中国激光

基于稳定轻量级网络的机器人抓取检测方法

徐志超[1]，薛俊鹏[1*]，孙鹏飞[2]，宋泽宇[1]，于长志[2]，卢文博[1]

[1]四川大学空天科学与工程学院，四川 成都 610065；

[2]中国工程物理研究院机械制造工艺研究所，四川 绵阳 621999

摘要 利用深度学习实现视觉图像的实时检测和姿态解算是引导机器人智能抓取的重要手段。针对机器人抓取检测中对准确性、实时性和稳定性的需求问题，构建了一种基于稳定轻量级网络的新型机器人抓取检测方法。将实例归一化层用于网络的卷积层和残差块中，每次只考虑单张图片的一个通道，不仅减少了运算量，还能有效利用单张图片的每个像素信息，在提升每个图像实例之间检测稳定性的同时加快模型收敛的速度；将特征金字塔网络分层结构融入上采样，并结合多维度语义信息增加对多尺度物体检测的准确度和稳定性。所构建的轻量级模型在交并比(IOU)为0.25时准确度为94.4%，画面传输速度为40.8 frame/s，并且在IOU低于0.5时准确度仍保持在80%以上。实验证明了在轻量级网络中加入实例归一化和特征金字塔的有效性。

关键词 机器视觉；机器人抓取；轻量级网络；物体检测；姿态检测

中图分类号 TP391 **文献标志码** A **DOI**：10.3788/CJL221397

1 引 言

复杂环境下的机器人柔性智能制造技术面临诸多挑战，如适应工作空间变化的快速响应、错误感知、噪声和控制的不准确性，以及抵抗机器人本身的扰动等[1-2]。其中，在真实工作场景中对多目标的抓取检测是机器人执行抓取的一个重要步骤，涉及抓取定位[3]和姿态解算[4]。由于机器人抓取检测的复杂性，近年来大部分研究也将关注点立足于用深度学习作为研究机器人抓取检测的重要方法，解决的问题在于提高机器人抓取检测的准确性、实时性和稳定性。

研究机器人实时检测和抓取的深度学习方法主要分为基于滑动窗口的方法、基于边界框的方法和像素级方法三类。基于滑动窗口的方法是用一个类似方框的分类器对整幅图像进行逐行扫描，对扫到的框内每个图像块进行概率评估，得到可以构成抓取的对象。Jiang等[5]提出了7维矩形抓取框，第一步使用线性分数函数进行快速搜索，将搜索矩形数量从数千万个减少到100个；第二步使用更复杂的特征来细化搜索，得到最优的抓取框。在前者研究的基础上，Lenz等[6]在滑动窗口方法中融入了深度学习的思想，在RGB-D数据中进行检测，第一个网络采用滑动窗口的方法，第二个网络则是一个分类器，评估滑动窗口中是否有最优的抓取框。虽然滑动窗口的方法可以有效进行抓取评估，但迭代次数多，响应速度较慢，难以满足实时性的需求。

基于边界框的方法在于先标记出数据集中的抓取框，训练神经网络，通过神经网络中的区域建议模块[7]得到可能的最优抓取，经回归模型进行修正，得到更加接近真实值的定向锚盒的中心点、宽、高以及基于水平位置的角度。Redmon等[8]在Krizhevsky等[9]研究的基础上，设计了可捕获的边界框，将抓取检测和目标分类结合在一起。Kumra等[10]在基于边界框方法的基础上，提出了多模态和单模态的网络架构，使用密集的50层残差网络ResNet[11]，以RGB图像为输入，取得了良好的实验结果。李正明等[12]还对抓取角度表示方法进行了优化，采用Faster R-CNN Inception-V2网络模型，把目标角度输出为分类后的标签形式，结合回归后的位置坐标，实现了对单个或多个目标准确的实时定位和姿态检测。虽然边界框对目标检测有积极的作用，但模型的体系结构复杂，需学习大量的模型参数。

基于像素级全卷积神经网络[13]的抓取方法相比于其他两种方法有两方面的优势：(1) 直接在图片的像素级基础上进行抓取预测[14]，生成抓取姿态；(2) 具有更轻量级的网络模型[15]，处理的参数更少。Morrison等[16]提出了一种基于U-net[17]的轻量级全卷积网络(FCN)，独立于对象的抓取综合模型，机器人抓取姿态可以从图像中按像素级生成，虽然最终结果表明该模型参数量少、实时性好，但精度较低。肖树国[18]在此方法的基础上选择图像分割和传统图像处理相结合的方法来进行目标检测，提出了改进的FCN模型，对分

收稿日期：2022-11-04；**修回日期**：2022-12-26；**录用日期**：2023-02-16；**网络首发日期**：2023-03-09

基金项目：教育部“春晖计划”合作科研项目(2020703-8)、四川省科技计划项目(2023YFG0181)

通信作者：*jpxue@scu.edu.cn

割图像的毛刺进行腐蚀剔除，使用中值滤波进行降噪解决视差图的空洞问题，提高了网络的实时性和抓取精度。Kumra 等[19]提出了一种新的生成残差卷积神经网络，可以在增加网络层数的情况下，抑制梯度消失和网络退化等问题，与 Morrison 等[16]的 FCN 相比，准确度更高。

机器人抓取检测需要在保持高准确度和实时快速性的条件下，同时降低参数量并保证其稳定性的问题。在上述研究的基础上，针对这一问题，本文首先在康奈尔数据集[5]上使用随机裁剪、缩放和旋转创建了一个增强数据集，接着选择基于 U-net 的轻量级网络结构并加入残差块增加提取特征的网络层数，同时抑制梯度消失和维度误差。使用实例归一化（IN）[20]来设计残差块和网络中每个卷积块的归一化层，增加检测图像实例的稳定性，加快模型收敛。之后，将特征金字塔网络（FPN）结构[21]思想整合到网络中，通过横向连接层，将自上而下和自下而上的特征图连接起来，融合多维度信息提升输出特征图的定位能力和语义信息。同时，采用 Huber 损失函数[22]分析计算结果，处理梯度爆炸的问题，避免异常值干扰。最后在康奈尔数据集上进行不同种类物体的测试，并增加了对未知物体的外部测试，证明了所提方法的有效性。

