北京航空航天大学学报
Journal of Beijing University of Aeronautics and Astronautics
ISSN 1001-5965,CN 11-2625/V

《北京航空航天大学学报》网络首发论文

题目：基于深度学习的视觉检测及抓取方法
作者：孙先涛，程伟，陈文杰，方笑晗，陈伟海，杨茵鸣
DOI：10.13700/j.bh.1001-5965.2022.0130
收稿日期：2022-03-08
网络首发日期：2022-05-26
引用格式：孙先涛，程伟，陈文杰，方笑晗，陈伟海，杨茵鸣．基于深度学习的视觉检测及抓取方法[J/OL]．北京航空航天大学学报．
https://doi.org/10.13700/j.bh.1001-5965.2022.0130

基于深度学习的视觉检测及抓取方法

孙先涛[1]，程伟[1]，陈文杰[1]，方笑晗[1]，陈伟海[2,*]，杨茵鸣[1]

(1.安徽大学 电气工程与自动化学院，合肥 230601；2.北京航空航天大学 自动化科学与电气工程学院，北京 100191)
*通信作者 E-mail： whchen@buaa.edu.cn

摘要 针对现有机器人抓取系统对硬件设备要求高、难以适应不同物体及抓取过程产生较大有害扭矩等问题，本文提出了一种基于深度学习的视觉检测及抓取方法。采用通道注意力机制对 YOLO-V3 进行改进，增强网络对图像特征提取的能力，提升复杂环境中目标检测的效果，平均识别率较改进前增加 0.32%。此外，针对目前姿态估计角度存在离散性的问题，提出一种基于 VGG-16 主干网络嵌入最小面积外接矩形(MABR)算法，进行抓取位姿估计和角度优化。改进后的抓取角度与目标实际角度平均误差小于 2.47°，大大降低两指机械手在抓取过程中对物体所额外施加的有害扭矩。本文利用 UR5 机械臂、气动两指机械手、Realsense D435 相机及 ATI-Mini45 六维力传感器等设备搭建了一套视觉抓取系统，实验表明本方法可以有效地对不同物体进行抓取分类操作、对硬件的要求较低、并且将有害扭矩降低约 75%，从而减小对物体的损害，具有很好的应用前景。

关键词 深度学习；神经网络；目标检测；姿态估计；机器人抓取
中图分类号 TP242 文献标志码 A DOI: 10.13700/j.bh.1001-5965.2022.0130

A visual detection and grasping method based on deep learning

SUN Xiantao[1], CHENG Wei[1], CHEN Wenjie[1], FANG Xiaohan[1], CHEN Weihai[2,*], YANG Yinming[1]

(1. School of Electrical Engineering and Automation, Anhui University, Hefei, 230601, China;

2. *School of Automation Science and Electrical Engineering, Beihang University, Beijing, 100191, China)*
*E-mail: whchen@buaa.edu.cn

Abstract To solve the problems of high hardware cost, difficult to adapt to different objects, and large harmful torque during grasp for the existing robotic grasping systems, this paper proposes a deep learning based visual detection and grasping research. The channel attention mechanism is used to enhance the ability of the network to extract image features and improves the effect of target detection in complex environments using the improved YOLO-V3. It is found that the average recognition rate is increased by 0.32% compared with that before the improvement. In addition, for the discrete problem of estimated orientation angle, an embedded minimum area bounding rectangle (MABR) algorithm based on VGG-16 backbone network is proposed to estimate and optimize the grasping position and orientation. The average error between the improved predicted grasping angle and the actual angle of the target is less than 2.47°. It greatly reduces the additional harmful torque applied by the two finger gripper to the object in the grasping process. This article uses a UR5 robotic arm, a pneumatic two-finger robotic gripper, a Realsense D435 camera and an ATI-Mini45 six-axis force/torque sensor to build a visual grasping system. Experimental results show that this method can effectively grasp and classify objects, has low requirements for hardware, and reduces harmful torque by about 75%, thereby reducing damage to the grasped objects, and has a good application prospect.

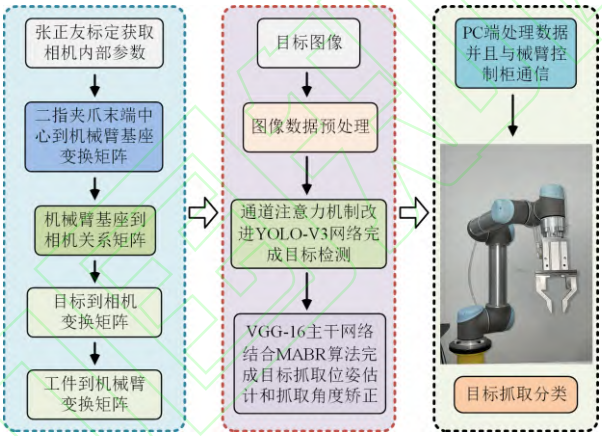
Key words deep learning; neural network; object detection; pose estimation; robotic grasping

收稿日期：2022-03-08
基金项目：国家自然科学基金（52005001）.
作者简介：孙先涛， 男，副教授，硕士生导师。研究方向：机器人抓取，欠驱动机械手，精密操作。程伟， 男，硕士研究生。主要研究方向：机器视觉，深度学习，机器人抓取。陈文杰， 男，教授，博士生导师。主要研究方向：智能抓取，欠驱动机械手，外骨骼智能控制。陈伟海， 男，教授，博士生导师。主要研究方向：机器人视觉，机器人抓取，高精密运动机械设计与控制。

Fund： National Natural Science Foundation of China (52005001)
网络首发时间： 2022-05-26 15:42:01 网络首发地址： https://kns.cnki.net/kcms/detail/11.2625.V.20220525.1827.003.html

随着科学技术不断进步，机器人在全球迎来了新的发展浪潮，占有越来越大的市场规模，在诸多行业中扮演重要角色。尤其在工业领域中的工件抓取分类、装配等重复性运动较多的场景，机器人被广泛应用。与传统人工相比，机器人具有较高准确性、稳定性和投资回报率的优势。近些年，在人工智能快速崛起和智能硬件不断迭代的基础上，计算机视觉和机器人紧密地联系在一起，机器人可通过相机这样的“眼睛”来获取物体图像信息，从而实现了与外界环境的交互。

