DOI:10.3979/j.issn.1673-825X.202408070205

基于深度值引导的机械臂多尺度抓取检测方法

刘想德,杨超旋,郑　凯,张　毅,蒋　菲
(重庆邮电大学 国家信息无障碍工程研发中心,重庆 400065)

摘　要:针对 4DoF 抓取检测的性能,改进了抓取表示方法,并提出了基于深度值引导的机械臂多尺度抓取检测框架(DGM-Grasp)。在编解码网络的基础上,融入多尺度跨空间注意力下采样模块,以更好地聚焦抓取特征;为了提取多尺度语义信息,设计了渐进式多尺度特征融合-解码模块;通过提出的深度值引导的抓取筛选模块解决抓取过程中的碰撞问题。DGM-Grasp 在 Cornell 和 Jacquard 两个单目标数据集上准确率分别达到 98.6%和 95.25%,检测用时可降低至 21 ms;在多目标数据集上也取得了良好的效果;消融实验和真实抓取实验成功率达到 96%。实验充分验证了 DGM-Grasp 的泛化能力和性能。

关键词:机械臂;深度学习;抓取检测;特征融合;深度图像;抓取表示
中图分类号:TP242　　**文献标志码**:A　　**文章编号**:1673-825X(2025)05-0717-12

A multi-scale robotic grasp detection method based on depth-guided mechanisms

LIU Xiangde, YANG Chaoxuan, ZHENG Kai, ZHANG Yi, JIANG Fei
(Research and Development Center for Information Accessibility, Chongqing University of Posts and Telecommunications, Chongqing 400065, P. R. China)

Abstract: To enhance the performance of 4-DoF grasp detection, this paper improves the grasp representation and proposes a depth-guided multi-scale grasp detection framework (DGM-Grasp) for robotic manipulators. Built upon an encoder - decoder architecture, the framework integrates a multi-scale cross-spatial attention down-sampling module to better focus on grasp-relevant features. To extract semantic information at different scales, a progressive multi-scale feature fusion and decoding module is designed. In addition, a depth-guided grasp filtering module is introduced to address collision problems during the grasping process. Experimental results show that DGM-Grasp achieves accuracies of 98.6% and 95.25% on the Cornell and Jacquard single-object datasets, respectively, while reducing detection time to 21 ms. The method also performs effectively on multi-object datasets, achieving a 96% success rate in ablation and real-world grasping experiments. These results demonstrate the superior generalization ability and performance of DGM-Grasp.

Keywords: robotic arm; deep learning; grasp detection; feature fusion; depth image; grasp representation

收稿日期:2024-08-07　**修订日期**:2025-09-05　**通讯作者**:杨超旋　1173528901@qq.com
基金项目:国家自然科学基金项目(51905065)
Foundation Item: National Natural Science Foundation of China (51905065)

0 引 言

机械臂抓取目前在工业、服务和医疗等领域不断提高影响力,如何实现鲁棒目标的准确定位抓取位姿检测并提高物体的实际环境抓取成功率仍然面临一系列问题。

早期的抓取检测方法以分析法[1]和感知生成法[2]为代表,它们都依赖既定的物体物理数据和数学模型进行推理,仅适用于相对稳定的环境,面对具有形状、大小和数量等变化较大的特性的鲁棒目标,无法适应。

将深度学习引入抓取检测已逐渐成为研究热点。基于点云的 6DoF 机器人抓取研究是方向之一。文献[3]发布了开源的大规模 6DoF 抓取数据集 GraspNet-1Billion,并提出一个以点云为输入的端到端抓取检测网络 GraspNet。这些 6DoF 抓取检测方法需要高性能的传感器提取完整精确的三维点云信息,同时进行更复杂的几何掩膜候选标注以保证检测的鲁棒性,并具有相当大的计算量。因此,相比于 6DoF 抓取方法,准确快速的 4DoF 抓取方法更能适应当前任务需求,具有研究与应用潜力。4DoF 领域中,文献[4]应用深度学习,提出了一种实时性较差的两阶段抓取检测网络,一阶段通过滑动窗口生成一系列候选抓取位姿,二阶段使用分类的方法选择最佳抓取位姿。文献[5]采用了类似 AlexNet 的 CNN 架构,实现了抓取检测的单阶段回归方法,检测用时降到 76 ms,但是准确率较低。文献[6]提出一种基于热力图匹配的方法 GG-CNN,以深度信息输入至没有任何全连接层的 CNN,输出 4 张热力图,从而生成像素级抓取位姿。文献[7]基于 GG-CNN,提出了 GR-CNN,可以处理 RGB-D 多模态信息,增加了多个 ResBlocks[8],提升了特征提取能力。

文献[9]提出一种轻量级无关对象的可分离卷积网络 GARDSCN,使用于实时抓取检测。文献[10]同样专注于轻量级网络,通过解耦合的抓取质量网络生成初始抓取矩形,并采用自适应过滤和椭圆拟合优化方法调整生成的抓取矩形。

特征融合方面,文献[11]集成注意力机制和多尺度特征融合,使网络能够充分关注抓取区域并根据物体尺度灵活调整抓取区域。TDMAG-Net[12]引入双分支解卷积来消除编解码结构卷积的棋盘伪影问题,并设计了多维注意模块来调整特征图的全局、局部和通道维度,提供更具区分性的特征。文献[13]通过残差注意力模块、特征融合模块和特征增强金字塔模块增强网络的特征提取能力。

SE-ResUNet[14]在 U-Net 中结合通道注意力和残差模块,提升了网络性能。文献[15]设计了结构先验注意(SPA)模块,与基础特征提取模块及残差连接结合,形成了类似 U-Net 抓取检测网络,实现抓取检测。DSNet[16]结合了 NLP 领域的 Transformer 分支和 U-Net 分支,通过瓶颈点处的双向桥接与交叉注意力机制,实现了局部特征和全局表示的保留。

文献[17]也将 Transformer 应用在抓取检测上,提升了全局信息建模能力,但自注意力局部特征提取能力较弱,计算复杂度高,实时性受到制约。

在抓取表示相关研究方面,文献[18]无需物体三维模型,直接从图像中预测抓取点,单个点能够较准确地表示抓取位置。文献[4]在文献[19]7 维有向抓取框的基础上进行改进,提出了 5 维抓取框,包括中心坐标、宽度、高度和旋转角度。文献[6]将五维抓取框简化为四维度,并假设抓取矩形内的所有像素都具有最佳且相同的抓取质量,但缺乏对抓取中心点的突出表示。文献[20]提出了一种基于二维高斯的抓取表示方法以解决矩形抓取表示抓取定位模糊的问题,但是需要更多计算资源。文献[21]提出了一种同时适用于平行夹爪和三指夹爪的定向箭头表示模型(oriented arrow representation, OAR)模型,并提出自适应抓取属性模型(adaptive grasping attribute, AGA)模型,解决训练中的角度冲突,该文发现真实抓取失败的主要原因来自目标物体或相邻物体碰撞阻碍,文献[22-23]也提出了相同的观点。以上研究的抓取表示方法均未考虑平行夹爪夹指本身的存在,容易导致碰撞,影响抓取稳定性。

