and Technology, Wuhan 430081, China)

**Abstract**: GSNet relies on graspness to distinguish graspable areas in cluttered scenes, which significantly improves the accuracy of robot grasping pose detection in cluttered scenes. However, GSNet only uses a fixed-size cylinder to determine the grasping pose parameters and ignores the influence of features of different sizes on grasping pose estimation. To address this problem, this study proposes a multi-scale cylinder attention feature fusion module (Ms-CAFF), which contains two core modules: the attention fusion module and the gating unit. It replaces the original feature extraction method in GSNet and uses an attention mechanism to effectively integrate the geometric features inside the four cylinders of different sizes, thereby enhancing the network's ability to perceive geometric features at different scales. The experimental results on GraspNet-1Billion, a grabbing pose detection dataset for large-scale cluttered scenes, show that after the introduction of the modules, the accuracy of the network's grasping poses is increased by up to 10.30% and 6.65%. At the same time, this study applies the network to actual experiments to verify the effectiveness of the method in real scenes.

**Key words**: point cloud; robot grasping pose detection; multi-scale feature fusion; cluttered scene; attention mechanism

① 基金项目: 国家重点研发计划 (2022YFB4700400); 国家自然科学基金 (62073249)
收稿时间: 2023-11-13; 修改时间: 2023-12-11; 采用时间: 2024-01-12; csa 在线出版时间: 2024-04-01
CNKI 网络首发时间: 2024-04-03

## 1 介绍

对物体进行抓取是机器人领域的关键技术, 在工业、仓储物流和医疗等领域有着广泛的应用. 在给定视觉输入的情况下检测潜在的抓取位姿是目前机器人视觉领域的一个关键问题. 然而, 由于复杂的环境造成的不确定性, 以及难以提取物体几何形状, 为机器人生成稳定的抓取位姿依然有很大的难度. 提高机器人抓取通用性、准确性和效率是该领域研究人员的长期追求. 关于抓取姿势检测的方法可以分为基于模型的方法和无模型的方法.

基于模型的方法利用物理分析来找到合适的抓取位姿[1,2], 这类方法需要准确的对象 3D 模型, 使用物理分析工具生成抓取姿态, 然后将预先生成的抓取位姿存储到数据库中, 使用时根据物体数据在数据库中进行查找. 然而这类方法对未知物体的泛化能力较差, 限制了在实际场景中的应用, 所以近年来相关研究都集中于无模型的方法.

无模型的方法主要基于深度学习, 对未知物体有着良好的泛化能力, 主要包括抓取位姿采样方法和直接回归抓取位姿方法. 抓取位姿采样方法使用采样算法生成一个或多个抓取样本, 同时学习一个函数来估计抓取样本的质量. 抓取位姿采样方法需要计算每个样本的得分, 并通过样本的编码信息做出抓取决策. PointNetGPD[3]先对被抓取物体潜在的候选抓取位姿进行采样, 然后使用基于点的神经网络对其进行评分. Mousavian 等人[4]使用了抓取得分相对于抓取位姿的导数, 进行梯度上升将潜在的不良样本细化为高质量抓取. 但是抓取位姿采样方法会生成大量无效抓取, 同时计算量较大, 无法适用于多物体的杂乱场景, 因此后续研究多使用端到端的网络直接从图像中回归抓取位姿. 早期的端到端方法[5,6]受到目标检测相关研究的影响, 在被抓取对象上生成边界框作为抓取位姿, 但是这类方法的抓取位姿自由度低, 限制了抓取的精度, 后续的研究[7−9]使用神经网络直接为机器人生成 6-DoF 抓取位姿. Ni 等人[10]使用 PointNet++[11]作为主干网络, 为每个点预测类别、分数和抓取位姿, Li 等人[12]提出了一个三分支的网络, 使用 3 个解码器分别进行实例分割、抓取位姿计算和置信度检测, 提高了在物体遮挡碰撞场景的抓取成功率. Gou 等人[13]使用 RGB 图像生成的热图来确定抓取的接近向量, 之后再结合深度图像计算夹爪的开口宽度和到抓取点的距离.

针对杂乱场景点云数据量大难以区分可抓取区域的问题, Wang 等人提出一个两阶段的网络 GSNet[14], 其中利用一种基于几何线索的指标抓取度 (graspness) 用以区分场景中的可抓取点, 显著提高了杂乱场景的网络的推理速度, 在大规模抓取位姿检测数据集 GraspNet-1Billion[15]上取得了现有方法中最好的结果, 将准确度提升了约 30%. 但是 GSNet 在确定抓取位姿参数时, 仅使用了固定大小的圆柱体区域内的特征, 缺乏对多尺度几何特征的提取能力, 会使网络忽略一些尺度更小的几何形状, 而在进行实际的精细抓取操作中, 物体微小的几何细节会对抓取成功率有较大的影响. 最近 Ma 等人提出了一个实现了多尺度平衡的 6 自由度抓取位姿生成网络[16], 提出了多尺度圆柱体分组 (multi-scale cylinder grouping, MsCG) 模块、尺度平衡学习 (scale balanced learning, SBL) 损失和一种对象平衡采样 (object balanced sampling, OBS) 策略, 解决了尺度不平衡的情况下小尺度样本的抓取检测问题. 其中多尺度圆柱体分组模块 (MsCG) 模块提取了抓取点位置的不同尺度的局部几何特征, 提高了网络对几何特征的感知能力. 但是物体不同尺度的几何特征对抓取结果会有不同的影响权重, 而多尺度圆柱体分组模块 (MsCG) 仅对不同尺度的几何特征进行了简单的拼接, 没有进行有效的特征融合.