2 基本原理

2.1 抓取姿态表示

网络模型流程图如图 1 所示。其原理是 Morrison 等[23]提出的更加先进的抓取表示法，通过 RGB 彩色图像输入稳定轻量级网络后产生抓取姿态。

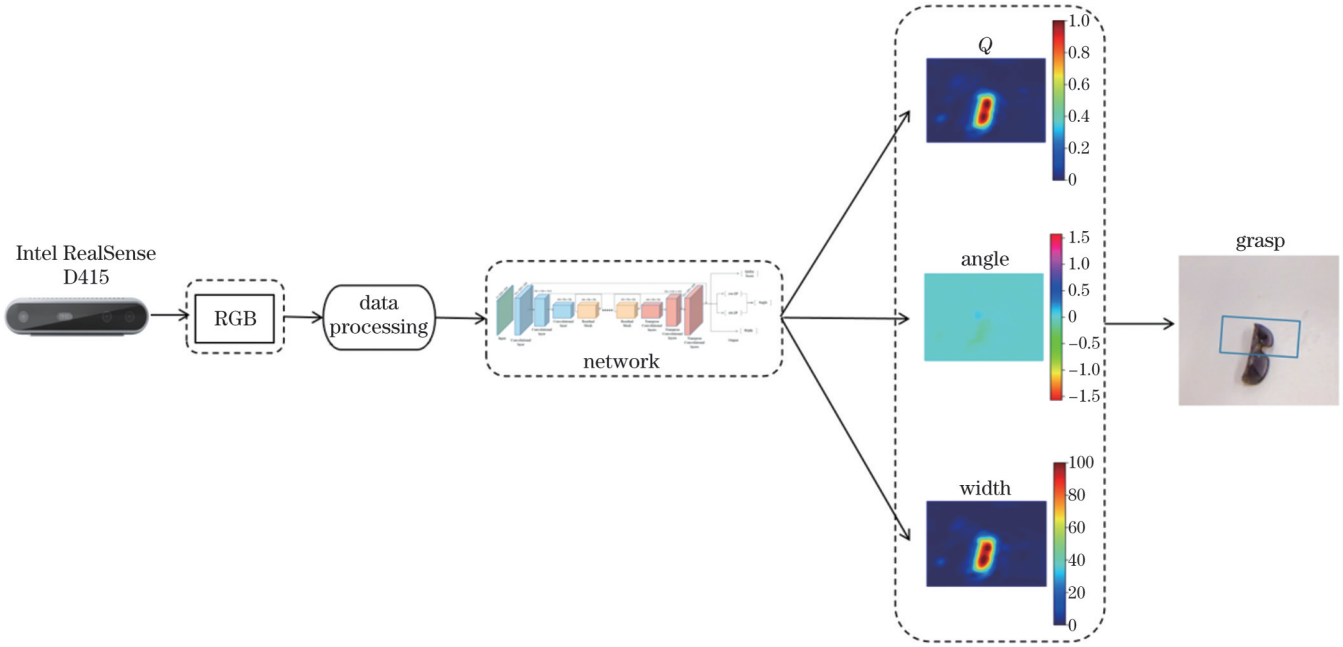

图 1 网络模型流程图

Fig. 1 Flow chart of network model

抓取姿态图如图 2 所示。定义一个垂直于 x-y 平面的抓取，本文通过一个固有参数的深度相机检测图像 $I=\mathbb{R}^{H\times W}$，其中 H 为高，W 为宽。用深度学习网络预测图像中每个像素，在生成的抓取图像中，一个抓取被描述为

$$X_i=\left(x,y,\phi_i,W_i,Q_i\right),\tag{1}$$

式中：$\left(x,y\right)$ 为图像坐标的像素中心点；i 为某例抓取框的标记；ϕ_i 代表机器人夹爪在 $\left(x,y\right)$ 处以水平方向为零度，绕 z 轴的旋转角度；W_i 为抓取物体时所需要的宽度的度量，并表示为像素范围内的值，可以通过使用深度相机的参数和测量的深度值转换成实验所需的物理量，$W_{\max}$ 是对机器人夹爪的最大宽度的描述，默认情况下为 150 pixel；Q_i 代表图像中每一点 $\left(x,y\right)$ 的抓取质量，表示为 0～1 之间的分数值，抓取成功率越大，分数值越高。

由于机器人夹爪是对足结构，在 $\pm\frac{\pi}{2}$ 弧度左右对称，因此角度 ϕ_i 的范围规定为 $\left[\frac{-\pi}{2},\frac{\pi}{2}\right]$。同时，为了使数据在 $\pm\frac{\pi}{2}$ 左右的取值唯一，且为了消除不连续性，本文将角度编码为一个单位向量的两个分量 $\cos 2\phi_i$ 和 $\sin 2\phi_i$，使网络能更好地学习该参数[24]。最后角度 ϕ_i 可以通过如下公式算出：

$$\phi_i=\frac{1}{2}\arctan\frac{\sin 2\phi_i}{\cos 2\phi_i}。\tag{2}$$

2.2 针对语义分割的网络设计

基于像素级神经网络的深度学习方法主要是运用语义分割对图像的每一个像素进行语义分类，属于生

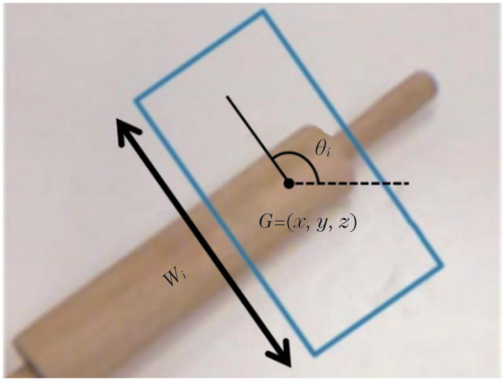

图 2　抓取姿态图

Fig. 2　Schematic of grasp posture

成架构，Ulyanov 等[20]关于图像风格迁移的研究同样如此。对此，本文的优化设计分为两块，即实例归一化和特征金字塔网络的应用。

首先，归一化是深度学习中不可缺少的模块，泛指把数据特征转换为相同尺度的方法，使大部分位置的梯度方向近似于基本指向最小值的最优搜索方向，提高训练效率。设 $x\in\mathbb{R}^{N\times C\times W\times H}$ 是包含一批 N 幅图像的输入张量，H 和 W 分别代表张量的高和宽，C 代表张量的通道数。在训练本文模型中属于生成架构的图像时，应该注重图像实例本身的每个像素。

传统的批量归一化（BN）一次考虑整批图像，会丢失单张图片的像素信息，具体公式如下：

$$\mu_r=\frac{1}{HWN}\sum_{t=1}^{N}\sum_{l=1}^{W}\sum_{m=1}^{H}x_{trlm},\tag{3}$$

$$\sigma_r^2=\frac{1}{HWN}\sum_{t=1}^{N}\sum_{l=1}^{W}\sum_{m=1}^{H}\left(x_{trlm}-\mu_r\right)^2,\tag{4}$$

$$x_{trjk}'=\frac{x_{trjk}-\mu_r}{\sqrt{\sigma_r^2+\varepsilon}},\tag{5}$$

$$y_{trjk}=\gamma x_{trjk}'+\beta,\tag{6}$$

式中：μ_r 为均值；σ_r^2 为方差；$trlm$（$trjk$）表示该张量的第 $trlm$（$trjk$）个元素，t 是批处理中图像的索引，r 是特征通道数，l 代表一张图片一个通道的宽，j 代表一个包含 T 张图片的输入张量的宽，m 代表一张图片一个通道的高，k 代表一个包含 T 张图片的输入张量的高。