目前，深度学习被广泛应用于各行各业，结合深度学习的智能抓取也成为国内外研究的热门领域[1]。相比于传统固定点、手动示教或利用简单视觉识别定位的抓取方法[2,3]，结合深度学习的机器人抓取具有更高的准确率、随机性和应用价值。Mallick 等[4]通过深层卷积网络语义分割法实现物体的检测和定位，利用机械臂完成物体的分拣工作。白成超等[5]通过改进的 YOLO(You Only Look Once)算法实现目标检测，实现机械臂的抓取动作。黄怡蒙等[6]对 Tiny-YOLOV3 目标检测的结果进行三角函数转换，并控制机械臂完成物体抓取。以上学者仅利用目标检测的方法完成物体抓取，并没有获取物体有效抓取点的位姿，抓取具有一定的局限性。Jiang 等[7]通过两步走模型框架，使用支持向量机(Support Victor Machine, SVM)排序算法预测物体的抓取点和角度。Chu 等[8]通过 ResNet-50 主干网络结合抓取建议框图实现物体抓取位姿预测。夏浩宇等[9]提出了基于 Keypoit RCNN 改进模型的抓取检测算法，实现对管纱的有效抓取。后者相比于前者提高了抓取成功率，但是存在预测抓取角度离散的问题，导致机械手容易与物体产生偏角，在抓取过程中容易改变物体当前状态甚至造成物体损坏，具有一定的干扰性。



针对上述问题，从视觉检测及机器人抓取工作实际需求出发，本文结合了目标检测和抓取位姿估计算法，并对目标检测和抓取角度进行改进，提高了机器人抓取物体的准确性和稳定性。整个抓取系统如下图 1 所示：1) 准备阶段，首先利用张正友标定法[10]获取相机的内部参数，接着通过探针法设置两指机械手末端中心，然后通过手眼标定获取机械臂和相机的坐标转换关系矩阵；2) 图像处理阶段，计算机首先对目标图像进行预处理，接着将处理后的数据传入到两个通道中：通道一采用通道注意力模块改进的 YOLO-V3[11]对物体进行目标检测；通道二采用 VGG-16[12]主干网络和最小面积外接矩形(Minimum Area Bounding Rectangle, MABR)[13]算法对物体的抓取位姿进行预测和抓取角度连续化矫正；3) 控制阶段，PC 端与控制柜建立通信，并发送抓取点坐标和机械手偏转角度信息，进行抓取分类动作。

图 1 抓取系统图

Fig1. Grasping system diagram

1 目标检测

相比于传统基于模型或人工标签式的目标检测技术[14]，现阶段的目标检测算法结合深度学习的优势在识别准确率和运行速度上得到了极大提高[15]，也更加满足当前机器人抓取和分类放置操作的工作需求。自 2014 年 Girshick 等[16]提出基于区域的卷积神经网络(Region proposals Convolutional Neural Networks, R-CNN)以来，该方向的目标检测算法不断地被改进，出现了以 Fast R-CNN[17]和 Faster R-CNN[18]为代表的先通过区域推荐再进行目标分类的两步走目标检测算法、以及以 YOLO[19]为代表的采用一个网络直接进行预测输出的目标检测算法等。

1.1 YOLO-V3 模型

自从 2016 年 Redmon 等[20]提出 YOLO 算法以来，此算法不断地被优化，准确率也在逐步提升。该算法结合了候选区域调整和网络预测结果优化两个步骤，有端到端的网络结构特性，具有输入一张图像直接输出预测结果的功能，其最大特点是整个网络的运行速度很快。2018 年 Redmon 等[11]再次提出了 YOLO-V3 算法，网络结构如图 2 所示，达到了兼顾检测准确率和实时性的效果。该算法引入了多尺度预测模块，对象分类器由以前 softmax 函数改为 logistic 函数进行输出预测，并且借鉴特征金字塔网络(Feature Pyramid Network, FPN)的思想来对小、中、大物体预测。

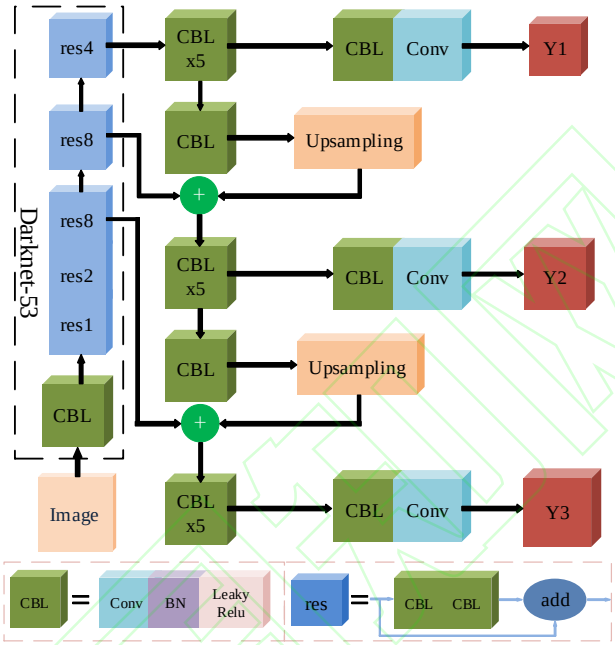


图 2 YOLO-V3 结构图
Fig2. YOLO-V3 structure diagram

Draknet-53 使用残差网络对图像特征进行更深层地提取，由 53 个卷积单元模块(Conv BN Leaky ReLU, CBL)组成，每个卷积单元模块由一个 Conv 卷积层、一个批量归一化(Batch Normalization, BN)层和一个 Leaky ReLU 激活函数组成，达到最大提取特征的同时也避免了过拟合的目的。当原始图像输入到网络中，首先会被模型划分成 N × N 个网格；接着网络会对每个网格预测 S 个 anchor box 候选预测框，得出相应的 bounding box(bbox)框图，并对这些框图进行 confidence 置信度打分；最后记录(*x*, *y*, *w*, *h*, *P*(Object)**IOU*)这五个元素。其中 *x*, *y*, *w*, *h* 分别表示预测框图的中心像素坐标和框图的长，宽；*P*(Object)表示网络预测该位置是 Object 某个物体的概率，*IOU* 表示框图与框图重叠的程度。