针对 GSNet 忽略了小尺度形状特征以及多尺度特征难以融合的问题, 本文提出了一个新的多尺度圆柱体注意力特征融合模块 (Ms-CAFF), 使用圆柱体特征提取模块作为隐式的位置编码, 同时使用注意力机制对 4 个不同尺度的特征进行聚合. 同时将模块引入 GSNet, 优化了网络对局部几何特征的感知能力. 实验证明, 本文的方法显著提高了网络性能, 通过使用注意力机制聚合多尺度局部几何信息, 提高了抓取位姿计算的准确度, 改进后的网络在大规模抓取位姿检测数据集 GraspNet-1Billion 上的结果优于其他方法.

本文的主要贡献如下.

(1) 针对多尺度特征难以融合的问题, 提出了一个多尺度圆柱体注意力特征融合模块 (Ms-CAFF), 包含注意力特征融合模块以及重新设计的门控单元, 使用注意力机制融合了抓取点附近 4 个不同大小圆柱体空间内的物体形状特征.

(2) 基于多尺度圆柱体注意力特征融合模块, 提出了一个基于注意力机制的多尺度特征融合抓取位姿检测网络, 将多尺度圆柱体注意力特征融合模块加入

GSNet 中, 显著提高了网络抓取位姿检测的性能.

(3) 在大规模抓取位姿检测数据集 GraspNet-1Billion 上与其他现有方法对比了抓取位姿检测的准确度, 验证了本文方法对多尺度几何特征的提取能力, 同时通过设计实际机器人抓取实验验证了提出算法在真实场景当中的有效性.

## 2 方法

GSNet 提取物体几何形状的过程中, 仅使用了一个固定大小的圆柱体区域进行特征提取, 这种方法使网络更倾向于范围内较大尺度的特征, 而忽略一些尺度更小的几何形状. 在 2D 图像领域卷积神经网络中感受野 (RF) 表示网络中神经元能感受到的图像范围, 感受野越小, 网络提取的特征则趋向于局部和细节, 而感受野越大, 则会使网络更倾向于更为全局、语义层次更高的特征, 需要合理的设置感受野的大小以适用于不同的任务. 同样的, 在点云的相关研究中也有类似的概念, PointNet++使用了多尺度分组 MSG (multi-scale grouping), 对不同半径的子区域进行特征提取后进行特征堆叠, 有效的捕获了点云的局部特征.

以上分析表明, 合理的聚合不同尺度大小的信息能够提高网络的效果, 而捕获多尺度模式的一种简单但有效的方法就是应用具有不同尺度的分组层, 并使用有效的方法将不同尺度的特征进行融合. 对于机器人抓取检测, 物体被抓取位置不同大小的形状都对最终的抓取检测有着重要的影响, 本文所提出的多尺度注意力特征融合 (Ms-CAFF) 模块使用注意力机制区分了 4 个不同半径圆柱体内部形状特征对机器人抓取的重要程度.

本节将说明修改之后的整体网络结构及各个模块的功能. 之后将介绍多尺度注意力特征融合模块的原理和细节.

### 2.1 网络结构

本文的网络结构如图 1 所示, 网络分为两段结构, 上半部分用于计算抓取度, 从而区分场景中的可抓取区域. 下半部分使用点和视图特征确定最后的抓取位姿.

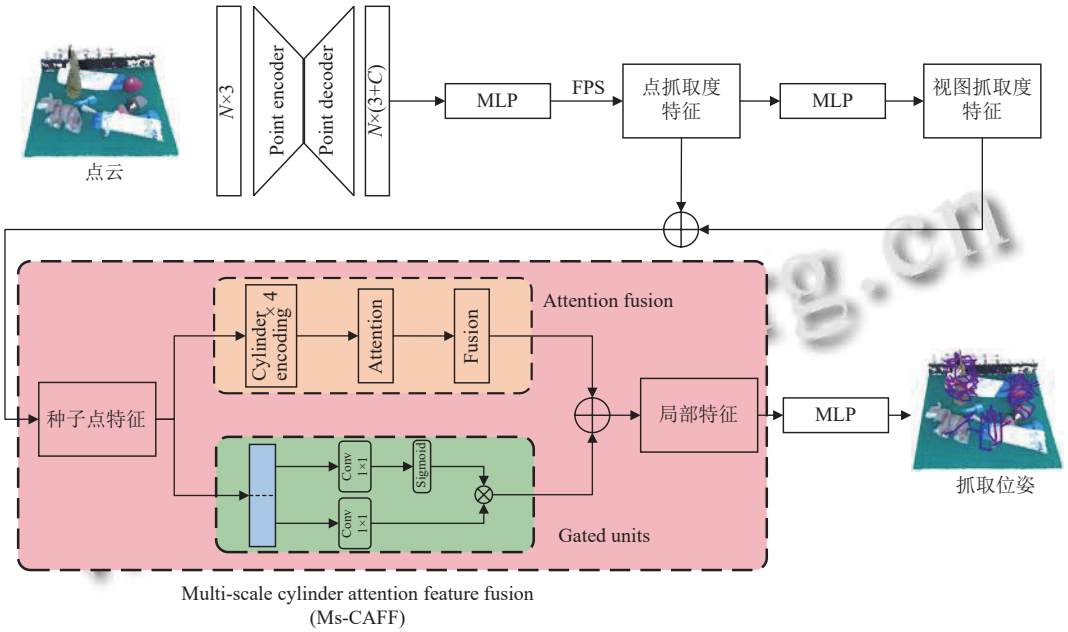


图 1 本文提出的多尺度特征融合抓取位姿检测网络结构图

主干网络采用 MinkowskiEngine, 利用 3D 稀疏卷积提取点云特征, 使用点集 $P=\{p_i \mid i=1,\cdots,N\}$ 作为网络的输入, 提取 $C$ 维度的点特征, 输出 $N\times(3+C)$ 的点特征用于后续处理, 通过多层感知器 (MLP) 计算每个点的抓取度得分:

$$S^P=\{s_i^p \mid s_i^p \subset [0,1], i=1,\cdots,N\} \tag{1}$$

其中, $S^P$ 包含 2 维对象分类分数以及 1 维抓取度分数.