实例归一化层如图 3 所示。IN 将归一化应用于单批图像，而不是整批图像，所以 IN 归一化公式中省去了 $\sum_{t=1}^{N}x_{trlm}$ 批量求和步骤，对单个像素的处理更有针对性。本文根据 Ulyanov 等[20]的实验结果，用 IN 替代卷积层和残差块中的 BN，沿着通道计算每张图的均值 μ 和方差 σ^2，对输入 x 做归一化，有 $x'=\left(x-\mu\right)/\sqrt{\sigma^2+\varepsilon}$，其中 x' 为归一化之后的输入，ε 为固定常量。之后，加入缩放和平移变量 γ 和 β 来调整最终尺度，最后输出 $y=\gamma x'+\beta$，提高本文的神经网络的性能。具体公式如下：

$$\mu_{tr}=\frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H}x_{trlm},\tag{7}$$

$$\sigma_{tr}^2=\frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H}\left(x_{trlm}-\mu_{tr}\right)^2,\tag{8}$$

$$x_{trjk}'=\frac{x_{trjk}-\mu_{tr}}{\sqrt{\sigma_{tr}^2+\varepsilon}},\tag{9}$$

$$y_{trjk}=\gamma x_{trjk}'+\beta。\tag{10}$$

对比以上公式中可以看出，BN 是对整批的所有像素求均值和标准差，所以每次的计算结果容易引入噪声，不稳定，而 IN 是对单批图像的所有像素求均值和标准差，信息源来源于图片本身，相当于全局信息的一次整合和调整，对于训练来说是更稳定的一种方法。

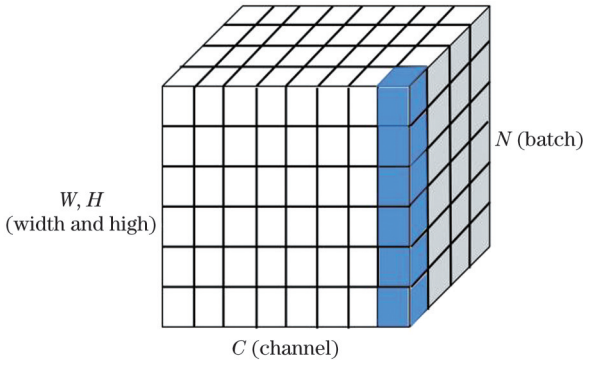

图 3　实例归一化层

Fig. 3　Instance normalization layer

其次，除可学习参数外，还存在众多影响网络性能的超参数。常见的超参数包括网络结构、优化参数以及正则化系数。本文基于 FPN 结构的设计属于网络结构方面的优化，因为卷积网络本身具有金字塔特征层次结构，就有低级到高级的语义，所以本文以单尺度图像作为输入，由于第一层卷积网络占用的内存很大，所以 FPN 从第二层卷积网络开始。首先是一个自下而上、步长为 2 的下采样，由于其下采样的次数少，因此该过程生成的特征图具有语义弱、分辨率高、易于目标定位的特点。然后是一个自上而下、步长同样为 2 的上采样，该过程生成的特征图具有语义强、分辨率低、易于目标识别的特点。最后加入 1×1 卷积层的横向连接层，将两种类型的特征图相结合。迭代此过程，生成不同维度定位能力强、语义信息丰富的分辨率映射。FPN 可以增强自上而下层的目标定位能力[21]。通常对于 FPN，自上而下的每层可以定义一个金字塔级别 P_Γ，其中 Γ 代表特征金字塔每层的编号。针对语义分割的特征金字塔网络如图 4 所示。通常最后一层具有最强的感受野，并且为了满足本文所设计的语义分割模型需要高分辨率图像，只选择最后一层 P_n 分辨率为 224×224 的特征图，n 代表 FPN 网络的最后一层。

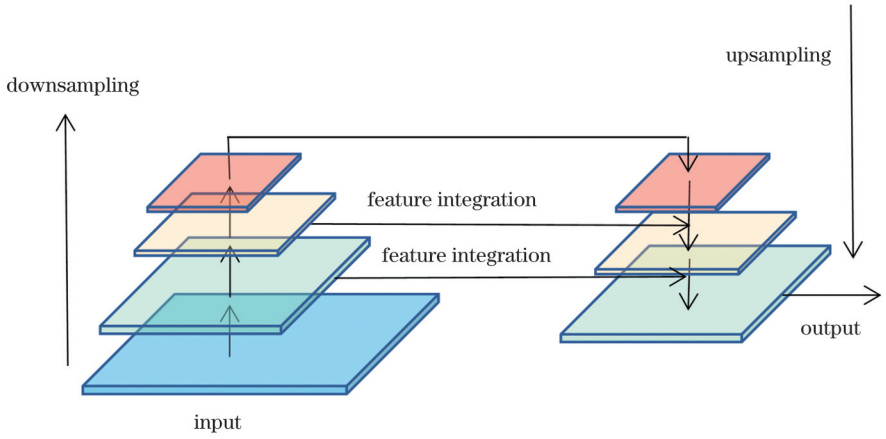

图 4　针对语义分割的特征金字塔网络

Fig. 4　Feature pyramid network for semantic segmentation

3　稳定轻量级网络检测方法

3.1　基于全卷积的稳定轻量级检测网络

本文网络模型如图 5 所示。输入是一个 3×224×224 的 RGB 色彩图像，前三个卷积层用于特征提取，由于本文只采用了 RGB 三通道，所以每个卷积层使用了如图 6(a)所示的 IN 层进行每张图片每个通道的归一化，充分利用单张图片的像素信息，并利用 ReLU 函数增强其非线性化能力。经过三层特征提取后，图像维度降为 64×56×56。众所周知，深度学习的精度会随着层数的增加而增加。但当网络模型超过一定量的层数时，会出现梯度消失和维数误差的问题，从而导致精度的饱和以及下降。本文基于实验测试结果加入了五层残差块，采用跳过连接的结构增加输入到输出的支路来解决这个问题。为了提升残差块主路的特征提取性能，同样加入了如图 6(b)所示的 IN 层。为了更容易地解析每个像素点，获得更高的分割准确率，在五层残差块后加入如图 6(c)所示的反卷积层进行上采样，并将相同维度的自上而下模块和自下而上模块横向连接，使高层信息可以融合浅层的边缘信息。利用这一网络结构，输出图像可以获得与输入图像同样的大小，并且定位能力强，语义信息丰富。

本文将输出结果分成四部分进行高斯滤波处理，分别为抓取质量分数、$\cos 2\phi_i$ 和 $\sin 2\phi_i$ 以及机器人夹爪的所需宽度。由于本文模型的结构，可以直接在原始图像上输出静态或者实时的抓取框。