为了提升抓取检测模型性能并解决平行夹爪的碰撞问题,本文提出了基于深度值引导的多尺度抓取检测方法(depth-guide multi-scale grasp detection, DGM-Grasp),主要研究内容和贡献如下。

1)基于五维度抓取表示,提出了考虑平行夹爪夹指区域的改进的六维度抓取表示。依据该抓取表示提出一种基于深度值引导的抓取位姿筛选方法,用于对抓取位姿集合进行筛选和重排,从而在避免碰撞的同时获得最优的抓取位姿。

2)提出了一种多尺度跨空间注意力下采样模块,使模型能够更好地关注目标的抓取特征,从而更有效地区分物体与背景。

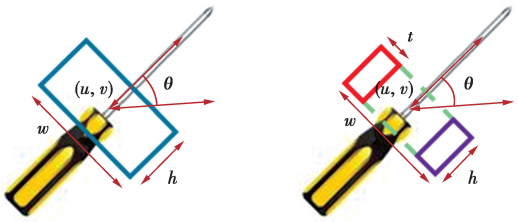
3)设计了一种渐进式多尺度特征融合与解码

模块,不仅引入了多尺度语义信息,还缓解了因非相邻特征层融合效果差的问题,同时能够起到解码器的作用。

实验表明,DGM-Grasp 在单目标数据集 Cornell 和 Jacquard 上准确率分别达到 98.6%和 95.25%,检测用时降至 21 ms;在多目标数据集和实际抓取实验上同样取得了良好的效果。

1　问题表述

为了更有效地理解机器人抓取位姿检测算法,首先需要定义抓取问题的表示方法。本文将抓取问题定义为平行夹爪的 4DoF(four degrees of freedom)平面抓取,提出一种改进抓取表示方法,将生成的像素抓取表示通过手眼矩阵转换成实际机械臂的抓取位姿完成抓取,如图 1 所示。



a　五维度抓取表示　　b　改进的抓取表示

图 1　两种抓取表示示意图

Fig.1　Diagram illustrating two types of grasping representations

自从文献[19]提出用旋转矩形框表示抓取姿势以来,更为广泛使用的方法是文献[4]中的五维度抓取表示,即

$$g=(u,v,\theta,h,w) \tag{1}$$

图 1a 是抓取位姿的矩形表示。抓取位姿 g 中,(u,v)为抓取矩形中心点的像素坐标,即平行夹爪的位置,θ 为抓取矩形旋转角度, h 为抓取矩形的高度,w 为抓取矩形的开合宽度,该表示方法可以很好地表示抓取位姿。但是,平行夹爪夹指本身还具有厚度,忽略夹指厚度,容易发生平行夹爪与目标物体或相邻物体的碰撞,从而导致抓取稳定性降低,为解决这一问题并筛选出最优的抓取位姿,本文提出了一种改进的六维度抓取表示方法,即

$$g=(u,v,\theta,h,w,t) \tag{2}$$

图 1b 中,u、v、θ、h、w 的定义与五维度抓取表示相同,t(thickness)为平行夹爪的夹指厚度,改进后的表示方法由 3 个连续的矩形框构成,两侧的矩形代表夹指的高度和厚度,即物理尺寸,中间的矩形代表单次夹取范围,这种抓取位姿表示方法能够获得夹指矩形内像素的信息,为后续抓取位姿筛选重排做准备。

为了避免网络学习的混乱,便于训练和验证,本文将抓取位姿表示的 h 和 t 设置为固定参数,即平行夹爪的实际物理尺寸在图像坐标系下的映射,参照 Morrison 等[6]的抓取定义方法,将网络中的像素级抓取表示定义为

$$G=(Q,W_g,\Theta)\in\mathbb{R}^{3\times H\times W} \tag{3}$$

式(3)中 Q、W_g、Θ 为热力图,热力图中的每个像素分别对应一个抓取质量分数、抓取宽度和抓取角度,其中抓取质量分数取值范围为[0,1],分数越大,代表抓取成功率越高;抓取宽度取值范围为$[0,W_g^{\max}]$,$W_g^{\max}$ 表示平行夹爪的最大物理开合尺寸在像素坐标下的映射;抓取角度由 $\cos(2\theta)$ 和 $\sin(2\theta)$ 两张热力图通过公式 $\theta=\frac{1}{2}\arctan\left(\frac{\sin(2\theta)}{\cos(2\theta)}\right)$ 计算得出,取值范围为$\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$。

然后,通过 2.4 节的抓取筛选策略,对 G 中的抓取位姿进行筛选重排,得到最优抓取位姿 g,基于变换矩阵 $\boldsymbol{T}_{ci}$、$\boldsymbol{T}_{rc}$和深度信息 z 可以将二维图像中的抓取框变换为相机坐标下,继而转换为机器人坐标系下的抓取位姿,公式为

$$G_r=\boldsymbol{T}_{rc}(\boldsymbol{T}_{ci}g)=(x,y,z,\theta,w) \tag{4}$$

2　方法描述

本文提出的 DGM-Grasp 主要包括 3 个部分:多尺度跨空间注意力下采样(multi-scale cross-spatial attention downsampling,MCAD)、渐进式多尺度特征融合-解码(asymptotic multi-scale feature fusion-decoder,AMFFD)、基于深度值引导的抓取位姿筛选方法(depth guided grasp selector,DGGS)。

DGM-Grasp 可以看作由编码器和解码器组成的端到生成式神经网络,输入为 RGB 图像和 Depth 图像拼接而成的 4 通道多模态 300pixel×300pixel 图像,如图 2 所示。方法整体结构和运行流程如下。

首先,用核尺寸为 9×9 的较大卷积核、步长为 1 的卷积操作将 4 通道图像的通道数提升至 16,输入至编码器进行下采样。编码器由 3 个 MCAD 模块组成,将特征图的通道数提升至 128,尺寸降至 37×37。编解码结构由残差层进行连接,残差层由 5 个残差块与 1 个 EMA[24]注意力模块串行组成。通过残差块的跳跃链接,可以充分提取深层特征并解决

梯度消失问题,同时将残差块中的 ReLU 激活函数替换为非线性的 Mish 激活函数。与 ReLU 激活函数面对负值直接置零不同,Mish 激活函数允许一定的负梯度,同时更加平滑,能够捕获更加复杂的特征。

然后,渐进式多尺度特征融合-解码模块对输入的 3 张不同层级的特征图进行特征融合和解码,将特征图还原至原始尺度,通过多次卷积操作得到表示抓取位姿的 3 张热力图。最后,通过 DGGS 模块对抓取位姿进行筛选重排,输出最优抓取姿态。