对象分类分数用于区分场景中的可抓取对象, 抓取度分数则用于筛选得分大于 0.1 的点, 接着使用最远点采样算法 (FPS) 采样 $M$ 个点用于后续的抓取位姿计算.

对于抓取位姿的计算方法, 由于使用深度学习方法难以直接回归旋转矩阵, GSNet 将抓取位姿解耦为

接近方向和平面内旋转角度. 使用 Fibonacci 格子从单位球体中采样 $N_V$ 个接近方向 $V=\{v_j \mid j=1,\cdots,N_V\}$. 使用点得分作为输入通过多层感知器得到视图抓取度得分.

$$S^v=\{s_i^v \mid s_i^v \subset [0,1], i=1,\cdots,N\} \tag{2}$$

输入为 $M$ 个种子点的特征, 输出 $M\times C$ 的残差特征用于后续网络的抓取位姿计算以及 $M\times V$ 的视图抓取度得分, 选择得分高的视图作为最终抓取位姿的接近向量.

多尺度圆柱体注意力特征融合模块以 $M$ 个种子点坐标及接近向量确定圆柱体中心的位置和偏转方向. 使用主干网络提取的特征以及点云坐标作为输入, 聚合抓取点位置 4 个不同大小的圆柱体空间内的聚合点级和视图级特征. 最后将得到的局部融合特征通过多层感知器预测夹爪的旋转角度和接近距离, 确定最终的抓取位姿. 下面本文将详细介绍多尺度注意力特征融合 (Ms-CAFF) 模块.

## 2.2 多尺度注意力特征融合模块

为了解决多尺度空间内物体形状特征难以融合的问题的本文提出了多尺度注意力特征融合模块. 模块分为两个部分: 注意力特征融合模块和门控单元. 注意力特征融合模块使用圆柱体编码操作, 用于完成对不同大小圆柱体空间内的特征进行隐式位置编码任务; 之后使用注意力机制对编码后的特征进行有效的融合, 实现小尺度局部形状和大尺度形状的融合. 门控单元则是对模块的输入特征做控制筛选, 从而进一步区分不同尺度的形状特征中对抓取成功率影响较大的部分.

### 2.2.1 注意力特征融合

注意力特征融合模块的结构如图 2 标注所示, 主要包含圆柱体编码和注意力特征融合. $C_p\in\mathbb{R}^{M\times C}$ 为主干网络提取的点特征, 通过多层感知器将点特征映射为视图特征 $C_v\in\mathbb{R}^{M\times C}$, 之后将两者求和得到特征 $C\in\mathbb{R}^{M\times C}$, 如式 (3) 和式 (4) 所示.

$$C_v=\mathrm{MLP}(C_p) \tag{3}$$

$$C=C_p+C_v \tag{4}$$

在一个确定的抓取位置上, 平行夹持器的有效操作空间为夹爪所能包含的空间大小, 即如图 3 所示的以夹持器宽度为底面直径、夹持器高度为高度的圆柱体. 此圆柱体空间内的形状特征直接决定了抓取物体的成功率. 但是自注意力机制是一种全局操作, 对位置信息是不敏感的, 而点云形状特征并不直接包含相关的位置信息, 所以本文参考卷积隐式编码位置信息的方法, 基于 PointNet++中的 Set Abstraction 操作代替卷积操作构建位置编码, Set Abstraction 通过 Sampling 和 Grouping 逐级的降采样, 有效地捕获了点云不同规模不同层次的局部邻域几何结构. 之后再通过一个一维卷积层隐式编码圆柱体空间内部的形状特征, 同时为了平衡计算量和精度, 使用了 0.25、0.5、0.75 和 1 倍夹持器宽度的圆柱体来提取并编码抓取点附近的物体形状特征.

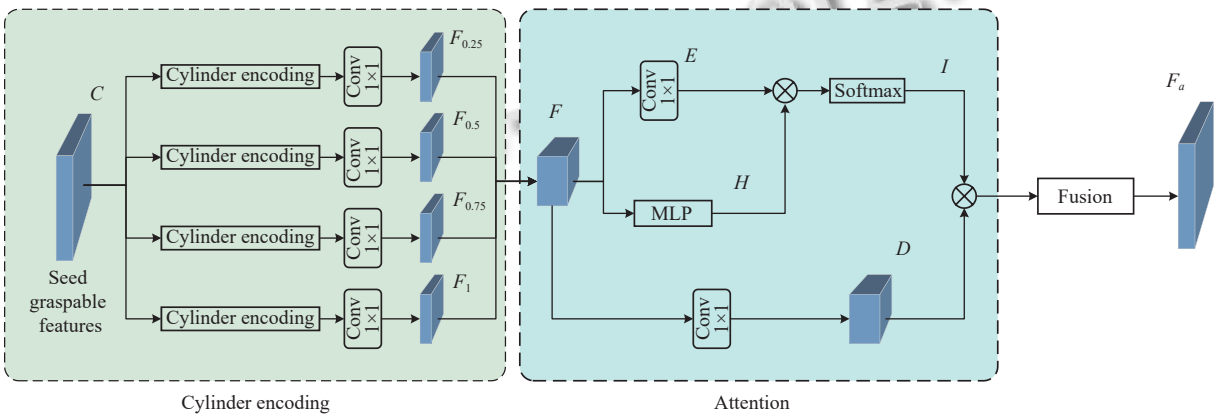


图 2 注意力特征融合模块

如图 3 所示, $F_i\in\mathbb{R}^{C\times M}$ 为圆柱体编码模块所编码的局部几何特征, $P\in\mathbb{R}^{N\times 3}$ 为点云坐标, $CE$ 为 (cylinder encoding) 圆柱体特征编码, 基于上面提到的 Set Abstraction 和一维卷积隐式构建位置编码, $i$ 表示圆柱体的半径长度, 分别编码 0.25、0.5、0.75 和 1 倍夹持器宽度的圆柱体内部的特征. 使用了残差特征 $C$ 以及点云坐标 $P$ 作为输入数据, 如式 (5) 所示.