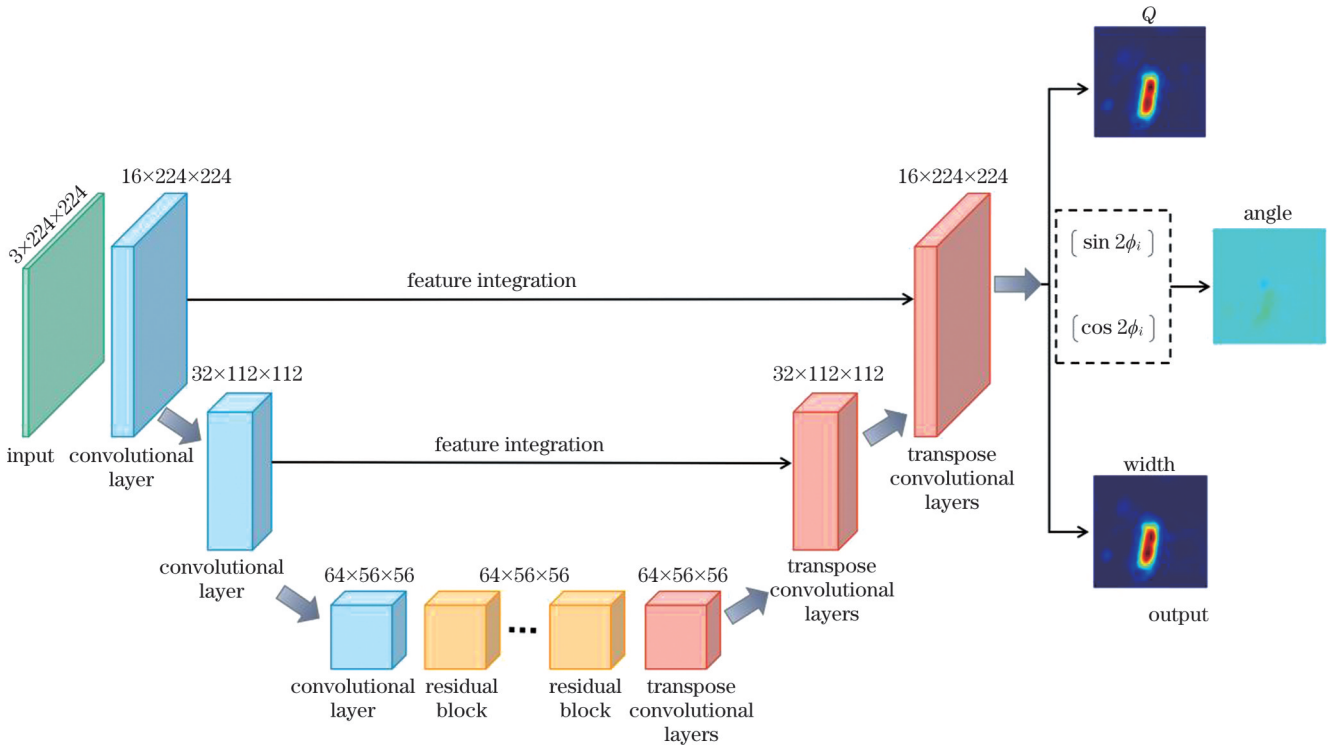

图 5　基于全卷积网络的高精度轻量级抓取网络

Fig. 5　High precision lightweight grasping network based on fully convolutional network

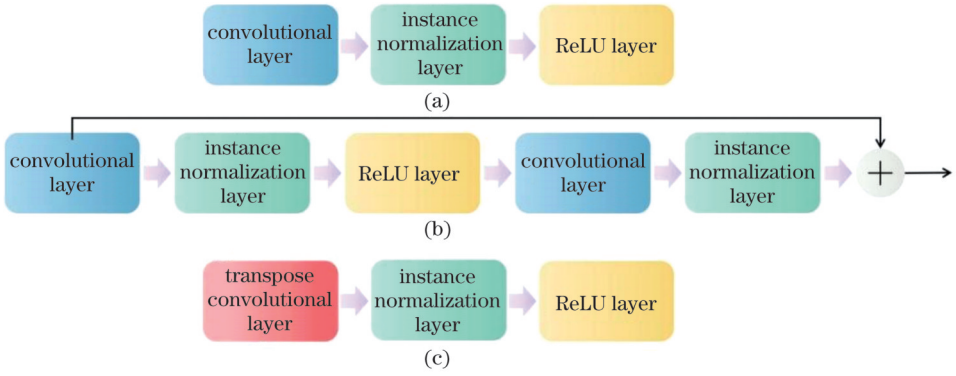

图 6 实例归一化示例。(a)卷积层;(b)残差块;(c)转置卷积层

Fig. 6 Examples of instance normalization. (a) Convolutional layer; (b) residual block; (c) transpose convolutional layer

3.2 训练方法

如果一个数据集中有样本 $A=\{A_1,\cdots,A_\varphi\}$、输入场景图像 $B=\{B_1,\cdots,B_\varphi\}$ 和图像中成功抓取 $X_i=\{x_1^1,\cdots,x_{\rho_1}^1,x_1^2,\cdots,x_{\rho_\varphi}^\varphi\}$,本文可以通过训练端到端的模型来学习一种映射函数 $\beta(B,A)=X_i$,对于成功抓取 X_i 的抓取质量分数 $Q_i\in[0,1]$,Q_i 越大则对应的 X_i 的负对数似然函数越小。因此训练目标应是映射函数可以使本文网络 X_i 的负对数似然函数最小,公式如下:

$$X_i=-\frac{1}{\varphi}\sum_{i=1}^{\varphi}\frac{1}{\rho_i}\sum_{b=1}^{\rho_i}\log_{10}\beta\left(x_i^b|B^i\right),\tag{11}$$

式中:φ 为样本数;ρ_i 为每个样本的成功抓取数。

使用 Adam 优化器[25]和标准反向传播神经网络对模型进行训练。学习速率设置为 7×10^{-4},使用批量大小为 8。

3.3 损失函数

通过分析不同损失函数在该网络下的运行结果,发现采用 Huber 损失函数[22]可有效处理梯度爆炸和异常值干扰的问题,并且在训练过程中具有较强的稳定性,可以获得最好的检测效果。

Huber 损失函数 S_α 表示如下:

$$S_\alpha=\begin{cases}0.5\left(G_\alpha-\overline{G_\alpha}\right)^2 & \text{if}\left|G_\alpha-\overline{G_\alpha}\right|<1\\ \left|G_\alpha-\overline{G_\alpha}\right|-0.5 & \text{otherwise}\end{cases},\tag{12}$$