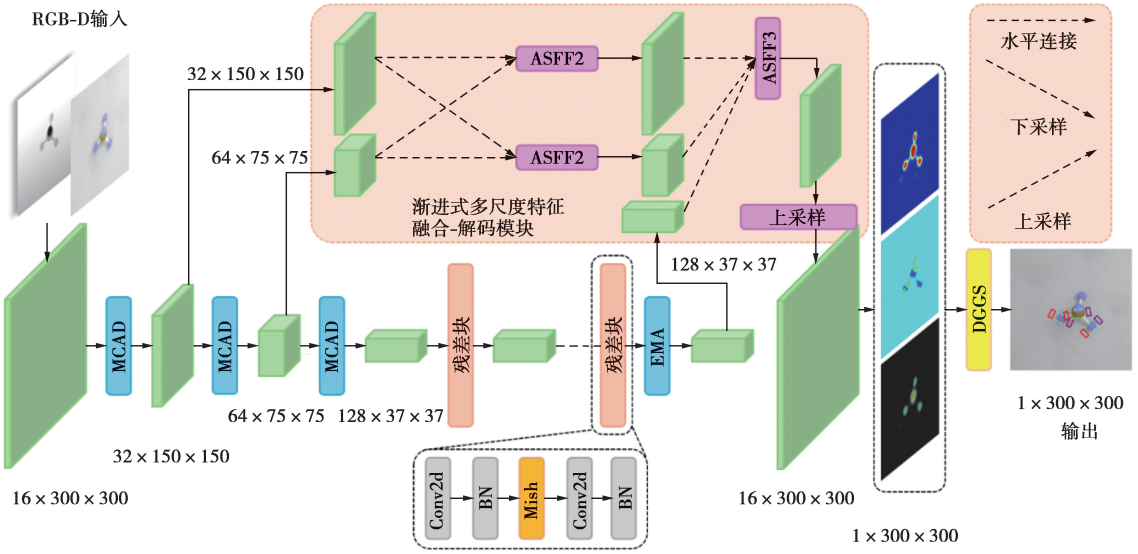


图 2 基于深度值引导的多尺度抓取检测方法的结构

Fig.2 Structure of depth-guide multi-scale grasp detection

2.1 多尺度跨空间注意力下采样模块

本文设计的 MCAD 模块如图 3 所示。下采样层由尺寸为 4×4、步长为 2 的卷积层和批量归一化(batch normalization, BN)层组成,在其后串联 EMA[24](efficient multi-scale attention)注意力机制,有助于在网络早期便能够有选择性地关注重要的物体抓取特征并抑制其他无关的特征。

EMA[24]是一种轻量级的高效多尺度注意力机制,提升计算效率并避免了通道降维。结合通道和空间注意力的优势,利用特征分组并行和跨空间学习融合多尺度特征信息,联系全局上下文,帮助抓取检测模型理解图像的整体结构和语义信息,提升网络对复杂场景的识别能力。EMA 注意力机制首先将输入特征图 $X \in \Re^{C\times H\times W}$ 沿通道维度分成 G 组,每组包含 $C//G$ 个通道,H 和 W 分别表示输入特征的空间高度和宽度。EMA 包含 3 个分支,其中 2 个 1×1 分支负责编码通道注意力,捕获长距离依赖关系,对每个子特征图进行水平和垂直方向的一维全局平均池化操作,分别编码水平和垂直方向的全局信息,公式为

$$z_c^H(H)=\frac{1}{W}\sum_{0\leqslant i\leqslant W}x_c(H,i) \tag{5}$$

$$z_c^w(W)=\frac{1}{H}\sum_{0\leqslant j\leqslant H}x_c(j,W) \tag{6}$$

将两个编码后的特征图进行拼接,并使用 1×1 卷积融合特征。最后,将输出重新分解为 2 个一维向量,并分别使用 Sigmoid 函数进行非线性激活,得到每个通道的注意力权重,通过逐元素乘法聚合原始分组特征和两个子权重,实现不同子网络之间通道交互特征的动态重校准。3×3 分支负责捕获多尺度特征表示,扩大特征空间。它使用 3×3 卷积核进行特征提取,并保留原始的空间分辨率。

1×1 分支和 3×3 分支的输出通过跨空间学习进行融合,突出重要的空间特征。它首先对 2 个分支的输出分别进行二维全局平均池化操作,编码全局空间信息,公式为

$$z_c=\frac{1}{H\times W}\sum_{j}^{H}\sum_{i}^{W}x_c(i,j) \tag{7}$$

然后,将 1×1 和 3×3 分支池化后的特征图分别与池化前的另一分支进行矩阵点积操作得到两个空间注意力图。最后,将两个空间注意力图相加,并进行非线性激活后与输入特征图相乘,得到最终的特征图输出。

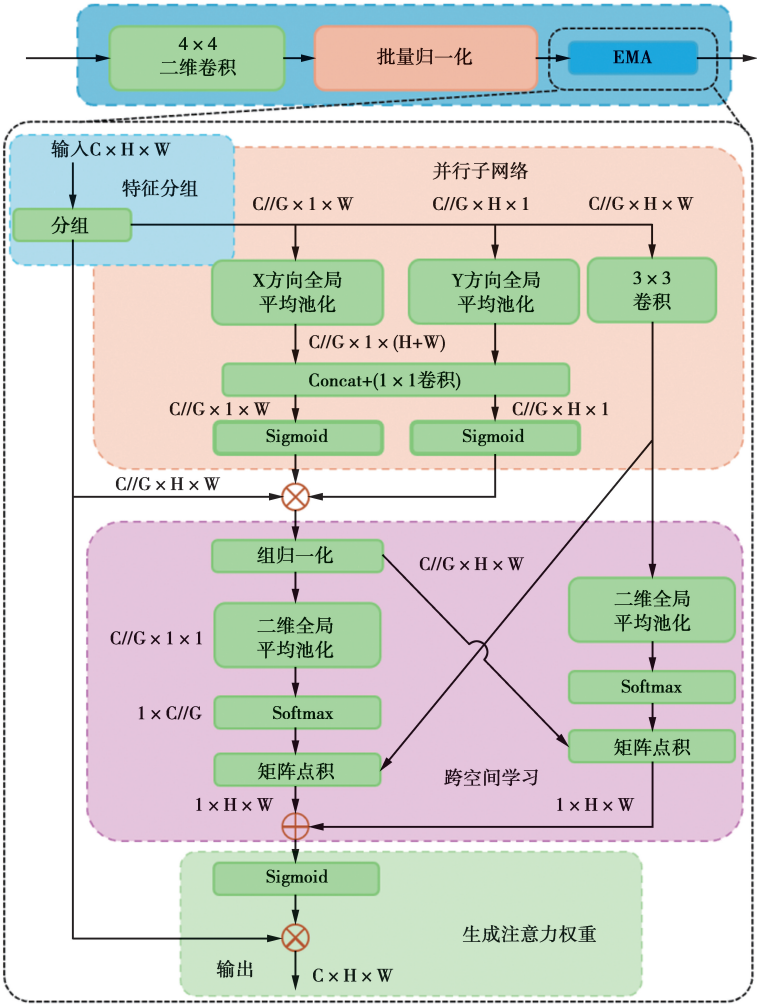


图 3　MCAD 模块结构图

Fig.3　MCAD module structure diagram

2.2　渐进式多尺度特征融合-解码模块

抓取过程中,待抓物体的种类较多,尺度差异较大,同一种类型的抓取点可能以不同的尺度出现。较小尺度的抓取点容易因为具有较少的像素信息而导致在特征提取过程中被忽略,影响抓取检测的效果。现有的抓取检测网络较少考虑图像的多尺度信息,或者非相邻特征层的融合效果不佳。