$$F_i=CE(C,P),\quad i=0.25,0.5,0.75,1 \tag{5}$$

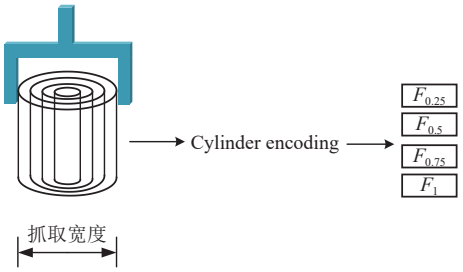


图 3 圆柱体编码示意图

机器人抓取需要针对不同的物体形状设定不同的抓取参数, 关键在于有效的提取物体的特征, 使用更大尺度的圆柱体会关注更大规模的形状信息, 但会丢失细节特征, 而小尺度的圆柱体空间会包含更为小的局部几何信息, 有利于对抓取位姿进行精细的调整. 对于大体积物体, 大尺度形状信息能够有助于找到合适的抓取位置, 而小尺度形状信息则对拥有复杂的表面形状或者局部形状有较大变化的物体抓取.

有效地对不同尺度形状信息进行平衡和对机器人抓取成功率有着重要的影响, 因此本文提出了多尺度注意力特征融合方法, 通过融合不同尺度的特征, 解决不同尺度物体形状特征的融合问题.

$$F = concat(F_{0.25}, F_{0.5}, F_{0.75}, F_1) \tag{6}$$

本文采用原始 Transformer 中所引入的自注意力机制, 首先将圆柱体编码模块所得到的 4 个位置编码和特征信息$F_i$进行连接操作, 得到注意力特征融合模块的输入$F$. 将$F$经过一维卷积后得到$D, E$, 同时让$F$经过全连接层得到特征映射$H$, 再将其转置并与$E$进行矩阵运算得到矩阵相似度权重, 再通过 *Softmax* 激活函数对权重进行归一化, 转换为注意力特征图$I$, 通过计算形状特征的相似性以区分不同尺度的物体形状的重要性. 其中$I$代表特征图中两个对应位置之间的相似度关系, 其中为了平衡计算效率将$d_k$设置为 32. 最后将特征映射$D$与$I$进行加权求和, 得到注意力权重的值向量的加权和$F_a$.

$$I = Softmax\left(\frac{E \cdot H^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{7}$$

$$F_a = I \cdot D \tag{8}$$

2.2.2 门控单元

在得到注意力特征$F_a$后, 需要将注意力特征附加到种子点特征之上, 得到增强的多尺度形状语义信息. 由于点云数据量较大, 如果直接使用全部的数据作为门控信息会增加计算量, 同时为了进一步过滤不重要的信息, 本文在参考 GLU 单元后设计了一个新的门控融合单元, 如图 4 所示.

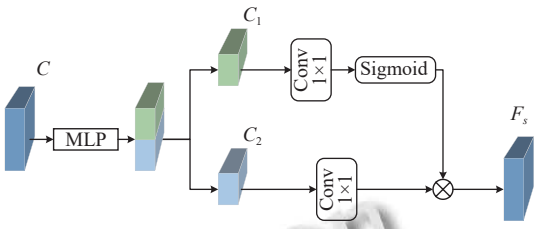


图 4 门控单元

首先通过全连接层对种子点特征$C$进行维度转换, 之后将$C$输入门控单元, $W$和$V$是不同的一维卷积核, $b$和$c$是偏置参数, 特征$C$被拆分为$C_1$和$C_2$, $C_1$经过参数$W$和$b$的卷积处理后, $C_2$由参数为$V$和$c$的卷积处理并通过 Sigmoid 激活函数的输出门得到门控信息用来控制最终的输出结果, 将两者进行矩阵逐元素乘积得到门控单元输出的特征$F_s$. 式 (9) 中*表示卷积运算.

$$F_s = (C_1 * W + b) \otimes \sigma(C_2 * V + c) \tag{9}$$

最后将经过门控单元过滤筛选后的全局特征与注意力特征融合模块输出的特征相加, 得到注意力特征融合 (Ms-CAFF) 模块的输出$F_{\mathrm{CAFF}}$.

$$F_{\mathrm{CAFF}} = F_a + F_s \tag{10}$$

整个多尺度注意力特征融合模块的输出如式 (11) 所示, *GU* (gated units) 代表门控融合操作.

$$F_{\mathrm{CAFF}} = Attention(CE(C)) + GU(C) \tag{11}$$

### 2.3 损失函数

本对于网络训练过程中的损失函数, 为防止梯度爆炸并加快训练速度, 本文采用 Smooth L1 平滑的平均绝对值误差函数. $x_n$和$y_n$分别代表预测值和标签值, $L_s$和$L_w$分别代表模块预测的抓取分数和抓取宽度的损失, 其中抓取得分损失$L_s$会将预测的所有抓取位姿的得分损失求平均值. 最后将两者以系数$\alpha$和$\beta$相加得到整个模块的损失, 在实验中分别设置为 10 和 15.

$$L_{w/s} = \begin{cases} 0.5(x_n, y_n)^2, & |x_n - y_n| < 1 \\ |x_n - y_n| - 0.5, & \text{otherwise} \end{cases} \tag{12}$$

$$L_{\text{Ms-CAFF}} = \alpha L_w + \beta L_s \tag{13}$$

## 3 实验分析

为了验证本文方法的有效性, 与其他主流抓取位

姿检测方法对比了公开数据集 GraspNet-1Billion 上的结果. 同时对网络的训练过程进行了可视化, 与使用 MsCG 模块替换圆柱体分组模块 (cylinder grouping) 的 GSNet 网络对比了网络损失和收敛速度. 为了验证修改后的网络对具有复杂几何形状物体生成抓取质量的改进情况, 进行了抓取结果可视化分析. 同时通过对比不同模块参数量和推理时间评估方法的复杂度和计算速度. 最后设计机器人抓取平台的实际实验, 在真实场景中验证了本文方法生成抓取位姿的有效性.