式中:α 代表第 α 个抓取;G_α 是由网络生成的抓握;$\overline{G_\alpha}$ 是地面真值抓握。

本文中的总损失函数 L 定义为

$$L\left(G_\alpha,\overline{G_\alpha}\right)=\frac{1}{\varphi}\sum_0^\varphi L_\alpha,\tag{13}$$

$$L_\alpha=L_{\text{pos}}+L_{\sin}+L_{\cos}+L_W,\tag{14}$$

式中:L_{pos}、$L_{\sin}$ 和 $L_{\cos}$、L_W 均为 Huber 损失函数,分别对应机器人夹爪中心位置 (x,y)、旋转角度 ϕ_i、和张开宽度 W_i。

4 实验与结果分析

本文网络模型的运行在 Ubuntu 18.04.6 LTS 系统中进行。该系统运行的硬件条件为英特尔酷睿 i7-9700CPU(3.00 GHz×8)和 NVIDIA GeForce RTX 2080Ti(11 GB)显卡。软件平台为 CUDA 11,Python 3.6 版本,PyTorch 环境,PyCharm 编译平台,通过 OpenCV 实现结果可视化。

4.1 训练测试数据集

实验中的训练测试集采用康奈尔抓取数据集的扩展版本,包括 1035 张 RGB 图像,分辨率为 640 pixel×480 pixel,240 个不同的真实物体,5110 个正抓取和 2909 个负抓取。标注的地面真实值由几个抓取矩形表示每个对象的抓取可能性。但是康奈尔数据集是一个用于训练网络模型的小数据集。因此,通过随机裁剪、缩放和旋转,创建了一个增强数据集,产生有效的 51000 个抓取示例。在训练过程中,只考虑来自数据集中的正标记抓取。

4.2 抓取检测方法及结果评判标准

为了合理地评估本文的实验结果,参考了 Jiang 等[5]提出的矩形度量和 Morrison 等[16]的工作。抓取选择结果评判标准是首先通过识别 Q_i 中的 Q 值最大像素 s^* 确定图像空间中最好的机器人夹爪中心点位置坐标,然后计算相应的 ϕ_i 和 W_i,并将所得的输出抓取矩形框与地面真实值进行比较。当输出的抓取矩形框满足以下两个条件时,便认为生成的抓取选择是有效的:(1)地面真实的抓取矩形与预测的抓取矩形之间的交并比(交集与并集的比值,IOU)得分大于 25%;(2)地面真实的抓取矩形与预测的抓取矩形的抓取方向之间的角度偏移量小于 30°。

4.3 稳定性测试结果分析

稳定性是机器人抓取检测中的一个重要指标,决定了机器人是否能在复杂的动态环境中具有鲁棒性[23]。

用不同 IOU 的标准评价网络性能[26],并与其他网络模型(具体包括 Chu 等[7]、Morrison 等[16]和 Kumra 等[19]设计的网络)进行比较。各种方法在不同 IOU 标准下的准确度如表 1 所示。本文设计的网络在 IOU 为 0.4 时准确度最高。不同方法的稳定性对比如图 7 所

示。随着IOU标准的升高，本文设计的网络的准确度变化平稳，即使在IOU为0.5时，该网络的准确度也能达到80.9%，说明本文设计的网络性能更加稳定，具有一定的竞争力。

表1 各种方法在不同IOU标准下的准确度

Table 1 Accuracy of various methods under different IOU criteria

unit: %

IOU threshold	0.25	0.30	0.35	0.40	0.45	0.50
Chu *et al.* [7]	96.0	94.9	92.1	84.7	—	—
Morrison *et al.* [16]	87.6	84.2	82.5	79.7	72.3	70.0
Kumra *et al.* [19]	92.1	89.9	88.8	84.3	83.1	79.8
Ours	94.4	91.0	88.8	85.4	84.3	80.9

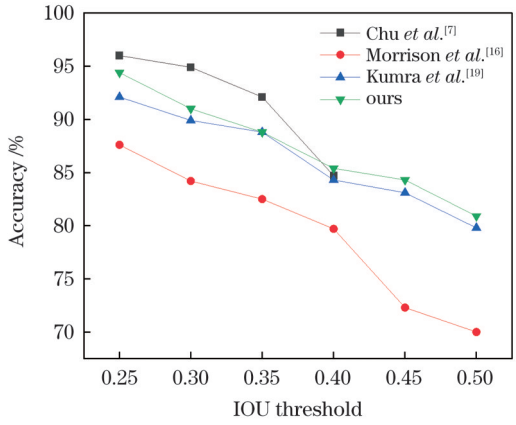

图7 各种方法的稳定性对比

Fig. 7 Comparison of stability among various methods

4.4 对比实验与分析

在机器人抓取的实时检测中，准确度和快速性无疑是衡量网络模型质量的两个重要因素[6]：准确度决定了机器人抓取时定位的能力，快速性则决定了所设计的神经网络是否可以进行机器人抓取检测的实时操作。此外，在资源有限的机器人系统下，保证取得同样的高准确度和良好的快速性所要求的参数量也是目前最新的评价标准之一[15]。表2列出了不同方法的准确度、速度和总参数量。本文设计的网络模型拥有94.4%的准确度、40.8 frame/s的速度和478900的总参数量。Chu等[7]设计的模型有最高准确度，采用50层残差块结构进行特征提取，前40层残差块输入到区域建议网络(RPN)结构进行特征提取，之后输出到两个全连接层，生成每个像素锚点的指定抓取建议和估计抓取框的抓取质量分数，然后通过感兴趣区域(ROI)池化[28]提取有效的抓取框，输入到后10层残差块，最后再通过两个全连接层输出五维抓取框。因其具有更多的残差块和分类回归处理，所以准确度达到了96.0%。本文模型的速度相比Chu等[7]的模型提高了4.9倍，总参数量减少了27705311。Morrison等[16]设计的网络模型仅有3个下采样卷积层和3个上采样卷积层，每个卷积层后都没有加入归一化层统一输入分布。实验结果表明，其速度比本文网络高2.4 frame/s，且参数量最少，但准确度只有87.6%。Kumra等[19]设计的方法模型有3个下采样卷积层、5个残差块和3个上采样卷积层，其归一化方式都是BN层，所以相比于本文网络计算量更大，速度低了2.8 frame/s，但本文模型融入了FPN进行横向连接的特征融合，因而参数量更大。Bergamini等[29]设计的网络基于比例块和残差块，比例块是含有1个IN步长为2的卷积层，残差块中是两部分含有IN的卷积层，最后输入通过跳跃连接与输出相连。整体框架采用FCN模型，第一部分通过5个比例块和2个残差块进行特征提取，第二部分通过2个比例块和1个卷积层预测每个锚点的得分，第三部分通过2个比例块和1个卷积层输出每个抓取矩形参数，速度上与本文模型只相差0.8 frame/s，但参数量大、准确度较低。总之，本文提出的网络模型达到了较高的准确度和速度，同时总参数量少，可以提高资源利用率。

表2 不同方法的准确度、速度和总参数量(IOU为0.25)

Table 2 Accuracy, speed and total parameter quantity of different methods (IOU is 0.25)