受文献[25]启发,本文提出了 AMFFD。该模块能够从浅层特征开始,逐渐引入最抽象的深层特征完成多尺度特征融合,并兼顾上采样,恢复特征图分辨率,起到解码器的作用。这种融合方式能够让不同层级的语义信息在渐进融合的过程中更加接近,提升非相邻特征层融合效果。其结构如图 2 所示。

AMFFD 接受 3 个不同层级的特征图作为输入,来自解码器的第一、第二个多尺度跨空间注意力下采样模块 MCAD 和残差层的输出,分别作为浅层特征、中层特征和深层特征。首先,为了对齐维度并为特征融合做准备,根据所需的采样率使用不同次数的卷积操作,使用一次核尺寸为 4、步长为 2 的卷积或转置卷积操作实现 2 倍下采样或上采样,使用 2 次相同的转置卷积操作实现 4 倍上采样,最大程度提取特征。随后,通过自适应空间特征融合[26](adaptive spatial feature fusion,ASFF)为对齐维度后的浅层特征和中层特征进行自适应融合,然后再得到的两个新特征与深层特征进行 3 个不同层级的特征融合。令 $\boldsymbol{x}_{ij}^{n\to l}$ 表示位置(i,j)处从 n 层调整到 l 层的特征向量,融合结果表示为 $\boldsymbol{y}_{ij}^{l}$,三层级融合公式为

$$\boldsymbol{y}_{ij}^{l}=\alpha_{ij}^{l}\cdot\boldsymbol{x}_{ij}^{1\to l}+\beta_{ij}^{l}\cdot\boldsymbol{x}_{ij}^{2\to l}+\gamma_{ij}^{l}\cdot\boldsymbol{x}_{ij}^{3\to l}\tag{8}$$

式(8)中:α_{ij}^{l}、β_{ij}^{l}和 γ_{ij}^{l}表示 3 个不同层级的特征图相对于 l 层的空间重要性权重,由分别以 $\lambda_{\alpha_{ij}}^{l}$、$\lambda_{\beta_{ij}}^{l}$和

$\lambda_{\gamma ij}^{l}$为控制参数的 SoftMax 函数计算得到;权重标量 λ_{α}^{l}、λ_{β}^{l} 和 λ_{γ}^{l} 是 3 个层级调整后的特征向量使用 1×1 卷积操作得到的;$\alpha_{ij}^{l}+\beta_{ij}^{l}+\gamma_{ij}^{l}=1$,且 $\alpha_{ij}^{l},\beta_{ij}^{l},\gamma_{ij}^{l}\in[0,1]$,计算公式为

$$\alpha_{ij}^{l}=\frac{e^{\lambda_{\alpha ij}^{l}}}{e^{\lambda_{\alpha ij}^{l}}+e^{\lambda_{\beta ij}^{l}}+e^{\lambda_{\gamma ij}^{l}}} \tag{9}$$

2.3 基于深度值引导的抓取位姿筛选方法

研究者大多选择抓取质量热力图中最高置信度坐标对应的抓取位姿,但这并不一定是最优的抓取位姿,依然有可能产生抓取碰撞等失败情况。

鉴于此,本文以改进的六维度抓取表示为基础,提出 DGGS,对抓取位姿集合进行筛选重排,解决抓取碰撞问题,输出最优抓取位姿。DGGS 接受 3 张热力图作为输入,分别是抓取质量、抓取宽度和抓取角度热力图,相当于一组抓取位姿的集合。

首先,通常情况下认为,对于平行夹爪,当待抓物体的形状与夹持器的物理容纳空间相匹配时,才能保证抓取的成功率;对于 4DoF 抓取,平行夹爪的两根夹指处的最小深度值大于夹取空间的最小深度值时才能成功抓取。利用本文提出的考虑夹指厚度的六维抓取表示,定义初始的筛选约束表示为

$$\begin{aligned} G^{*}&=\operatorname{argmax}_{G\in \mathrm{Top}_{100}} Q\\ \text{s.t.}&\begin{cases} d_{\mathrm{minA}}>d_{\mathrm{minC}}\\ d_{\mathrm{minB}}>d_{\mathrm{minC}}\\ Q>0.2\\ d_{\mathrm{m}}\geqslant 10 \end{cases} \end{aligned} \tag{10}$$

式(10)中:G^{*} 表示初始筛选出的最多 100 个符合约束的抓取位姿;d_{minA}、d_{minB}和 d_{minC}分别表示夹指 A、B 和夹取空间 C 的矩形表示区域的最小深度值;d_{m}(min distance)为抓取点之间的最小间隔,这里设置为 10 个像素。

因为矩形最小深度值不能完整表示整个矩形的深度分布,受文献[19]工作启发,通过非线性指标 $\tilde{d}$ 对初始筛选结果 G^{*} 中的抓取姿态进行重排,非线性指标计算公式为

$$\tilde{d}=\frac{d_{\mathrm{avgA}}d_{\mathrm{avgB}}}{d_{\mathrm{avgC}}^{2}} \tag{11}$$

式(11)中:d_{avgA}、d_{avgB}和 d_{avgC}别表示夹指 A、B 和夹取空间 C 的矩形表示区域的平均深度值,当夹取空间矩形中待抓物体所占比例越大,d_{avgC}的值越小,$\tilde{d}$ 的值越大,此时抓取位姿的开合宽度与待抓物体越匹配;d_{avgA}、d_{avgB}和 d_{avgC}的差距越大,$\tilde{d}$ 的值越大,夹取空间能容纳的深度越大。同时,$\tilde{d}$ 同样具有区分抓取成功与否的功能,通过计算 Cornell 抓取数据集[4]中 2912 个标签中正标签和负标签的 $\tilde{d}$ 值的分布可以认识到这一点。

负标签的值分布以 1 呈高斯分布,而正标签的值则分布在 1 的右侧且普遍大于负标签的值,如图 4 所示。综上所述,根据 $\tilde{d}$ 值大小作为 G^{*} 的重排指标,能够筛选出最优的抓取位姿。