### 3.1 数据集实验

● 数据集. 大规模杂乱场景抓取位姿检测数据集 GraspNet-1Billion, 包含 Realsense/Kinect 相机在 190 多个杂乱场景的不同视角拍摄的 97 280 张 RGBD 图像, 通过力封闭指标计算进行抓取位姿标注与评估. 本文的对比的基准方法为 GSNet, 测试场景根据对象类别被分为 3 类: 已见过的物体 (Seen)、未见过但相似的物体 (Similar) 以及未见过的新物体 (Novel), 以对比不同方法的泛化能力.

● 评价指标. 使用 *Precision@k* 作为评估指标, 衡量排名前 $k$ 的抓取的精准度. *TP* (true positive) 代表成功抓取, *FP* (false positive) 代表失败抓取, $AP_\mu$ 表示给定摩擦系数 $\mu$ 时 $k$ 从 1–$N$ 的平均 *Precision@k*, $N$ 设定为 50. *AP* 由 $AP_\mu$ 的均值计算得到, $\mu$ 的范围为 0.2–1.2, 以 0.2 为间隔.

$$Precision@k = \frac{TP@k}{TP@k + FP@k} \tag{14}$$

$$AP_\mu = \frac{1}{N}\sum_{k=1}^{N} Precision@k \tag{15}$$

● 实验参数设定. 模型采用 PyTorch 框架进行训练以及测试, 操作系统为 Ubuntu 18.04, 在 RTX3080Ti 上训练了 10 轮, 批大小 batch size 设置为 2, 每个批次的学习率设置为 0.000 5, 同时使用 Adam 优化器优化训练过程, 夹持器宽度为 $r$=0.05 m, 圆柱体编码模块半径分别设置为 0.25、0.5、0.75 和 1 倍的夹持器宽度, 高度范围为[−0.02 m, 0.04 m].

实验结果如表 1 所示.

表 1 在 GraspNet-1Billion (RealSense/Kinect) 上与其他方法的结果对比, CD 为碰撞检测 (%)

| 方法 | Seen | | | Similar | | | Novel | | |
|---|---|---|---|---|---|---|---|---|---|
| | *AP* | $AP_{0.8}$ | $AP_{0.4}$ | *AP* | $AP_{0.8}$ | $AP_{0.4}$ | *AP* | $AP_{0.8}$ | $AP_{0.4}$ |
| GG-CNN[17] | 15.48/16.89 | 21.84/22.47 | 10.25/11.23 | 13.26/15.05 | 18.37/19.76 | 4.26/6.19 | 5.52/7.38 | 5.93/8.78 | 1.86/1.32 |
| Chu等人[18] | 15.97/17.59 | 23.66/24.67 | 10.80/12.74 | 15.41/17.36 | 20.21/21.64 | 7.06/8.86 | 7.64/8.04 | 8.69/9.34 | 2.52/1.76 |
| GPD[19] | 22.87/24.38 | 28.53/30.16 | 12.84/13.46 | 21.33/23.18 | 27.83/28.64 | 9.64/11.32 | 8.24/9.58 | 8.89/10.14 | 2.67/3.16 |
| PointGPD[4] | 25.96/27.59 | 33.01/34.21 | 15.37/17.83 | 22.68/24.38 | 29.15/30.84 | 10.76/12.83 | 9.23/10.66 | 9.89/11.24 | 2.74/3.21 |
| GraspNet[11] | 27.56/29.88 | 33.43/36.19 | 16.95/19.31 | 26.11/27.84 | 34.18/33.19 | 14.23/16.62 | 10.55/11.51 | 11.25/12.92 | 3.98/3.56 |
| Gou等人[13] | 27.98/32.08 | 33.47/39.46 | 17.75/20.85 | 27.23/30.40 | 36.34/37.87 | 15.60/18.72 | 12.25/13.08 | 12.45/13.79 | 5.62/6.01 |
| Li等人[12] | 36.55/— | 47.22/— | 19.24/— | 28.36/— | 36.11/— | 10.85/— | 14.01/— | 16.56/— | 4.82/— |
| Ma等人[16] | 58.95/— | 68.18/— | 54.88/— | 52.97/— | 63.24/— | 46.99/— | 22.63/— | 28.53/— | 12.00/— |
| Ma等人[16]+CD | 63.83/— | 74.25/— | 58.66/— | 58.46/— | 70.05/— | 51.32/— | 24.63/— | 31.05/— | 12.85/— |
| GSNet[14] | 65.70/61.19 | 76.25/71.46 | 61.08/56.04 | 53.75/47.39 | 65.04/56.78 | 45.97/40.43 | 23.98/19.01 | 29.93/23.73 | 14.05/10.60 |
| GSNet+CD | 67.12/63.50 | 78.46/74.54 | 60.90/58.11 | 54.81/49.18 | 66.72/59.27 | 46.17/41.89 | 24.31/19.78 | 30.52/24.60 | 14.23/11.17 |
| 本文方法 | 71.76/63.90 | 82.44/73.77 | 67.59/56.25 | 63.81/52.03 | 75.84/62.52 | 56.71/44.37 | 27.01/20.12 | 33.82/24.77 | 15.56/12.11 |
| 本文方法+CD | **74.68/65.76** | **85.58/76.53** | **68.98/58.36** | **65.11/55.83** | **77.91/65.48** | **58.22/45.62** | **28.20/21.17** | **34.95/25.72** | **16.14/12.82** |