1.2 算法优化

虽然 YOLO-V3 已经具有很好的检测性能，但在目标图像复杂和尺度多样化情况下，检测准确率和识别度仍然有提升空间。在目标检测中，图像的特征提取至关重要，网络需要去除图像中的干扰因素，突出检测目标的特征。通过对 YOLO-V3 检测模型加入通道注意力机制模块来增强网络提取特征的效果，改善相机拍摄目标多尺度和图像场景复杂情况下的目标检测识别度和准确率，通道注意力机制模块如图 3 所示。

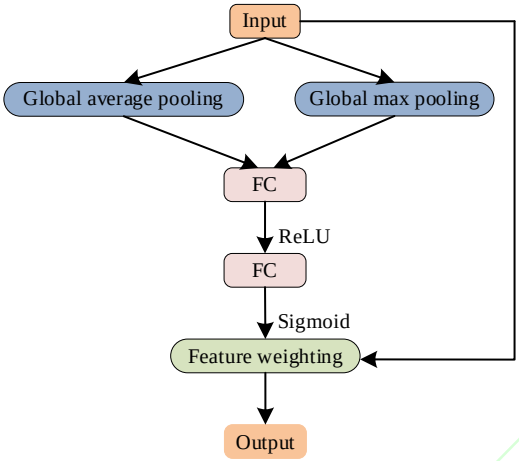


图 3 通道注意力模块图
Fig3. Channel attention module diagram

首先，通过全局平均和最大池化将全局信息压缩为一维矢量，接着通过一个全连接层降低网络维度、一个 ReLU 激活函数和全连接层恢复到原来的网络维度，然后利用 sigmoid 函数归一化特征权重，最后通过特征加权获得权重矩阵。利用权重矩阵对原网络结构中提取的目标特征进行重构，对有利特征信息进行加分无关特征进行减分处理，从而提升目标检测的效果。

2 抓取位姿估计

2.1 五维抓取框

机器人抓取系统想要准确抓取物体，首先需要获取目标可抓取点的位姿信息，现阶段位姿表示方法以平面的 3 自由度(Degrees of Freedom, DOFs)[21]和空间的 6DOFs[22]位姿为主。其中 3DOFs 位姿由目标抓取点的平面坐标(x, y)和偏转角度 θ 组成，而 6DOFs 位姿由目标抓取点的空间坐标(x, y, z)和旋转向量(rx, ry, rz)组成。在工业应用中多以工作台上的物体抓取为主，与空间位姿相比，平面的位姿估计方法更加高效、实用，因此本文采用 3DOFs 位姿表示法扩展而成的五维抓取框图进行位姿估计，如图 4 所示。

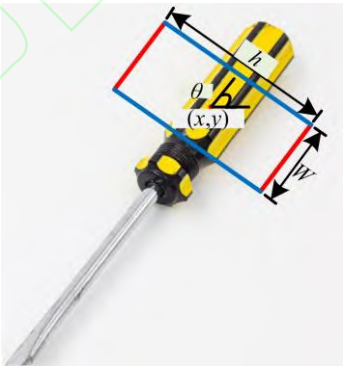


图 4 五维抓取框图
Fig4. Five dimensional grasping frame diagram

抓取框图的中心点坐标(x, y)为平行两指机械手的抓取点；θ 为机械手相对于水平轴的一个偏转角度；h, w 分别表示机械手的开合范围和宽度。五维抓取位姿的函数表示方法如以下式所示。

$$g=\{x,y,w,h,\theta\} \tag{1}$$

2.2 抓取位姿估计模型

目前基于深度学习的抓取位姿估计算法，本质是对 RGB 或 RGD 图像进行回归预测和分类预测。

首先利用主干网络对图像进行特征提取，然后对目标抓取点的$\{x,y,w,h\}$四维信息做回归预测以及抓取角度θ做分类预测[23]。本文拟用 Chu 等[8]提出的多目标、多抓取检测思路进行抓取位姿估计。结合实际研究方案，为了满足对单个物体位姿估计、机械手平稳抓取物体和提高系统运行速度的需求，对原网络结构进行了改进，以实现网络结构简单化和高效运行的目的。

如图 5 所示，本文抓取位姿估计算法由原框架对多个目标的位姿估计的双层网络，替换为对单个目标进行位姿估计的单层网络；并使用 VGG-16 网络替换 ResNet-50 进行特征提取，相比于由 49 个卷积层和 1 个全连接层组成 ResNet-50 网络，VGG-16 由 13 个卷积层和 3 个全连接层组成，网络结构深度不足 ResNet-50 的 1/3。单层的 VGG-16 网络结构满足对单个不同目标的准确预测且估计速度得到提升。

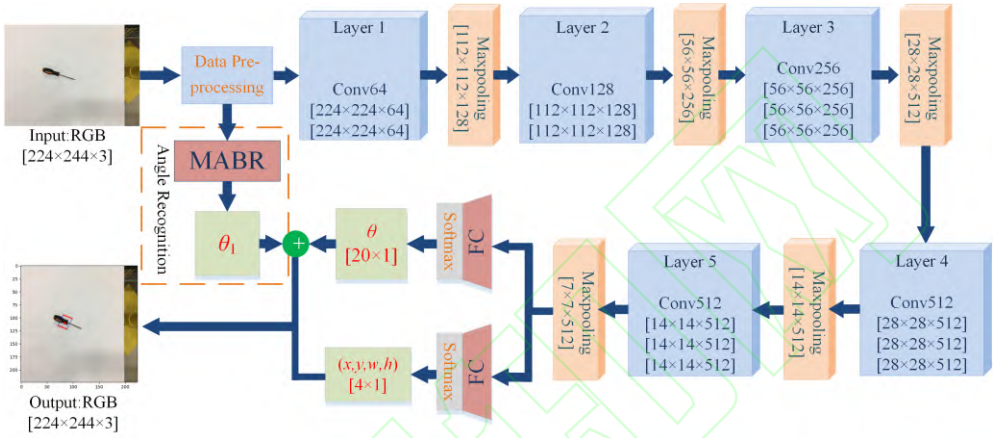


图 5 抓取位姿估计网络结构图

Fig5. Grasping pose estimation structure diagram

抓取角度θ被量化为等长R个间隔，并设置θ为 0 时为非抓握模式，即此时没有预测合适的抓取位姿。网络总损失函数L_{gcr}由抓取角度分类预测损失和四维边框预测损失组成，如下式所示。