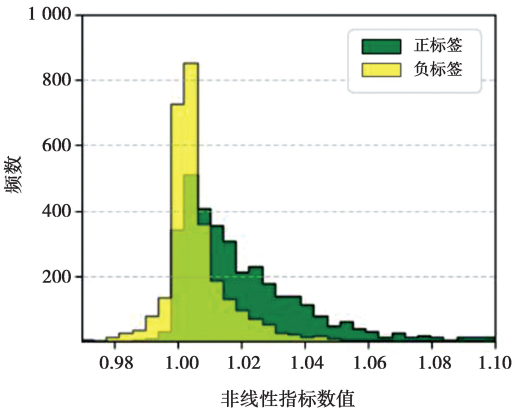


图 4 Cornell 数据集正、负标签 $\tilde{d}$ 值分布

Fig.4 Cornell dataset positive and negative label $\tilde{d}$ value distribution

2.4 损失函数

本文方法使用的损失函数为

$$L=L_{\mathrm{quality}}+L_{\mathrm{angle}}^{\cos}+L_{\mathrm{angle}}^{\sin}+L_{\mathrm{width}}+L_{\mathrm{GIoU}} \tag{12}$$

式(12)中:L_{quality}采用均方误差 MSE(mean squared error)损失函数,$L_{\mathrm{angle}}^{\cos}$,$L_{\mathrm{angle}}^{\sin}$和 L_{width}的定义与 L_{quality}相同。L_{quality}的定义为

$$L_{\mathrm{quality}}=MSE(q_{g},\hat{q}_{g}) \tag{13}$$

式(13)中:q_{g} 和 $\hat{q}_{g}$ 分别是抓取质量热力图和标注抓取矩形(ground truth labeling,GT)的中间三分之一的图像掩模。为了使模型更关注抓取矩形两侧表示夹指的部分,本文还加入基于沙漏形匹配机制的 L_{GIoU}[22]损失,定义为

$$L_{\mathrm{GIoU}}=\frac{1}{2}-\left(I_{\mathrm{IoU}}-\frac{\varepsilon-u}{\varepsilon}\right) \tag{14}$$

式(14)中:I_{IoU}是沙漏型预测抓取矩形 g 和 GT 矩形的交并比;ε 是包含 g 和 GT 矩形的最小框的面积;u 是 g 和 GT 矩形的并集。

3　实验结果与分析

本文方法均使用 GGCNN 工程中的以数据集矩形中心 1/3 区域中全像素标记抓取点,同一区域内的抓取点具有相同的抓取角度和宽度的方法进行训练。

实验图像部分:文中仅具有基于深度值引导的抓取位姿筛选方法 DGGS 模块的抓取检测方法使用六维度抓取表示,其他对比方法使用五维度抓取表示,但为了方便对比,实验图像中均使用本文提出的六维抓取表示进行抓取结果显示。

3.1　训练平台及参数细节

DGM-Grasp 在环境为 Ubuntu 16.04,处理器为 Intel © Core™ i5-13600KF,在单个 NVIDIA GeForce RTX 4070 的 PC 平台上进行训练和测试。训练框架为 Pytorch 和 CUDA 11.3,训练 Epochs 为 100,每个 Epoch 进行 1000 次迭代,使用 Adam 优化器,初始学习率为 0.001,每 15 个训练轮次衰减一次,衰减参数为 0.1,batch size 设置为 8。

3.2　评估指标

为了更方便比对,本文使用[19]的方法评估 DGM-Grasp 的效果,当预测抓取矩形 g 同时满足以下 2 个条件时,认为预测结果正确。

$$\begin{cases} \left| g_\theta^{GT} - g_\theta \right| < 30° \\ J(g^{GT}, g) = \dfrac{\left| g^{GT} \cap g \right|}{\left| g^{GT} \cup g \right|} > 25\% \end{cases} \tag{15}$$

即预测抓取矩形 g 与 GT 矩形的角度差异不超过 30°且二者的 *Jaccard* 指数(交并比)大于 25%。

3.3　单目标抓取位姿检测

本文使用 2 种经典的单目标抓取数据集 Cornell[4] 数据集和 Jacquard[27] 数据集来评估 DGM-Grasp 的性能,单目标抓取检测结果如图 5 所示。

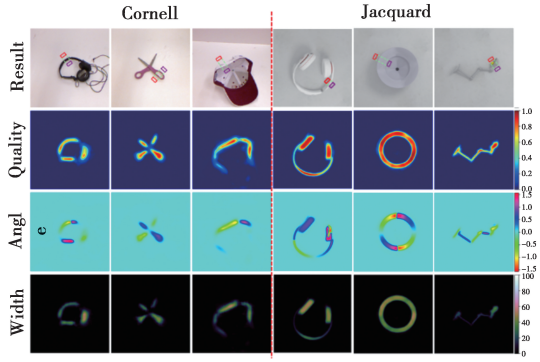


图 5　单目标数据集抓取检测结果

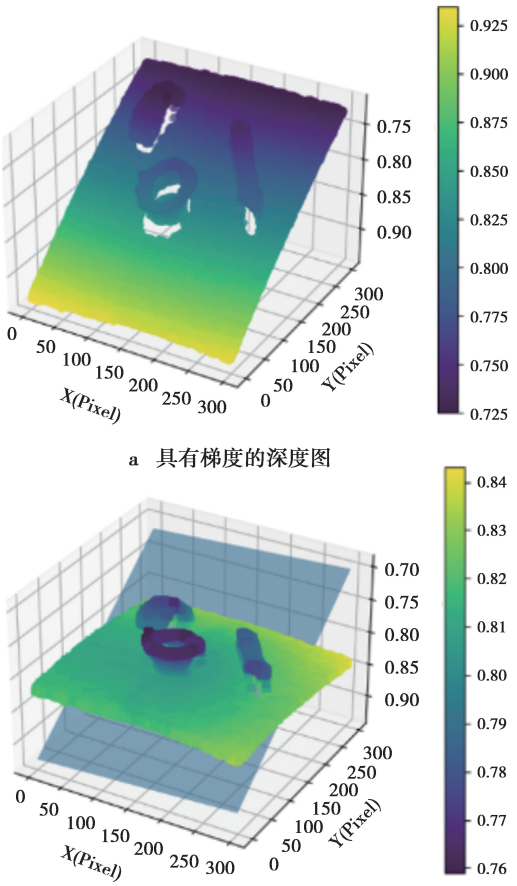
Fig.5　Single-target dataset grasp detection results

Cornell 数据集包含 885 组 RGB-D 图像。Cornell 数据集图像由于相机视角的原因,深度值分布具有过大梯度。由于本文所提方法对深度图像数据较为敏感,因此在进行 DGGS 时,对其深度图像做矫正处理,确保其深度图像深度值为以桌面为平面的垂直深度。首先将深度图像转换成点云 $\boldsymbol{P}$,然后使用随机抽样一致性(random sample consensus,RANSAC)算法对点云进行拟合得到平面 p,根据平面与垂直方向夹角绕点云质心 $\boldsymbol{c}$ 旋转点云 $\boldsymbol{P}$,公式为

$$p = RANSAC(\boldsymbol{P}) \tag{16}$$

$$\boldsymbol{P}' = \boldsymbol{R}(\boldsymbol{r},\theta)(\boldsymbol{P} - c) + c \tag{17}$$