本文改进后的方法在 GraspNet-1Billion 的 RealSense 和 Kinect 相机上均取得了目前最好的效果, 无论是否添加碰撞检测, 结果均优于其他现有方法, 其中在见过的物体上, 分别在 RealSense 和 Kinect 相机提高了 7.56% 和 2.26%, 在相似的物体上分别提高了 10.30% 和 6.65%, 同时也优于目前在此指标上效果最好的方法. 在最困难的新物体上, 改进后的网络也全部优于基准方法, 提升了 3.89% 和 1.39%. 实验表明, 本文改进后的网络提升了生成抓取位姿的准确度, 说明多尺度注意力特征融合模块增强了对局部不同尺寸的形状感知能力.

### 3.2 损失对比

图 5 和图 6 为训练过程中抓取得分以及预测的抓取宽度的损失函数变化情况, 由于数据集中将完全不可接受的抓取位姿分数和抓取角度分别标注为−1 和 0, 导致了图 5 和图 6 中的曲线有一定差异. 曲线经过了平滑处理, 与 GSNet 和使用 MsCG 模块替换圆柱体分组模块 (cylinder grouping) 的 GSNet 进行了对比. 图 5 和图 6 中 Ms-CAFF 模块相比于 MsCG 模块和 GSNet

网络收敛速度更快, 同时损失值始终低于其他方法.

### 3.3 方法复杂度对比

表 2 为不同模块和方法的参数量和训练时间对比.

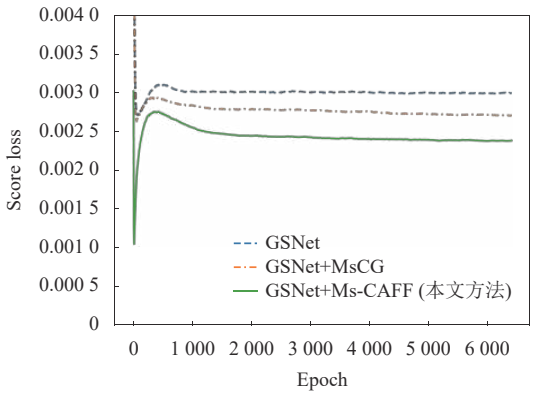


图 5 平滑后的抓取得分损失

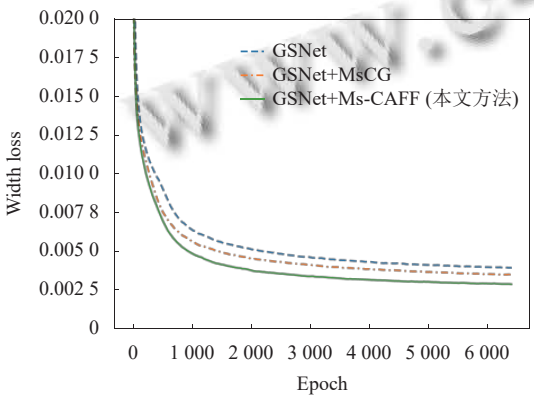


图 6 平滑后的抓取角度预测损失

表 2 不同方法的参数量和计算量

| 模型 | 参数量 (M) | 计算量 (G) | 训练时间 (h) |
|---|---|---|---|
| GSNet | 15.35 | 12.069 | 86 |
| GSNet+MsCG | 16.41 | 14.341 | 43 |
| 本文方法 | 19.56 | 17.338 | 42 |

GSNet 需要大约 86 h 才完成训练, 而本文改进后的方法将训练时长缩短了一半. 同时相比于 MsCG 模块也减少了 1 h, 训练时长与参数量和网络结构有关, 这表明本文方法虽然增加了参数, 但加入模块对结构的改进使得网络更加有效地捕捉数据特征从而更快地收敛到最佳. 多尺度注意力特征融合模块相比与基准方法和 MsCG 模块参数量和计算量均有一定程度的增加, 但仍然控制在可接受的范围内. 表 3 为网络推理单个场景的抓取位姿所需要的时间.

表 4 和表 5 为不同模块的推理时间对比. 测试数据被分为 Seen、Similar 和 Novel 这 3 类, 每类包含 30 个场景, 本文分别从每类中随机挑选 10 个场景进行了测试, 单场景采样 15 000 个点云, 计算平均的推理时间. 可以看到, 本文的方法网络没有明显增加推理时间, 与其他方法差距最多控制在 0.03 s 左右, 不同方法的时间基本保持一致. 上述两个实验证明, 本文提出的多尺度注意力特征融合模块虽然参数量和计算量有一定的增加, 但依然控制在可以接受的范围内, 因为计算的形状特征都集中于夹持器大小的空间内, 特征数量有限但是对抓取成功率影响较大, 同时证明小尺度的几何特征对抓取位姿估计至关重要, 能够在不增加计算量的同时显著提高网络的准确性.