Method	Accuracy/%	Speed/(frame·s^{-1})	Total parameter quantity
Chu *et al.* [7]	96.0	8.3	28184211
Morrison *et al.* [16]	87.6	42.4	67604
Kumra *et al.* [19]	92.1	38.0	477860
Bergamini *et al.* [29]	87.1	40.0	10485760
Ours	94.4	40.8	478900

4.5 图像抓取识别检测

4.5.1 康奈尔抓取数据集上的评价结果

为了展现所提方法在现实场景中的性能，参考Bergamini等[29]的定性实验，将康奈尔数据集中的测试物体分成部分在图中的物体、半透明物体、反射性物体、分叉物体和形状不规则物体。图8为基于康奈尔抓取数据集的图像抓取识别定性检测结果，由抓取质量分数、角度和夹爪宽度组成的检测结果输出。从矩形抓取框的显示可以看出，本文抓取结果的Q_i值基本在0.9以上，Q值最大像素s^*对应于抓取中心的位置，相应的ϕ_i和W_i也符合真实的抓取情况。

4.5.2 外部测试

采用本文的模型对于未在康奈尔抓取数据集中出现的未知物体进行了和单目标抓取实时检测，结果如图9所示。本文所设计的网络对于反射性物体和部分在图中的不规则物体都能输出定位准确、姿态符合要求的抓取框。此外，如图9(c)所示对透明物体的实验结果所示，即使对于其他网络[19]不曾开展实时抓取检测实验的透明物体，本文所设计的网络也同样能输出

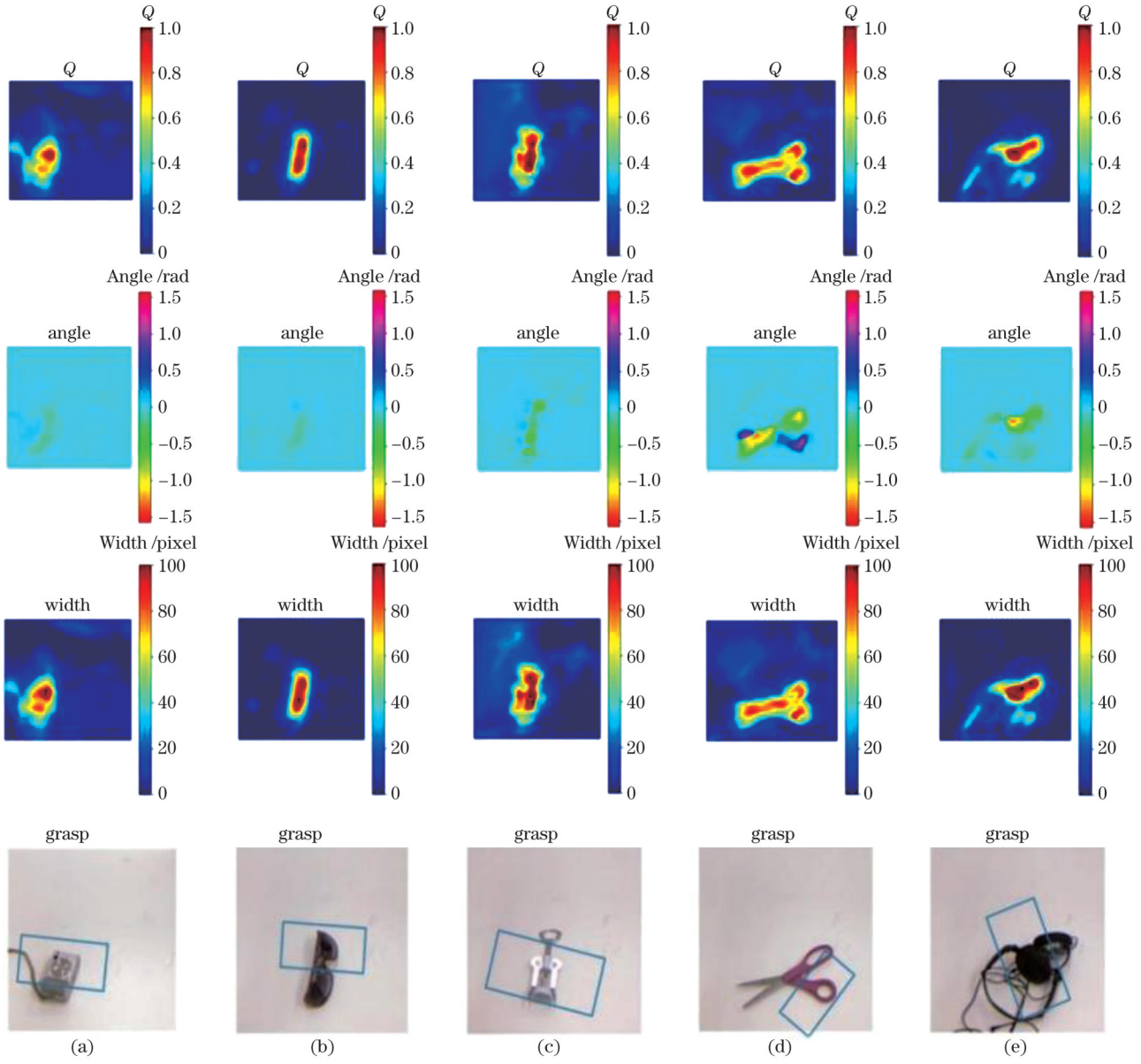

图 8 基于康奈尔抓取数据集的图像抓取识别定性检测结果。(a)部分在图中的物体;(b)半透明物体;(c)反射性物体;(d)分叉物体;(e)形状不规则物体

Fig. 8 Qualitative detection results of image grasping recognition based on Cornell grasping dataset. (a) Object partially in figure; (b) translucent object; (c) reflective object; (d) bifurcated object; (e) irregularly shaped object