式(17)中:$R(\boldsymbol{r},\theta)$ 代表由旋转轴 $\boldsymbol{r}$ 和 θ 计算出的旋转矩阵;$\boldsymbol{P}'$ 代表旋转后的点云。最后将旋转后的点云还原为深度图像,完成矫正处理。处理后的深度图像仅在 DGGS 过程中使用。对多目标数据集 Multi-Object 中的深度图像做相同处理,矫正前后的效果如图 6 所示。



a　具有梯度的深度图

b　深度矫正后的深度图

图 6　Multi-Object 数据集深度图像矫正前后效果

Fig.6　Multi-object dataset depth image correction effects before and after

由于Cornell数据集较小,可采用随机旋转、平移和裁剪的方式进行数据增强,因此本文同文献[4-5,17]等工作一样,采用五折交叉验证。同时使用图像分割(image-wise,IW)和对象分割(object-wise,OW)进一步测试方法的泛化性。不同抓取检测方法在Cornell数据集上的准确率和检测速度如表1所示。

表1 不同方法在Cornell数据集上的性能对比

Tab.1 Performance comparison of different methods on the Cornell dataset

算法	输入	准确率/%		用时/ms
		IW	OW	
SAE[4]	RGB-D	73.90	75.60	1350.0
GG-CNN[6]	D	73.00	69.00	19.0
ResNet-50x2[28]	RGB-D	89.21	88.96	103.0
GraspNet[29]	RGB-D	90.20	90.60	24.0
FCGN, ResNet-101[30]	RGB-D	97.70	96.61	117.0
GR-CNN[7]	RGB-D	97.70	96.70	20.0
TF-Grasp[17]	RGB-D	97.99	96.70	41.6
DSC-GraspNet[31]	RGB-D	98.30	97.70	14.0
本文算法	RGB-D	98.60	98.00	21.0

由表1可见,本文算法在准确率方面优于其他算法,在IW上达到了98.60%,在OW上达到了98.00%,验证了本文算法良好的泛化性。同时检测用时达到了21 ms,满足机械臂抓取任务的实时性要求。

Jacquard数据集包含5.4×10^4组RGB-D图像,其中90%用于训练,10%用于测试。为了测试所提方法的性能,与几种代表性的方法进行比较,它们在Jacquard数据集上的准确率如表2所示。本文算法取得了95.25%的准确率,优于其他方法。

表2 不同方法在Jacquard数据集上的准确率对比

Tab.2 Accuracy comparison of different methods on the Jacquard dataset

算法	输入模式	准确率/%
Jacquard[27]	RGB-D	74.20
GG-CNN[6]	D	84.00
FCGN, ResNet-101[28]	RGB-D	91.80
GR-CNN[7]	RGB-D	94.60
TF-Grasp[17]	RGB-D	94.60
DSC-GraspNet[31]	RGB-D	94.70
本文算法	RGB-D	95.25

3.4 多目标抓取位姿检测

为了进一步验证所提抓取检测方法的有效性,使用包含97组RGB-D图像的Multi-Object[32]和包含505组RGB-D图像的clutter[21]两个多目标数据集测试所提方法的性能,为了方便比较,同时使用多目标数据集训练GG-CNN,GR-CNN和DGM-Grasp,由于数据集较小,均采用数据增强。对Clutter数据集中3种不同的标注进行稀疏化处理至1/10,并格式化为Jacauard数据集的格式。当检测结果中至少有3个满足评估指标的抓取位姿时被认为检测准确,3种方法的准确率和检测用时如表3所示。

表3 不同方法在多目标数据集上的性能对比

Tab.3 Performance comparison of different methods on the multi-target dataset

算法	Multi-Object		Clutter	
	准确率/%	用时/ms	准确率/%	用时/ms
GG-CNN[6]	80	15	42.57	14
GR_CNN[7]	85	19	69.31	17
本文算法	95	24	81.19	21

3种方法的检测结果如图7所示。在多目标场景中,DGM-Grasp具有多尺度特征融合和注意力机制,能够有效地识别抓取特征,质量图贴合物体形状,角度和宽度准确。GG-CNN特征提取能力有限,难以完全区分物体和背景。GR-CNN能够分辨物体和背景,但是由于不具备注意力和多尺度特性,不能完全聚焦抓取特征,角度和宽度准确度一般。同时,二者生成的抓取检测结果中,存在无法抓取的抓取位姿,如Result图上黄色虚线圆圈所示,它们的夹指矩形处在物体本身或相邻物体上,夹指矩形的最小深度值明显等于或小于夹取空间的最小深度值,在抓取过程中夹指会与物体碰撞,导致抓取失败。而DGM-Grasp在DGGS的筛选重排作用下,虽然有时无法一次性为抓取区域内全部的物体都生成抓取位姿,但是避免了抓取阻碍的情况,保证每次抓取的成功率,从而提高实际抓取的整体成功率。

3.5 消融实验

为了验证各个模块对DGM-Grasp的贡献,本文进行了消融实验。以DGM-Grasp使用普通上采样、移除EMA注意力和DGGS后的模型作为基线模型,逐步增减模块,在Jacquard数据集上进行测试,部分消融实验方法的训练准确率如图8所示,全部结果

如表 4 所示。

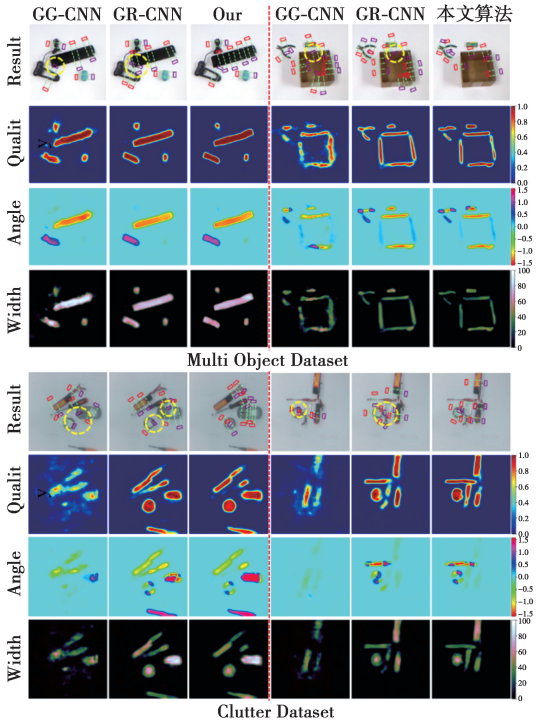


图 7　不同方法在多目标数据集上的检测结果

Fig.7　Detection results of different methods on the multi-target dataset

表 4　DGM-Grasp 消融实验结果

Tab.4　DGM-Grasp ablation experiment results

组合编号	模型组合				准确率/%	用时/ms
	基线模型	MCAD	AMFFD	DGGS		
1	√	×	×	×	93.19	55
2	√	√	×	×	94.48	56
3	√	×	√	×	94.66	56
4	√	×	×	√	93.32	55
5	√	√	√	×	95.14	57
6	√	√	×	√	94.69	56
7	√	×	√	√	94.81	57
8	√	√	√	√	95.25	58