$$L_{\text{gcr}}\left(\{(p_l,\beta_l)\}_{c=0}^{C}\right)=\sum_c L_{\text{gcr\_cls}(p_l)} +\lambda\sum_c 1_{c\neq 0}(c)L_{\text{gcr\_reg}\left(\beta_c,\beta_c^*\right)} \tag{2}$$

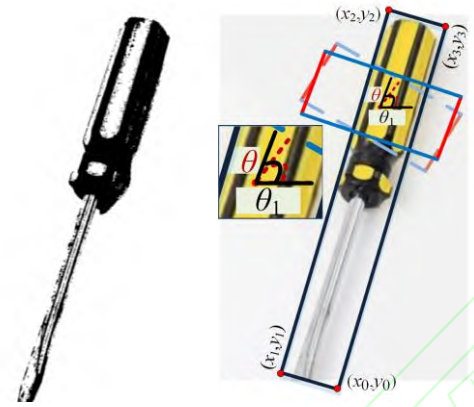
式中C定义抓取角度类别的总数为R+1，取值 19；p_l表示最后经过 Softmax 归一化指数函数层输出第l个角度的分类概率；β_l表示相应预测的抓取边框；$L_{\text{gcr_cls}}$表示抓取角度分类的交叉熵损失；$L_{\text{gcr_reg}}$表示权重为λ的边界框预测的回归损失；β_c, β_c^*分别表示预测的抓取框和真实的抓取框。

将一张 224 × 224 × 3 大小的原始图像输入到网络中，首先进行预处理，然后通过 VGG-16 主干网络进行五组 Conv 卷积特征提取操作和 Max Pooling 最大池化下采样操作，卷积核大小为 3 × 3，特征通道数由 64 逐步扩大到 512。输出端连接两组全连接(Fully Connected, FC)层和 Softmax 归一化指数函数分类器，分别进行抓取角度θ的分类预测和四维框图$\{x,y,w,h\}$的回归预测，最后组合输出相应的目标抓取位姿信息。

2.3 角度优化

虽然通过五维抓取框预测位姿可提高机器人抓取物体的准确率，但该方法存在预测角度离散的问题。位姿估计输出的抓取角度有一个主要分类问题，如图 6(b)虚线框所示，这导致机械手与物体存在较大角度偏差。对工作台上容易移动的物体抓取影响较小，因为物体滑动会消除角度误差；但对工作台上通过夹具固定而不易移动物体的抓取影响较大，因为角度偏差会导致机械手在抓取物体过程中产生一个有害扭矩，导致抓取失败，也容易改变物体当前状态造成再装配困难。

基于上述问题，本文在位姿估计中还引入 MABR 算法，如图 5 虚线所示。通过识别物体偏转角度来解决位姿估计抓取角度离散的问题，角度识别如图 6 所示，从而网络输出抓取物体最优角度，实现机械手对物体平稳地抓取。



(a)二值图　　(b)外接矩形框

图 6 角度识别原理图
Fig6. Angle recognition schematic diagram

首先对图像进行阈值分割，阈值设置为 183，当图像像素灰度值大于 183 时变为 255，反之变为 0，获得如 6(a)图所示的黑白二值图；然后，进行腐蚀、膨胀、开运算和闭运算等操作对二值图像进行去干扰处理；最后，利用最小面积外接矩形包围物体，输出矩形四个顶点坐标 $A_0(x_0, y_0)$，$A_1(x_1, y_1)$，$A_2(x_2, y_2)$，$A_3(x_3, y_3)$。得到包围框的四个顶点坐标后，通过三角函数变换求出矩形任意相邻两条边的长度 a 和 b，如下式所示。

$$a=\sqrt{(x_3-x_0)^2+(y_3-y_0)^2} \tag{3}$$

$$b=\sqrt{(x_3-x_2)^2+(y_3-y_2)^2} \tag{4}$$

然后，对边长 a，b 值进行大小判断，确定矩形框的长 h 和宽 w。如果 $a > b$,长度 h 等于 a，即物体的偏转角度 θ_1 为 α；反之长度 h 等于 b，物体偏转角度 θ_1 为 β。

$$\alpha=\arctan\left|\frac{y_3-y_0}{x_3-x_0}\right|\times\frac{180^\circ}{\pi} \tag{5}$$

$$\beta=\arctan\left|\frac{y_3-y_0}{x_3-x_0}\right|\times\frac{180^\circ}{\pi}+90^\circ \tag{6}$$

将求出的物体偏转角度 θ_1 与位姿估计的抓取角度 θ 进行比较，当二者出现偏差时将 θ_1 替换为两指机械手的抓取角度，从而网络输出最优的抓取角度。

3 实验结果与分析

本文的视觉抓取实验涉及图像处理与机械臂配合工作，系统采用 linux 下基于 Visual Studio Code 编译软件进行开发，确保系统和编译环境的统一性，便于图像处理与机械臂运动控制间的数据传输。接下来将依次进行目标检测、抓取位姿估计、物体角度矫正和真实机械臂抓取分类实验。