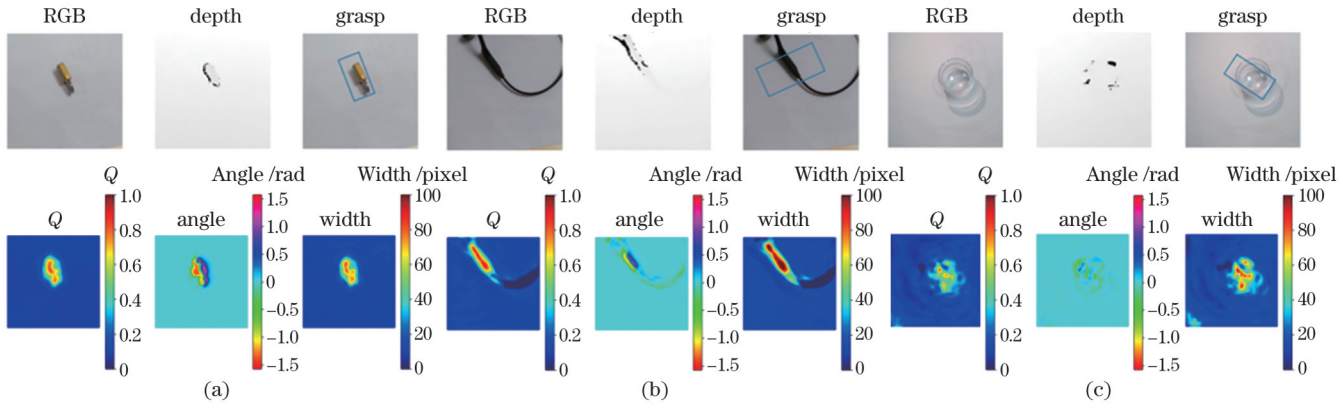

图 9 未知物体检测的单目标抓取实时检测。(a)反射性物体;(b)部分在图中的不规则物体;(c)透明物体

Fig. 9 Real-time detection for single-target grasp of unknown object detection. (a) Reflective object; (b) irregular object partially in figure; (c) transparent object

符合真实抓取情况的定位准确、姿态符合要求的抓取框，打破了目前深度相机无法准确测量透明物体的限制[30]。

5 结 论

本文提出了一种基于稳定轻量级网络的机器人抓取检测方法。首先，将IN加入到网络的卷积层和残差块中，解算单张图片中每个像素的抓取矩形，提升检测每个图像实例之间的稳定性，加速小批量模型的收敛速度；其次，融入FPN的思想，增强最终尺度特征图上的语义信息和定位能力，从而使网络检测目标更加精准和稳定；最后，采用Huber损失函数训练网络并在康奈尔数据集进行评估。定量结果表明，该方法的总参数量为478900，稳定性和已有的方法相比，可以在IOU小于0.5时都保持80%以上的准确度，且模拟曲线平稳。同时，在IOU为0.25的公认标准下，该方法具有94.4%的高准确度以及40.8 frame/s的速度。定量结果说明本文方法可以兼顾目前对机器人抓取检测的准确性、实时性和稳定性的要求。定性结果表明，本文设计的网络模型对不同属性的物体均有效果。综上，本文研究表明在轻量级网络中加入IN和FPN，能够有效提升机器人抓取检测的性能。

参 考 文 献

[1] 张磊，徐孝彬，曹晨飞，等. 基于动态特征剔除的图像与点云融合的机器人位姿估计方法[J]. 中国激光，2022，49(6)：0610001.
Zhang L，Xu X B，Cao C F，et al. Robot pose estimation method based on image and point cloud fusion with dynamic feature elimination[J]. Chinese Journal of Lasers，2022，49(6)：0610001.

[2] 崔海华，漏华铖，田威，等. 轨道式爬行机器人制孔基准的视觉高精度定位[J]. 光学学报，2021，41(9)：0915002.
Cui H H，Lou H C，Tian W，et al. High-precision visual positioning of hole-making datum for orbital crawling robot[J]. Acta Optica Sinica，2021，41(9)：0915002.

[3] 黄会明，刘桂华，段康容. 基于微振镜结构光投射器的机器人抓取[J]. 中国激光，2019，46(2)：0204002.
Huang H M，Liu G H，Duan K R. Robot bin-picking based on micro-electromechanical system structure light projector[J]. Chinese Journal of Lasers，2019，46(2)：0204002.

[4] 杜学丹，蔡莹皓，鲁涛，等. 一种基于深度学习的机械臂抓取方法[J]. 机器人，2017，39(6)：820-828，837.
Du X D，Cai Y H，Lu T，et al. A robotic grasping method based on deep learning[J]. Robot，2017，39(6)：820-828，837.

[5] 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，May 9-13，2011，Shanghai，China. New York：IEEE Press，2011：3304-3311.

[6] Lenz I，Lee H，Saxena A. Deep learning for detecting robotic grasps[J]. International Journal of Robotics Research，2015，34(4/5)：705-724.

[7] Chu F J，Xu R N，Vela P A. Real-world multiobject，multigrasp detection[J]. IEEE Robotics and Automation Letters，2018，3(4)：3355-3362.

[8] Redmon J，Angelova A. Real-time grasp detection using convolutional neural networks[C]//2015 IEEE International Conference on Robotics and Automation (ICRA)，May 26-30，2015，Seattle，WA，USA. New York：IEEE Press，2015：1316-1322.

[9] Krizhevsky A，Sutskever I，Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM，2017，60(6)：84-90.

[10] Kumra S，Kanan C. Robotic grasp detection using deep convolutional neural networks[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)，September 24-28，2017，Vancouver，BC，Canada. New York：IEEE Press，2017：769-776.

[11] He K M，Zhang X Y，Ren S Q，et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，June 27-30，2016，Las Vegas，NV，USA. New York：IEEE Press，2016：770-778.

[12] 李正明，章金龙. 基于深度学习的抓取目标姿态检测与定位[J]. 信息与控制，2020，49(2)：147-153.
Li Z M，Zhang J L. Detection and positioning of grab target based on deep learning[J]. Information and Control，2020，49(2)：147-153.

[13] Zhang H，Duan D Y. Computational ghost imaging with compressed sensing based on a convolutional neural network[J]. Chinese Optics Letters，2021，19(10)：101101.

[14] 邵斌，杨华，朱斌，等. 基于实时语义分割的红外小目标检测算法[J]. 激光与光电子学进展，2023，60(14)：1410006.
Shao B，Yang H，Zhu B，et al. Infrared small target detection algorithm based on real-time semantic segmentation[J]. Laser & Optoelectronics Progress，2023，60(14)：1410006.

[15] 马倩倩，李晓娟，施智平. 轻量级卷积神经网络的机器人抓取检测研究[J]. 计算机工程与应用，2020，56(10)：141-148.
Ma Q Q，Li X J，Shi Z P. Research on light-weight convolutional neural network for robotic grasp detection[J]. Computer Engineering and Applications，2020，56(10)：141-148.

[16] Morrison D，Leitner J，Corke P. Closing the loop for robotic grasping：a real-time，generative grasp synthesis approach[EB/OL]. [2022-10-05]. http://www.roboticsproceedings.org/rss14/p21.pdf.

[17] Ronneberger O，Fischer P，Brox T. U-net：convolutional networks for biomedical image segmentation[M]//Navab N，Hornegger J，Wells W M，et al. Medical image computing and computer-assisted intervention-MICCAI 2015. Lecture notes in computer science. Cham：Springer，2015，9351：234-241.

[18] 肖树国. 基于深度学习的目标抓取位姿确定方法研究[D]. 哈尔滨：哈尔滨工程大学，2019.
Xiao S G. Position and attitude determination based on deep learning for object grasping[D]. Harbin：Harbin Engineering University，2019.

[19] 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)，October 24-January 24，2021，Las Vegas，NV，USA. New York：IEEE Press，2021：9626-9633.