从表 4 可知,MCAD 和 AMFFD 对所提方法的准确度提升较为显著;AMFFD 需多次上采样和下采样,较为耗时,而 MCAD 和 DGGS 对检测用时的负担较小;一并嵌入 MCAD、AMFFD 和 DGGS 的抓取检测方法,准确率达到了 95.25%,虽然检测用时最长,达到了 58 ms,但仍能保证机械臂抓取任务的实时性。

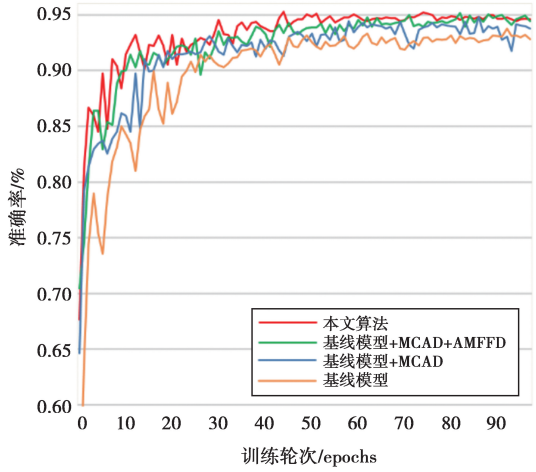


图 8　消融实验各方法训练准确率

Fig.8　Accuracy of training with various ablation methods

消融实验部分抓取检测效果如图 9 所示。相比于图 9a 中基线模型的抓取质量热力图,由于具有嵌入了 EMA 注意力的 MCAD 模块,图 9b-图 9d 均能够区分背景,聚焦抓取区域,并能够输出更准确的抓取宽度。由于 AMFFD 的融入,图 9c 和图 9d 与图 9b 相比,能够融合多层级特征图的语义信息。图 9d 与图 9c 相比,虽然热力图特征上几乎没有区别,但是由于 DGGS 的作用,输出的抓取位姿结果避免了图 9c 中的碰撞情况,抓取成功率更高。综上所述,各个模块都发挥了作用。

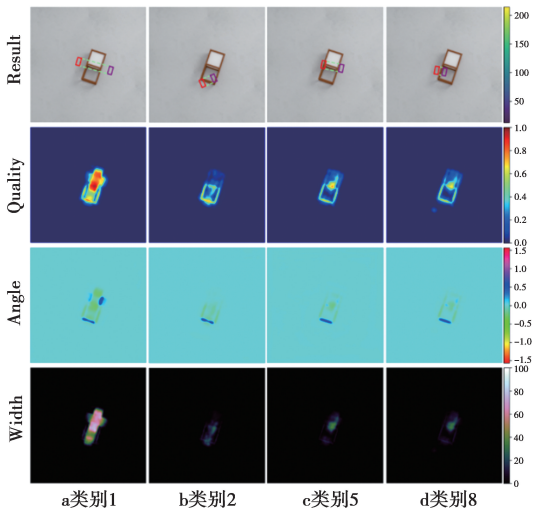


图 9　消融实验抓取检测效果

Fig.9　Ablation study on grasp detection performance

4　真实机械臂抓取实验

为了进一步验证 DGM-Grasp 在真实环境中的可用性和泛化性,参照文献[33]工作,本文搭建了

机器人抓取平台。实验平台由装有电动平行夹爪的 Baxter 机器人七自由度机械臂、垂直于桌面固定的 ORBBEC Astra pro RGBD 相机和抓取区域组成,如图 10 所示。

图 10 Baxter 机械臂抓取实验平台及物体

Fig.10 Baxter robot grasping experiment platform and targets

实验所选物体分为场景类型和形状类型,两者互有重合。场景类型包括胶带卷、刷子、指甲刀等生活类物体,扳手、钳子、三角架等工程类物体;形状类型为镂空、扁平、细长、怪异形状物体,能充分检验模型的性能和稳定性。对 20 种未知物体的 20 个组合进行真实抓取实验,每个组合至少在抓取区域随机放置 3 个未知物体。如果机器人能够将单个物体抓取并放置,就被认为当次抓取操作成功。真实机械臂抓取检测结果和抓取操作如图 11 所示。

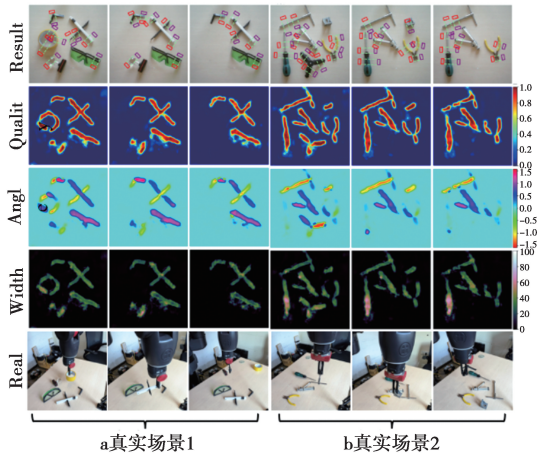


图 11 真实抓取实验结果

Fig.11 Real-world grasping experiment results

由图 11 可见,在真实场景下,DGM-Grasp 面对未知物体和环境,仍能够准确地识别出多目标物体的抓取特征,质量图贴合物体边缘,抓取角度和宽度准确,生成的抓取位姿均为实际可抓取位姿,保证了真实场景下的高准确率抓取,抓取成功率对比如表 5 所示。

表 5 不同方法真实抓取成功率对比

Tab.5 Comparison of real-world grasping success rates with different methods

方法	成功次数/实验次数	成功率/%
SAE[4]	89/100	89.0
GG-CNN[6]	110/120	92.0
GR-CNN[7]	334/350	95.4
TF-Grasp[17]	152/165	92.1
DGM-Grasp	193/200	96.5

实验进行总计 200 次抓取,DGM-Grasp 抓取成功率达到了 96.5%,明显高于其他抓取检测方法。抓取失败主要原因是面对部分物体时,抓取位置离物体质心较远,当物体重量较大时,容易脱落。

5 结束语

本文提出的 DGM-Grasp 能够聚焦物体的抓取特征,避免抓取阻碍,准确地预测出实际可抓的抓取位姿,在单目标数据集和多目标数据集上都取得了优秀的性能。DGM-Grasp 在真实抓取任务环境中进行测试,验证了其性能和泛化性。未来考虑在靠近物体质心抓取和复杂环境下抓取等方向继续研究。

参考文献:

[1] 刘亚欣,王斯瑶,姚玉峰,等.机器人抓取检测技术的研究现状[J].控制与决策, 2020, 35(12): 2817-2828.
LIU Y X, WANG S Y, YAO Y F, et al. Research status of robotic grasp detection technology[J]. Control and Decision, 2020, 35(12): 2817-2828.