3.1 目标检测

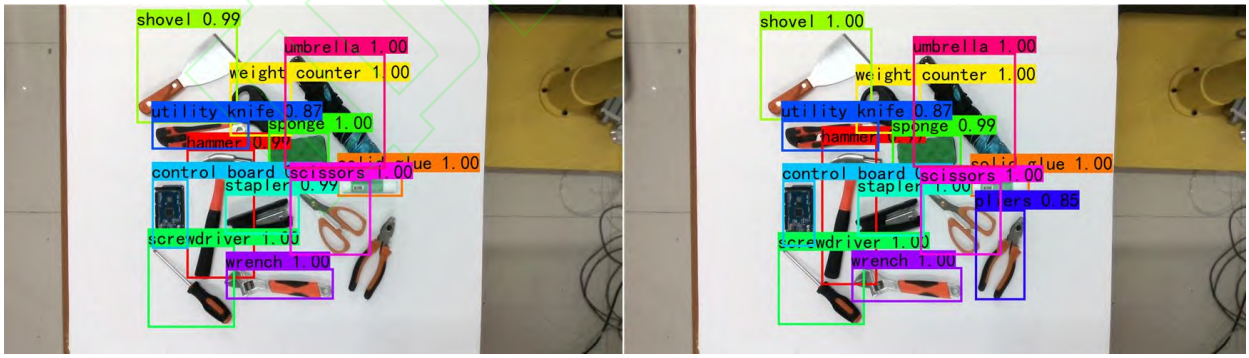
目标检测采用自建 VOC2007 格式数据集制作标签进行模型训练，一共选取 12 个类别，采集了 1490 张图像。类别如图 7 所示，分别是“hammer”、“solid glue”、“weight counter”、“shovel”、“sponge”、“screwdriver”、“stapler”、

“control board”、“pliers”、“wrench”、“scissors”、“umbrella”，选取其中的 1341 张图像用于模型训练，剩余的 149 张图像用于模型测试。

图 7 目标检测类别图
Fig7. Target detection category diagram

目标检测模型采用 Pytorch 深度学习框架，使用 NVIDIA 的 GTX1660 型号 GPU 对模型训练进行加速。由于主干特征提取网络具有特征通用性，因此也采用冻结训练方法二次加快模型训练的速度。将模型训练的参数设置如下，解冻前：1) 基础学习率 *lr* 为 0.001；2) 批量大小 *Batch_size* 为 8；3) 起始训练迭代 *Init_epoch* 为 0；4) 冻结训练迭代 *Freeze_ epoch* 为 50。解冻后：1) 基础学习率 *lr* 为 0.0001；2) 批量大小 *Batch_size* 为 4；3) 起始训练迭代 *Init_epoch* 为 50；4)解冻训练迭代 *UnFreeze_epoch* 为 100。

利用模型训练得出的权重对输入网络的图像进行目标检测，输出对应的预测种类和概率，结果如图 8 所示。相比于改进前，加入注意力机制模型的检测网络平均识别准确率(mean Average Precision, mAP)由 92.33%增加到 92.65%，提升 0.32%。并在网络置信度不变的情况下，降低模型在杂乱环境下了漏检的可能，检测效果更加突出，证明了改进模型的实际意义。



(a)改进前　　(b)改进后

图 8 目标检测结果图
Fig8. Target detection result diagram

3.2 抓取位姿估计

表1 算法对比表
Table1 Algorithm comparison table

算法	准确率/%		运行时间/s
	cornell 数据集	实验目标	
双层结构 ResNet-50	91.30	87.11	0.932
单层结构 ResNet-50	91.12	86.69	0.714
单层结构 VGG-16	90.89	87.19	0.286

抓取位姿估计采用 cornell 数据集制作模型训练所需的数据样本，将连续的 180 度等分成 19 个离散值制作抓取框的角度标签。该实验使用的深度学习框架和加速设备与目标检测保持一致，将模型训练的参数设置如下：1) 学习率 *lr* 为 0.0001；2) 批量大小 *batch_size* 为 8；3) 训练迭代 *epoch* 为 1000。并且输出端利用 MABR 算法识别的物体偏转角度来优化位姿估计的抓取角度。

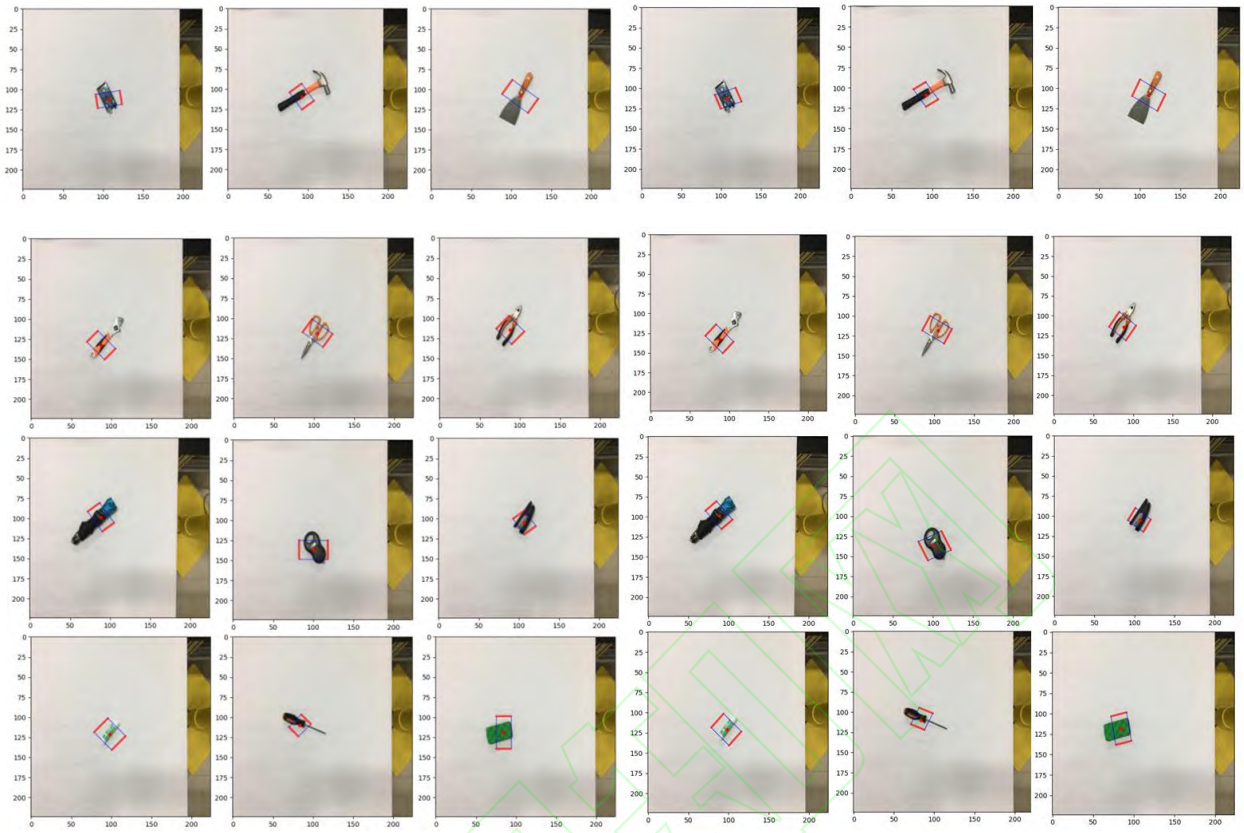
将本文的单层网络结构 VGG-16 方法分别与双层网络结构的 ResNet-50 和单层网络的 ResNet-50 进行单个物体抓取位姿估计对比实验，电脑配置为 GTX1660 显卡和 8 GB 运行内存，对比结果如表 1 所示。