表 3 不同方法的平均推理时间 (RealSense/Kinect) (s)

| 模型 | Seen | Similar | Novel |
|---|---|---|---|
| GSNet | 0.508 1/0.488 3 | 0.473 8/**0.496 4** | **0.443 0**/0.470 3 |
| GSNet+MsCG | **0.480 9**/**0.463 0** | 0.470 5/0.526 6 | 0.462 4/0.466 9 |
| 本文方法 | 0.487 1/0.471 8 | **0.466 4**/0.537 5 | 0.450 3/**0.440 1** |

表 4 单模块推理时间 (RealSense/Kinect) (s)

| 模块 | 时间 |
|---|---|
| Cylinder grouping (GSNet) | 0.000 922/0.000 845 |
| MsCG | 0.003 620/0.003 241 |
| 本文方法 (Ms-CAFF) | 0.003 423/0.003 380 |

表 5 不同模块的推理时间 (RealSense/Kinect) (s)

| 相机 | Attention | Gate fusion | Total |
|---|---|---|---|
| RealSense | 0.003 571 | 0.000 368 | 0.003 987 |
| Kinect | 0.004 829 | 0.000 404 | 0.005 316 |

### 3.4 抓取结果可视化分析

图 7 为抓取位姿可视化结果, 从数据集中挑选物体进行对比, 筛选了得分大于 0.5 的抓取位姿. 图 7(a)–图 7(c) 是具有复杂的局部几何形状的物体, GSNet 没有生成有效的抓取, 而本文提出的 Ms-CAFF 模块相比替换了 MsCG 模块的 GSNet 生成了更多的有效抓取, 图 7(d)–图 7(e) 为大型的有光滑的曲面物体. 图 7(f) 号物体则是局部几何形状有较大变化的物体, 只有 Ms-CAFF 模块生成了有效的抓取位姿. 可以看出受益于改进后基于注意力多尺度的特征提取方法, 网络在具有复杂几何形状的物体上生成的抓取位姿质量显著高于其他对比方法.

表 6 为不同方法在随机抽取的 3 个场景中生成的有效抓取数量 $N$ 以及有效抓取平均分 $S_h$ 和所有抓取的平均分 $S_a$, 由于生成抓取有一定的波动, 所列出的每个场景均进行了 5 次实验. 可以看出, 本文方法无论是生成的高得分抓取数量还是平均得分都领先于其他方法.

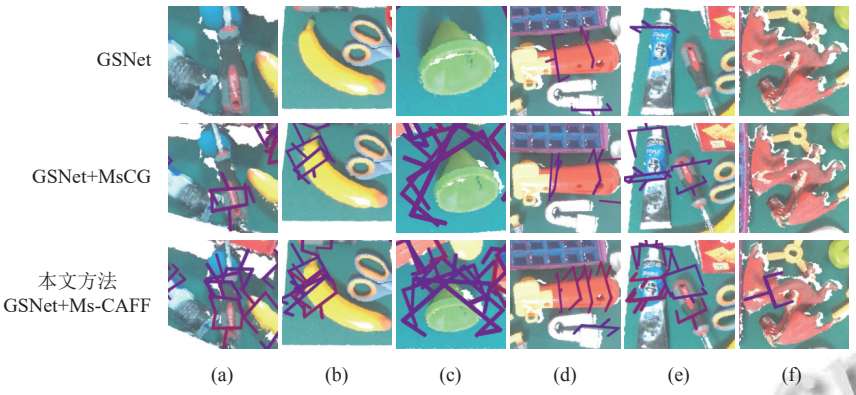


图 7 抓取位姿可视化

表 6 不同场景的有效抓取数量及平均分

| 方法 | Scene123 | | | Scene146 | | | Scene163 | | |
|---|---|---|---|---|---|---|---|---|---|
| | $N$ | $S_h$ | $S_a$ | $N$ | $S_h$ | $S_a$ | $N$ | $S_h$ | $S_a$ |
| GSNet | 11 | 0.5060 | 0.1523 | 13 | 0.4696 | 0.1613 | 7 | 0.1693 | 0.1693 |
| GSNet+MsCG | 25 | 0.5257 | 0.1743 | 15 | 0.4703 | 0.1971 | 9 | 0.1496 | 0.1496 |
| 本文方法 | 31 | 0.5402 | 0.2143 | 23 | 0.5576 | 0.2164 | 10 | 0.1816 | 0.1816 |

### 3.5 消融实验

本节将各模块组合进行比较实验. 将 GSNet 与 MsCG 结合, 并在 MsCG 基础上分别增加重新设计的门控单元和注意力特征融合模块, 以验证注意力特征融合模块和门控单元在不同物体上提升效果. 实验结果如表 7 所示. 重新设计的门控融合单元对特征的过滤使得网络的各项指标都有一定的提升的同时也保持了较快的推理速度. 而注意力机制的加入使得抓取成功率有了明显的提高, 尤其是在最为困难的 Novel 指标上, 分别提高了 1.28% 和 1.15%. 实验表明, 本文提出的多尺度圆柱体注意力特征融合模块中的注意力机制对不同尺度的形状特征进行了有效的提取和融合, 特别是改善了在陌生物体上生成抓取位姿的准确度.

表 7 消融实验结果 (RealSense/Kinect)(%)

| 模型 | *AP* (Seen) | *AP* (Similar) | *AP* (Novel) |
|---|---|---|---|
| GSNet | 67.12/63.50 | 54.81/49.18 | 24.31/19.78 |
| GSNet+MsCG | 70.37/64.72 | 59.33/51.76 | 26.82/19.89 |
| GSNet+GU | 71.78/64.95 | 60.61/52.84 | 26.96/20.03 |
| GSNet+Attention | 73.84/65.58 | 64.28/54.91 | 28.10/21.04 |
| 本文方法 | **74.68/65.76** | **65.11/55.83** | **28.20/21.17** |

### 3.6 真实场景实验

为了评估本文提出的方法在真实场景中的性能, 使用带有平行夹持器的机械臂和 Kinect v2 相机进行了机器人实验, 使用配备 GTX1050 和 Ubuntu 18.04 的计算机用于运行模型, 与基准方法 GSNet 进行对比. 挑选了 8 个具有各种形状的物体形成杂乱的场景. 实验场景及可视化结果如图 8 所示.


图 8 真实实验场景、点云和抓取位姿可视化图

首先通过 Kinect 相机采集场景点云, 然后进行直通滤波过滤场景外的点云, 使用最远点采样 (FPS) 将点云采样至 15 000 个点, 输入训练好的模型, 得到抓取点坐标和旋转矩阵等抓取参数. 机器人进行多次抓取, 直到抓取所有物体, 进行了 20 次实验取平均值, 使用物体个数和尝试次数相除作为抓取实验的成功率.