[2] BOHG J, MORALES A, ASFOUR T, et al. Data-driven grasp synthesis—a survey[J]. IEEE Transactions on Robotics, 2013, 30(2): 289-309.

[3] FANG H S, WANG C, GOU M, et al. GraspNet-1Billion: A large-scale benchmark for general object grasping [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 11444-11453.

[4] LENZ I, LEE H, SAXENA A. Deep learning for detecting robotic grasps[J]. The International Journal of Robotics Research, 2015, 34(4-5): 705-724.

[5] REDMON J, ANGELOVA A. Real-time grasp detection using convolutional neural networks[C]//Proceedings of the IEEE International Conference on Robotics and Automation(ICRA). Seattle, WA, USA: IEEE, 2015: 1316-

1322.

[6] MORRISON D, CORKE P, LEITNER J. Learning robust, real-time, reactive robotic grasping[J]. The International journal of Robotics Research, 2020, 39(2-3): 183-201.

[7] KUMRA S, JOSHI S, SAHIN F. Antipodal robotic grasping using generative residual convolutional neural network [C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas, NV, USA: IEEE, 2020: 9626-9633.

[8] HE K, ZHANG X, REN S, et al.Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 770-778.

[9] GAO H, ZHAO J, SUN C. A real-time grasping detection network architecture for various grasping scenarios[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(8): 1-12.

[10] FU K, DANG X. Light-weight convolutional neural networks for generative robotic grasping[J]. IEEE Transactions on Industrial Informatics,2024,20(4):6696-6707.

[11] YU S, ZHAI D H, XIA Y. Skgnet: Robotic grasp detection with selective kernel convolution[J]. IEEE Transactions on Automation Science and Engineering, 2022, 20(4): 2241-2252.

[12] REN G, GENG W, GUAN P, et al. Pixel-wise grasp detection via twin deconvolution and multi-dimensional attention[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 4002-4010.

[13] ZHOU Z, ZHU X, CAO Q. AAGDN: Attention-augmented grasp detection network based on coordinate attention and effective feature fusion method[J]. IEEE Robotics and Automation Letters, 2023, 8(6): 3462-3469.

[14] YU S, ZHAI D H, XIA Y, et al. SE-ResUNet: A novel robotic grasp detection method[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 5238-5245.

[15] CHEN L, NIU M, YANG J, et al. Robotic grasp detection using structure prior attention and multiscale features [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2024, 54(11): 7039-7053.

[16] ZHANG Y, QIN X, DONG T, et al. DSNet: Double strand robotic grasp detection network based on cross attention[J]. IEEE Robotics and Automation Letters, 2024, 9(5): 4702-4709.

[17] WANG S, ZHOU Z, KAN Z.When transformer meets robotic grasping: Exploits context for efficient grasp detection[J]. IEEE robotics and automation letters, 2022, 7(3): 8170-8177.

[18] SAXENA A, DRIEMEYER J, NG A Y.Robotic grasping of novel objects using vision[J]. The International Journal of Robotics Research, 2008, 27(2): 157-173.

[19] JIANG Y, MOSESON S, SAXENA A. Efficient grasping from RGBD images: Learning using a new rectangle representation[C]//2011 IEEE International Conference on Robotics and Automation (ICRA). Shanghai, China: IEEE, 2011: 3304-3311.

[20] CAO H, CHEN G, LI Z, et al. Efficient grasp detection network with Gaussian-based grasp representation for robotic manipulation[J]. IEEE/ASME Transactions on Mechatronics, 2022, 28(3): 1384-1394.

[21] WANG D, LIU C, CHANG F, et al. High-performance pixel-level grasp detection based on adaptive grasping and grasp-aware network[J]. IEEE Transactions on Industrial Electronics, 2021, 68(12): 12345-12356.

[22] QIN X, HU W, XIAO C, et al.Attention-based efficient robot grasp detection network[J]. Frontiers of Information Technology & Electronic Engineering, 2023, 24(10): 1430-1444.

[23] YAN Y, TONG L, SONG K, et al. SISG-Net: Simultaneous instance segmentation and grasp detection for robot grasp in clutter[J]. Advanced Engineering Informatics, 2023(58): 102189.

[24] OUYANG D, HE S, ZHANG G, et al.Efficient multi-scale attention module with cross-spatial learning[C]// 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023: 1-5.

[25] YANG G, LEI J, ZHU Z, et al.AFPN: Asymptotic Feature Pyramid Network for Object Detection[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2023, 53(12): 2184-2189.

[26] LIU S, HUANG D, WANG Y. Learning spatial fusion for single-shot object detection[EB/OL]. (2019-11-21)[2024-08-07]. https://arxiv.org/abs/1911.09516

[27] DEPIERRE A, DELLANDREA E, CHEN L.Jacquard: A large scale dataset for robotic grasp detection[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid, Spain: IEEE, 2018: 3511-3516

[28] KUMRA S, KANAN C. Robotic grasp detection using deep convolutional neural networks[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Vancouver, BC, Canada: IEEE, 2017: 769-776.

[29] ASIF U, TANG J, HARRER S. GraspNet: An efficient convolutional neural network for real-time grasp detection

for low-powered devices[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI). Stockholm, Sweden: IJCAI, 2018: 4875-4882.

[30] ZHOU X, LAN X, ZHANG H, et al.fully convolutional grasp detection network with oriented anchor box[C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid, Spain: IEEE, 2018: 7223-7230.

[31] ZHOU Z, ZHANG X, RAN L, et al.DSC-GraspNet: A lightweight convolutional neural network for robotic grasp detection[C]//2023 9th International Conference on Virtual Reality (ICVR). Xi'an, China: IEEE, 2023: 226-232.

[32] CHU F J, XU R, VELA P A. Real-world multiobject, multigrasp detection[J]. IEEE Robotics and Automation Letters, 2018, 3(4): 3355-3362.

[33] 刘想德.基于视觉引导的 Baxter 机器人运动控制研究[J].重庆邮电大学学报(自然科学版),2018,30(4):552-557.
LIU X D. Research on motion control of baxter robot based on visual guidance[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2018, 30(4): 552-557.

作者简介:

刘想德,副教授,硕士生导师,主要研究方向为工业机器人、机电系统运动控制技术和计算机集成制造系统等。E-mail: liuxd@cqupt.edu.cn。

杨超旋,硕士研究生,主要研究方向为深度学习和机器人视觉。E-mail:1173528901@qq.com。

郑 凯,副教授,硕士生导师,主要研究方向为机电系统故障诊断及预测、机器人导航与控制等,E-mail:zhengkai@cqupt.edu.cn。

张 毅,教授,博士生导师,主要研究方向为智能系统与移动机器人、智能物流技术与装备等。E-mail:zhangyi@cqupt.edu.cn。

蒋 菲,硕士研究生,主要研究方向为机器人运动控制和机器调度。E-mail:13609444101@qq.com。

(**编辑**:张 勇)