从对比结果能够得出，对于单个物体的抓取位姿估计，双层结构和更深层 ResNet-50 网络在估计准确率上并没有突出的表现，反而单层结构 VGG-16 的方法在运行时间上有明显的优势。

利用模型训练得出的权重对目标图像进行抓取位姿框图的预测，输出相应预测框抓取点的像素坐标(u, v)和两指机械手偏转角度 θ。并与未加 MABR 算法的抓取位姿估计方法进行对比实验，输出的数据如表 2，结果如图 9 所示。相比于改进前，改进后位姿估计的抓取角度连续化，更加趋于物体的偏转角度。通过实验测量，计算出改进后的预测抓取角度与目标的实际偏转角度平均误差小于 2.47 °。

表2 位姿估计结果表
Table2 Pose estimation table

目标	目标抓取点(u, v)/像素值		目标抓取角度/°		目标实际角度/°
	改进前	改进后	改进前	改进后	
控制板	(107, 112.2)		100	124	123
锤子	(92.3, 109.3)		30	18	23
铲子	(111.3, 108.5)		50	62	59
扳手	(87.5, 132.5)		40	46	44
剪刀	(104.5, 118.3)		50	53	54
钳子	(88.1, 114.5)		40	48	52
雨伞	(88.5, 98.4)		30	35	35
计重器	(100.5, 136.2)		90	135	127
订书机	(106.9, 104.7)		30	46	48
固体胶	(98.2, 120.9)		40	45	45
螺丝刀	(83.6, 110.2)		130	161	162
海绵	(84.6, 118.7)		180	13	14



(a)改进前　　(b)改进后

图 9 抓取位姿估计图

Fig9. Grasping pose estimation result diagram

3.3 机械臂抓取实验

本文所实施的抓取分类实验主要基于 UR5 机械臂、气动两指机械手及 Realsense D435 相机等搭建的平台实现。相机固定在平台的正上方，安装方式采用“眼在手外”的模式，并且使用 ATI-Mini45 六维力传感器搭建测力平台验证本文改进抓取角度方法的有效性，系统样机搭建如图 10 所示。

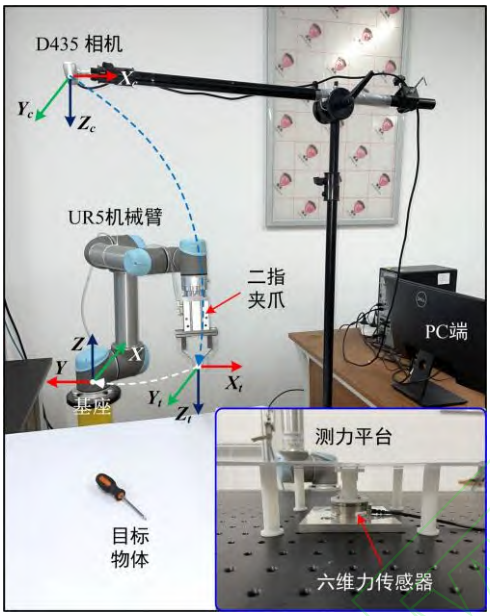


图 10 系统样机搭建
Fig10. System prototype setup

测力平台通过 3D 打印的卡具将硅胶、亚克力板与六维力传感器连接起来。抓取对象通过胶水固定在亚克力板上，硅胶起传导作用力和避免较大扭矩损坏设备的作用。抓取系统的坐标转换流程如下：相机首先获取图像的二维像素坐标，通过相机的深度信息和内参数据将图像像素坐标转换到基于相机坐标系下三维坐标；然后利用手眼标定的关系矩阵，将相机坐标系下的坐标转换成机械臂基座坐标系下的三维坐标，最终实现了抓取目标到机械臂基座坐标系下的坐标转换。

像素坐标(u, v)与深度相机坐标(x_c, y_c, z_c)之间的变换如下式所示。

$$\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = \begin{bmatrix} \alpha & 0 & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \boldsymbol{K}^{-1} z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{7}$$

式中，$\alpha=f/dx$ 和 $\beta=f/dy$ 表示 u, v 像素坐标系下的实际焦距；f 是相机固有焦距；dx 和 dy 是在 u 轴和 v 轴上的物理像素尺寸；K 是相机的内参矩阵；z_c 是相机坐标下的深度信息。

深度相机坐标(x_c, y_c, z_c)与机械臂基座坐标(x, y, z)之间的变换如下式所示。

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{T} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = \boldsymbol{T}_m \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} \tag{8}$$

式中，R 和 T 分别表示手眼关系的旋转矩阵和平移向量。

完成不同坐标系下的坐标转换后，PC 端通过获取的数据对机械臂进行相应的控制。抓取分类实验通过 Ubuntu16.04 系统中 Moveit！运动功能包实现，具体实验步骤如下：