实验结果如表 8 所示, 在现实场景中, 受到相机拍摄质量以及物体遮挡的影响, 导致输入的点云质量较

差, 且有较多的噪点, 使得网络生成的抓取位姿较不稳定, 会出现碰撞问题, 导致在较大物体上容易出现抓取失败的情况. 但是在所有物体抓取成功率上本文方法相比于 GSNet 依然取得了更好的结果.

表 8 真实场景的抓取实验

| 方法 | 物体个数 | 尝试次数 | 成功率 (%) |
|---|---|---|---|
| GSNet | 8 | 13 | 61.5 |
| 本文方法 | 8 | 10 | 80 |

## 4 结论与展望

针对 GSNet 仅使用固定大小的圆柱体提取特征、忽略物体微小的几何特征, 以及多尺度几何特征难以融合的问题, 本文提出了一个多尺度圆柱体注意力特征融合模块 (Ms-CAFF), 使用注意力机制有效地融合被抓取物体不同尺度的几何特征信息. 在大规模公开抓取数据集 GraspNet-1Billion 的实验中, 本文方法优于其他现有方法. 改进后的模型相比之前缩短了约一半的训练时间, 同时加快了网络的收敛速度, 在没有明显增加计算量的情况下提高了网络生成抓取位姿准确度. 可视化实验表明, 相比于其他方法, 本文方法在具有复杂形状的物体上生成的抓取位姿的数量和质量都有了显著提高. 实际场景的机器人抓取实验证明了本文方法在真实场景的有效性. 下一步将继续研究不同尺度特征融合的方法, 同时进一步提高网络模型对真实场景的适应能力.

## 参考文献


1 Miller AT, Allen PK. Examples of 3D grasp quality computations. Proceedings of the 1999 IEEE International Conference on Robotics and Automation. Detroit: IEEE, 1999. 1240–1246.

2 Dang H, Allen PK. Semantic grasping: Planning robotic grasps functionally suitable for an object manipulation task. Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vilamoura-Algarve: IEEE, 2012. 1311–1317.

3 Liang HZ, Ma XJ, Li S, *et al*. PointNetGPD: Detecting grasp configurations from point sets. Proceedings of the 2019 International Conference on Robotics and Automation. Montreal: IEEE, 2019. 3629–3635.

4 Mousavian A, Eppner C, Fox D. 6-DoF GraspNet: Variational grasp generation for object manipulation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 2901–2910.

5 Redmon J, Angelova A. Real-time grasp detection using convolutional neural networks. Proceedings of the 2015 IEEE International Conference on Robotics and Automation. Seattle: IEEE, 2015. 1316–1322.

6 Kumra S, Kanan C. Robotic grasp detection using deep convolutional neural networks. Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vancouver: IEEE, 2017. 769–776.

7 Zhao BL, Zhang HB, Lan XG, *et al*. REGNet: Region-based grasp network for end-to-end grasp detection in point clouds. Proceedings of the 2021 IEEE International Conference on Robotics and Automation. Xi'an: IEEE, 2021. 13474–13480.

8 Wei W, Luo YK, Li FY, *et al*. GPR: Grasp pose refinement network for cluttered scenes. Proceedings of the 2021 IEEE International Conference on Robotics and Automation. Xi'an: IEEE, 2021. 4295–4302.

9 Jeng KY, Liu YC, Liu ZY, *et al*. GDN: A coarse-to-fine (C2F) representation for end-to-end 6-DoF grasp detection. Proceedings of the 2020 Conference on Robot Learning. Cambridge: PMLR, 2021. 220–231.

10 Ni PY, Zhang WG, Zhu XX, *et al*. PointNet++ grasping: Learning an end-to-end spatial grasp generation algorithm from sparse point clouds. Proceedings of the 2020 IEEE International Conference on Robotics and Automation. Paris: IEEE, 2020. 3619–3625.

11 Qi CR, Yi L, Su H, *et al*. PointNet++: Deep hierarchical feature learning on point sets in a metric space. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 5105–5114.

12 Li YM, Kong T, Chu RH, *et al*. Simultaneous semantic and collision learning for 6-DoF grasp pose estimation. Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague: IEEE, 2021. 3571–3578.

13 Gou MH, Fang HS, Zhu ZD, *et al*. RGB matters: Learning 7-DoF grasp poses on monocular RGBD images. Proceedings of the 2021 IEEE International Conference on Robotics and Automation. Xi'an: IEEE, 2021. 13459–13466.

14 Wang CX, Fang HS, Gou MH, *et al*. Graspness discovery in clutters for fast and accurate grasp detection. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 15944–15953.

15 Fang HS, Wang CX, Gou MH, *et al*. GraspNet-1Billion: A large-scale benchmark for general object grasping. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 11441–11450.

16 Ma HX, Huang D. Towards scale balanced 6-DoF grasp detection in cluttered scenes. Proceedings of the 2023 Conference on Robot Learning. Auckland: PMLR, 2023. 2004–2013.

17 Morrison D, Corke P, Leitner J. Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach. Proceedings of the 2018 Robotics: Science and Systems XIV. Pittsburgh, 2018. 1–10.

18 Chu FJ, Xu RN, Vela PA. Real-world multiobject, multigrasp detection. IEEE Robotics and Automation Letters, 2018, 3(4): 3355–3362. [doi: 10.1109/LRA.2018.2852777]

19 Ten Pas A, Gualtieri M, Saenko K, *et al*. Grasp pose detection in point clouds. The International Journal of Robotics Research, 2017, 36(13–14): 1455–1473. [doi: 10.1177/0278364917735594]


(校对责编: 张重毅)