[20] Ulyanov D，Vedaldi A，Lempitsky V. Improved texture networks：maximizing quality and diversity in feed-forward stylization and texture synthesis[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，July 21-26，2017，Honolulu，HI，USA. New York：IEEE Press，2017：4105-4113.

[21] Lin T Y，Dollár P，Girshick R，et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，July 21-26，2017，Honolulu，HI，USA. New York：IEEE Press，2017：936-944.

[22] Meyer G P. An alternative probabilistic interpretation of the Huber loss[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)，June 20-25，2021，Nashville，TN，USA. New York：IEEE Press，2021：5257-5265.

[23] Morrison D，Corke P，Leitner J. Learning robust，real-time，reactive robotic grasping[J]. International Journal of Robotics Research，2020，39(2/3)：183-201.

[24] Hara K，Vemulapalli R，Chellappa R. Designing deep convolutional neural networks for continuous object orientation

estimation[EB/OL]. (2017-02-06) [2022-10-08]. https://arxiv.org/abs/1702.01499.

[25] Kingma D P, Ba J. Adam: a method for stochastic optimization [EB/OL]. (2014-12-22) [2022-10-09]. https://arxiv.org/abs/1412.6980.

[26] 郝宸, 田瑾, 韩华, 等. 基于注意力机制的实时性抓取检测算法[J]. 传感器与微系统, 2022, 41(1): 131-134.
Hao C, Tian J, Han H, et al. Real-time grab detection algorithm based on attention mechanism[J]. Transducer and Microsystem Technologies, 2022, 41(1): 131-134.

[27] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence, June 6, 2016, New York: IEEE Press, 2016: 1137-1149.

[28] Jiang B R, Luo R X, Mao J Y, et al. Acquisition of localization confidence for accurate object detection[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11218: 816-832.

[29] Bergamini L, Sposato M, Pellicciari M, et al. Deep learning-based method for vision-guided robotic grasping of unknown objects [J]. Advanced Engineering Informatics, 2020, 44: 101052.

[30] Shi C Q, Miao C X, Zhong X G, et al. Pixel-level grasp detection for unknown objects with encoder-decoder-inception deep network [C]//2022 IEEE 5th International Conference on Electronics Technology (ICET), May 13-16, 2022, Chengdu, China. New York: IEEE Press, 2022: 1153-1157.

Robot Grasp Detection Method Based on Stable Lightweight Network

Xu Zhichao[1], Xue Junpeng[1*], Sun Pengfei[2], Song Zeyu[1], Yu Changzhi[2], Lu Wenbo[1]

[1]*School of Aeronautics and Astronautics, Sichuan University, Chengdu* 610065, *Sichuan, China*;

[2]*Institute of Machinery Manufacturing Technology, China Academy of Engineering Physics, Mianyang* 621999, *Sichuan, China*

Abstract

Objective In many complex environments, intelligent manufacturing technology using robots faces many challenges. There are rapid response to changes in the workspace, false perception, error of noise and control, and even resistance to disturbances of the robot itself. The grasp detection of target in the real working scene is an important step in robot workflow, which involves grasp positioning and attitude estimation (Fig. 1). Due to the complexity of robot grasp detection, in recent years extensive research has committed to using deep learning method to achieve grasp detection. The objective is to improve the accuracy, timeliness and stability of robot grasp detection. Deep learning methods for real-time robot detection and grasp are divided into three main categories. They are sliding window-based, bounding box-based, and pixel-level methods. The sliding window-based method is effective for grasp detection. But the large number of iterations and the slow response speed make it difficult to meet the real-time requirements. Bounding box-based methods have a positive effect on target detection. But the structure of the model is complex and a large quantity of parameters should be learnt. In the present study, we report a robot grasp detection method based on a stable pixel-level lightweight network to reduce the number of parameters. We hope that our method can save more computation and memory during robot grasp detection. And it can ensure the high accuracy, timeliness and stability under complex conditions.

Methods Firstly, an enhanced dataset is created by using random cropping, scaling, and rotation on the Cornell dataset (Fig. 2). Then, we build a lightweight network based on U-net (Fig. 5). Residual blocks are added to increase the number of extracted feature layers in the network. They can suppress the gradient vanishing and dimensional error (Fig. 6). At the same time, instance normalization (IN) is used to design the normalization layer of the residual block and each convolutional block in the network (Fig. 3). It will increase the stability of the detected image instance and accelerates the model convergence (Fig. 6). In addition, the network integrates the idea from the feature pyramid network (FPN) structure (Fig. 4). We connect the top-down feature map with the bottom-up feature map through the horizontal connection layer. The former has strong semantics, low resolution and easy target recognition. The latter has weak semantics, high resolution and easy target localization. Meanwhile, the multi-dimensional information is integrated to improve the localization ability and semantic information of the output feature map (Fig. 5). Finally, in order to avoid the problem of gradient explosion and outlier interference, our study uses the Huber loss function to analyze the calculation results.

Results and Discussions In this study, all comparison methods uniformly use the Cornell dataset. And we use different intersection-over-union (IOU) standards to evaluate the network performance (Table 1). As the IOU standard becomes more strict, our network changes smoothly (Fig. 7). When the IOU is 0.5, the accuracy of our network can still reach 80.9%. Therefore, our network performance is more stable and competitive in robustness to complex environments. When the IOU is 0.25, the network model designed in this study has 94.4% accuracy, 40.8 frame/s speed and 478900 total parameter quantity (Table 2). The results show that our method can achieve high accuracy and speed when the total parameter quantity keeps small. The model is tested using random objects in the Cornell dataset (Fig. 8). And we also perform real-time detection of unknown objects (Fig. 9). The experiment shows that the network designed can output the grasp frame with accurate positioning and posture information for objects with different properties, including objects partially in the figure, translucent objects, reflective objects, bifurcated objects, irregularly shaped objects, and even transparent objects.

Conclusions This study is based on simultaneously achieving high accuracy and timeliness of robot grasp detection, reducing the

number of system parameters and ensuring the stability of the network. Referring to the pixel-based U-net lightweight network model, we design a robot grasp detection method based on stable lightweight network. Firstly, the quantitative results show that the simulated curve is stable and the total parameter quantity of the method is 478900. Our method can maintain more than 80% accuracy when the IOU is less than 0.5, outperforming the existing methods. At the same time, when the IOU is 0.25, it has a high accuracy of 94.4% and a speed of 40.8 frame/s. It is rational that this method can meet the current requirements for the accuracy, timeliness and stability of robot grasp detection. Moreover, the qualitative results show that the network model designed in this study is effective on objects with different properties. Especially, our method is not disturbed by the situation that the current depth camera cannot accurately measure transparent objects. Finally, it is shown that adding IN and FPN to the lightweight network can effectively improve the performance of robot grasp detection.

Key words machine vision; robot grasp; lightweight network; object detection; posture detection