步骤 1 设置 UR5 机械臂抓取拍照等待位姿，坐标为(x_0, y_0, z_0)，两指机械手偏转角度为 0 °；

步骤 2 相机获取目标图像，计算机处理数据，输出目标抓取点的坐标(x, y, z)和偏转角度 θ 信息；

步骤 3 控制两指机械手偏转 θ 角度，机械臂由等待位(x_0, y_0, z_0)移到抓取位(x, y, z)，准备抓取；

步骤 4 气动控制两指机械手闭合，完成物体抓取，然后机械臂根据目标检测结果进行相应分类放置；

步骤 5 完成放置操作后，机械臂回到初始拍照等待位置；

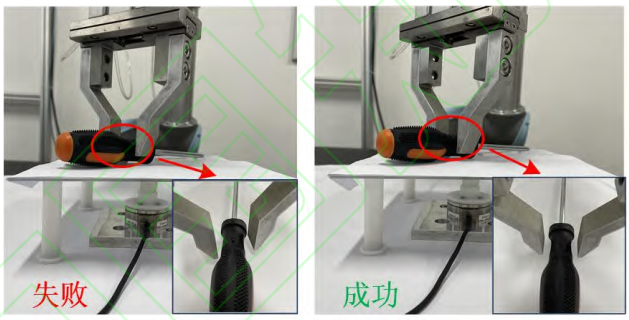
步骤 6 如果继续抓取，则返回步骤 1；否则，抓取任务结束。

现对实验目标设置 12 组改进前与改进后的对比抓取实验，来验证改进后两指机械手抓取的准确率和平稳效果，编号为实验 1~12，各目标抓取点的坐标、角度以及目标实际角度如表 3 所示。

表3 抓取实验数据表

Table3 Grasping experimental data table

编号	目标抓取点(*x*, *y*, *z*)/mm		目标抓取角度/°		目标实际角度/°	抓取扭矩/N·mm	
	改进前	改进后	改进前	改进后		改进前	改进后
实验 1	(153.41, -675.29, 102.35)		50	52	53	4	2.3
实验 2	(13122, -603.70, 99.16)		50	58	58	9.5	0.3
实验 3	(161.96, -558.44, 102.71)		140	157	156	15	2.6
实验 4	(111.15, -574.50, 98.79)		10	21	19	8	5
实验 5	(114.63, -732.19, 96.96)		30	39	47	19	11
实验 6	(102.68, -657.68, 100.41)		40	51	50	10.6	2.5
实验 7	(127.41, -675.63, 100.53)		40	46	45	4	1.5
实验 8	(155.39, -597.67, 105.50)		50	53	55	8	3.8
实验 9	(176.65, -690.90, 103.57)		90	111	113	17.5	4
实验 10	(131.77, -739.27, 100.34)		100	112	112	12.5	0
实验 11	(194.20, -687.68, 101.49)		30	36	35	5	2.5
实验 12	(127.47, -590.49, 100.38)		30	63	63	25	0



(a)改进前　　(b)改进后

图 11 抓取实验

Fig.11 Grasping experiments

从图 11(a)与(b)对比看出，改进后两指机械手与物体的有害夹角较改进前有明显改善，这说明本方法可以使机械手用更贴合物体的偏转角度进行抓取，从而减少了对物体的干扰，提高抓取成功率。

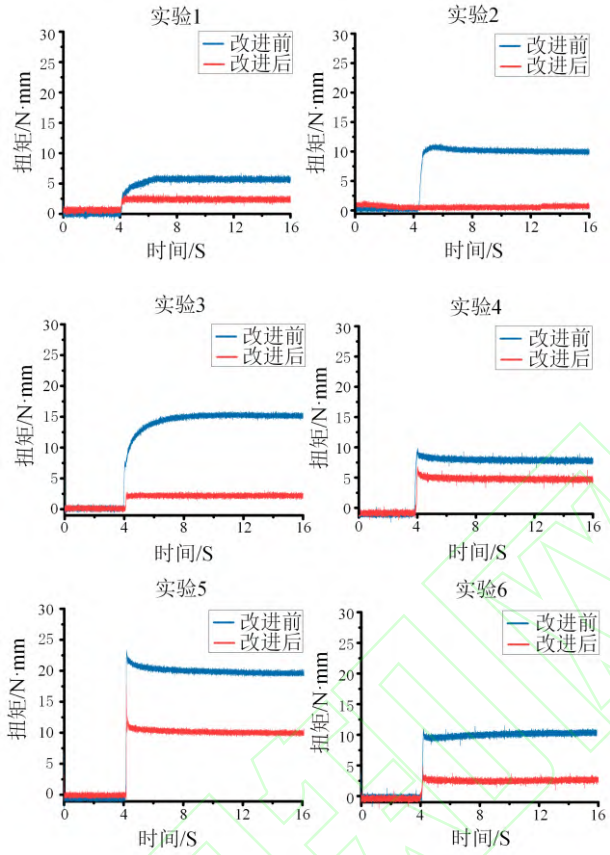


图 12 抓取扭矩图

Fig.12 Grasping torsion diagram

为了定量描述本文改进后的抓取效果，通过六维力传感器搭建的测力平台测量两指机械手抓取过程中物体 Z 轴方向受到的扭矩。实验中两指机械手的闭合与张开通过气缸控制，气压设置范围在 0.05 MPa ~ 0.08 MPa 之间。六维力传感器采样时间 16 s，频率为 5000 Hz，测量的 12 组对比实验数据如上表 3 中所示，其中前 6 组实验物体抓取所受的有害扭矩对比如图 12 所示。通过实验测量，设定当预测抓取角度与目标偏转角度误差大于 5 度，即产生的扭矩大于 10 N·mm 时，抓取失败，反之成功，如图 11 所示。从表 3 和图 12 的实验数据计算得出，改进后两指机械手抓取物体所产生的平均有害扭矩由改进前的 11.5 N·mm 降低到 2.9 N·mm，减少约 75%；抓取成功率由改进前 50%增加到 91%，提升约 40%，证明了本文改进机器人抓取目标方法的有效性。

4 结论

针对现有机器人抓取分类系统难以适应不同物体和抓取过程中产生较大有害扭矩的问题，提出基于深度学习视觉检测及抓取方法。

1) 通过实验测试，加入通道注意机制的 YOLO-V3 检测模型，对复杂环境中的目标检测效果更突出，平均识别率较改进前增加 0.32%。

2) 通过实验对比可知，VGG-16 主干网络嵌入 MABR 的算法，可以有效地对目标进行抓取位姿估计，并且输出的抓取角度更优，与目标的实际偏转角度平均误差小于 2.47°，具有很好的实际意义。

3) 通过 UR5 机械臂、气动两指机械手、Realsense D435 相机及 ATI-Mini45 六维力传感器等搭建的实验平台进行了验证，结果表明能有效地抓取不同物体，抓取过程中机械手产生的有害扭矩减少约 75%，显著减小两指机械手抓取过程对物体的干扰，抓取稳定性增加、成功率提高约 40%，具有很好

的应用前景。

参考文献(References)

[1] DU G,WANG K,LIAN S,et al.Vision-based robotic grasping from object localization,object pose estimation to grasp estimation for parallel grippers:a review[J].Artificial Intelligence Review,2021,54(3):1677-1734.

[2] 翟敬梅，董鹏飞，张铁．基于视觉引导的工业机器人定位抓取系统设计[J]．机械设计与研究，2014，30（5）：45-9．ZHAI J M，DONG P F，ZHANG T．Positioning and grasping system design of industrial robot based on visual guidance[J]．Machine Design&Research，2014，30（5）：45-49（in Chinese）．

[3] WEI H,PAN S,MA G,et al.Vision-guided hand-eye coordination for robotic grasping and its application in tangram puzzles[J].AI,2021,2(2):209-228.

[4] MALLICK A,DELPOBIL A P,CERVERA E.Deep learning based object recognition for robot picking task[C]//Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication(IMCOM2018).New York,USA,2018:1-9.

[5] 白成超，晏卓，宋俊霖．结合深度学习的机械臂视觉抓取控制[J]．载人航天，2018，3：299-307．
BAI C C,YAN Z,SONG J L.Visual grasp control of robotic arm based on deep learning[J].Manned Spaceflight,2018,3:299-307(in Chinese).

[6] 黄怡蒙，易阳．融合深度学习的机器人目标检测与定位[J]．计算机工程与应用，2020，56（24）：181-187．
HUANG Y M,YI Y.Robot object detection and localization based on deep learning[J].Computer Engineering and Applications,2020,56(24):181-187(in Chinese).

[7] 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(ICRA2011).Shang Hai,China,IEEE,2011:3304-3311.

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

[9] 夏浩宇，索双富，王洋，等．基于 Keypoint RCNN 改进模型的物体抓取检测算法[J]．仪器仪表学报，2021，42（4）：236-246．
XIA H Y,SUO S F,WANG Y,et al.Object grasp detection algorithm based on improved Keypoint RCNN model[J].Chinese Journal of Scientific Instrument,2021,42(4):236-246(in Chinese).

[10] ZHANG Z.Flexible camera calibration by viewing a plane from unknown orientations[C]//Proceedings of the seventh IEEE International Conference on Computer Vision(ICCV1999).Corfu,Greece,1999,1:666-673.

[11] REDMON J,FARHADI A.YOLOv3:An incremental improvement[J].arXiv e-prints,2018.

[12] SIMONYAN K,ZISSERMAN A.Very deep convolutional networks for large-scale image recognition[J].arXiv preprint,1409.1556,2014.

[13] SONG R,LI F,Fu T,et al.A Robotic automatic assembly system based on vision[J].Applied Sciences,2020,10(3):1157.

[14] 尹宏鹏，陈波，柴毅，等．基于视觉的目标检测与跟踪综述[J]．自动化学报，2016，42：1466-1489．
YIN H P,CHEN B,CHAI Y,et al.Vision-based object detection and tracking:a review[J].Acta Automatica Sinica,2016,42:1466-1489(in Chinese).

[15] 王玺坤，姜宏旭，林珂玉．基于改进型 YOLO 算法的遥感图像舰船检测[J]．北京航空航天大学学报， 2020， 46（6）：1184-1191．
WANG X K,JIANG H X,LIN K Y.Remote sensing image ship detection based on modified YOLO algorithm[J].Journal of Beijing University of Aeronautics and Astronautics,2020,46(6):1184-1191(in Chinese).

[16] ZHANG N,DONGAHUE J,GIRSHICK R,et al.Part-based R-CNNs for fine-grained category detection[J].European Conference on Computer Vision,2014,834-849.

[17] GIRSHICK R.Fast r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision(ICCV2015).Santiago,Chile,2015:1440-1448.

[18] REN S,HE K,GIRSHICK R,et al.Faster r-cnn:Towards real-time object detection with region proposal networks[J].Advances in Neural Information Processing Systems,2015,28:91-99.

[19] REDMON J,DIVVALA S,GIRSHICK R,et al.You only look once:Unified,real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR2016).Las Vegas,USA,2016:779-788.

[20] 刘元宁，吴迪，朱晓冬，等．基于 YOLOv3 改进的用户界面组件检测算法[J]．吉林大学学报（工学版），2021，215（3）：1026-1033．
LIU Y N,WU D,Zhu X D,et al.User interface components detection algorithm based on improved YOLOv3[J].Journal of Jilin University(Engineering and Technology Edition),2021,215(3):1026-1033(in Chinese).

[21] 熊军林，赵铎．基于 RGB 图像的二阶段机器人抓取位置检测方法[J]．中国科学技术大学学报，2020， 50：1-10．
XIONG J L,ZHAO D.Two-stage grasping detection for robots based on RGB images[J].Journal of University of Science and Technology of China,2020,50:1-10(in Chinese).

[22] TEKIN B,SINHA S N,FUA P.Real-time seamless single shot 6d object pose prediction[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR2018).Salt Lake,USA,2018:292-301.

[23] KUMRA S,KANAN C.Robotic grasp detection using deep convolutional neural networks[C]//Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS2017).Vancouver,Canada,2017:769-776.2017:769-776.