第 60 卷 第 6 期/2023 年 3 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于两阶段的机器人动态多物品定位抓取方法

孟月波[1*]，黄琪[1]，韩九强[2]，徐胜军[1]，王宙[1]

[1]西安建筑科技大学信息与控制工程学院，陕西 西安 710055；

[2]西安交通大学自动化科学与工程学院，陕西 西安 710049

摘要 为解决工厂流水线上不同种类动态物品的快速精准抓取问题，提出一种两阶段动态多物品定位抓取方法。第1阶段采用所提多尺度上下文感知的单分支融合语义分割网络获取目标物品的掩码区域：首先特征提取网络采用单分支结构，在保证提取丰富的空间信息和高层语义信息的同时，减小网络参数量；随后特征融合网络通过双边引导特征融合模块增强空间信息和语义信息的表达能力；最后设计特征增强网络，通过特征辅助收敛模块嵌入浅层和深层网络中，加快网络收敛速度。第2阶段采用基于轮廓点检测的快速位姿估计策略在掩码区域预测最佳抓取点位姿。在自建数据集上的测试及流水线平台抓取实验结果表明，所提方法能实时检测和预测物品抓取点位姿，精准完成物品抓取，其分割精度、预测时间和抓取成功率均优于对比方法。

关键词 机器视觉；机器人抓取；两阶段定位抓取算法；多尺度上下文感知；特征增强；位姿估计

中图分类号 TP242 **文献标志码** A **DOI**：10.3788/LOP213364

Robot Dynamic Object Positioning and Grasping Method based on Two Stages

Meng Yuebo[1*], Huang Qi[1], Han Jiuqiang[2], Xu Shengjun[1], Wang Zhou[1]

[1]*College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an* 710055, *Shaanxi, China*;

[2]*College of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an* 710049, *Shaanxi, China*

Abstract A two-stage dynamic multi-object positioning and grasping method is proposed to solve the problem of fast and accurate grasping of various types of dynamic objects on a factory assembly line. In the first stage, the proposed multiscale context-aware single-branch fusion semantic segmentation network is used to obtain the mask area of the target object: first, the feature extraction network adopts a single-branch structure, which reduces the number of network parameters while ensuring the extraction of rich spatial information and high-level semantic information; subsequently, the feature fusion network improves the expression ability of spatial data and semantic information through the bilateral guided feature fusion module; finally, the feature enhancement network is designed, and the feature assisted convergence module is embedded in the shallow and deep networks to accelerate the convergence speed of the network. In the second stage, a quick pose estimation strategy based on contour point detection is applied to predict the optimum posture of the grasping point in the mask region. The test results on the self-built dataset and the pipeline platform grab experiments demonstrate that the proposed method can detect and predict the position and posture of the object grab points in real time and accurately complete the object grab. Furthermore, its segmentation accuracy, prediction time, and grab success rate are better than the comparison method.

Key words machine vision; robot grab; two-stage positioning and grabbing algorithm; multi-scale context perception; feature enhancement; pose estimation

1 引 言

随着物流行业的崛起，商品流通速度加快，快递打包需求日益增长[1]。国内人工成本不断攀升，打包机器人渐渐应用于物流工厂流水线上，替代人工完成对物品的分拣包装，极大地提高工作效率。物品的精准

收稿日期：2021-12-27；**修回日期**：2022-01-16；**录用日期**：2022-01-27；**网络首发日期**：2022-02-09

基金项目：自然科学基础研究计划面上项目(2020JM-473，2020JM-472)、陕西省重点研究计划项目(2021SF-429)

通信作者：*mengyuebo@163.com

抓取是分拣包装的关键环节，因此关于物品快速定位抓取的研究具有重大意义。

现阶段，机器人定位抓取方法分为传统的方法和深度学习的方法两类。传统的定位抓取方法主要通过检测目标物体的边缘轮廓、几何结构等特征，在模板库匹配搜索确定抓取位姿。宋薇等[2]通过Canny算法检测物品边缘信息，采用倾角分层的chamfer距离匹配算法计算图像边缘和模板间的相似度匹配，最后运用爬山法局部最优的遗传算法搜索最佳的抓取位姿；张震等[3]采用Shi-Tomasi角点检测算法[4]提取物品特征点信息，通过改进局部二值模式（LBP）算法得到特征点对应的特征描述子与模板间的Hamming距离进行特征点匹配，估计单应性矩阵从而确定目标物体在实际场景中的位置和方向。但这些方法容易受光照、背景及遮挡影响，泛化能力较差。

近年来，卷积神经网络（CNN）因其具有获取深层次特征的能力，在实例分割、目标检测等领域广泛应用[5-8]，学者也将其应用到机器人抓取领域。文献[9-10]采用Faster RCNN通过预测物品边界框（bounding box）中心点的方式实现物体的定位抓取，但此类方法忽视姿态信息，使用范围受限，只适用于吸盘类夹具或者固定角度抓取场景等。

为了获取目标物品的姿态角度，有学者提出级联网络思想：Lenz等[11]首先采用小的深度网络生成目标物体抓取候选框，然后通过大的深度网络筛选可能性最大候选框，从而确定物体抓取位姿；陈丹等[12]提出两级Faster RCNN位姿估计网络，改进Faster RCNN用于准确获取目标物品区域，标准Faster RCNN用于目标物体区域最佳抓取位姿预测；李秀智等[13]改进YOLOV3网络提升其小目标识别与定位能力，多目标抓取检测网络生成诸多抓取矩形，利用交并比（IoU）区域评估算法筛选最优抓取矩形，将矩形中点和矩形相对水平方向的夹角作为抓取位姿；夏晶等[14]以R-FCN[15]定位抓取候选框并进行候选框角度粗估计，采用Angle-Net来进一步精细抓取角度，确定抓取位姿。级联网络采用滑动框遍历图像寻找最佳位姿，预测时间长，很难满足动态物体抓取的要求，同时对不规则物品位姿预测不准确。

语义分割方法通过像素级标注实现物品的精细化分割，不仅能够获得类别信息，还能够提取物体的精准空间位姿，为物品位姿预测与抓取提供新途径。陈奎烨等[16]采用Mask RCNN获取目标物品的掩码区域，将掩码区域的中心和掩码区域对应的最小外接矩形和水平方向的夹角作为抓取位姿；王德明等[17]利用Mask RCNN获取目标物品空间点云信息，通过改进的4点全等集（4PCS）算法和迭代最近点（ICP）算法将分割出的点云和目标模型的点云进行匹配和位姿精修。由于分割网络参数量大、计算时间长，文献[16]、[17]网络只适用于静态场景。本实验组认为在保证准确性的前提下减少网络参数，通过分割网络轻量化的方式进行精准位姿实时估计，是实现多品种物品定位抓取的有效手段。当前轻量化语义分割方法大多采用编码-解码结构[18-19]，通过限制输入图像的尺寸、降低图像的分辨率来提高推理速度，但同时也造成空间信息的缺失，导致较差的分割效果。Yu等[20]提出的BiSeNet使用双分支网络融合低层次空间特征和高层次语义特征，提高分割精度。Yu等[21]在BiSeNet的基础上提出BiSeNetV2，BiSeNetV2采用浅层宽通道空间分支网络和深层窄通道语义分支网络减少网络参数量。Zhao等[22]提出采用三分支结构的ICNet，该网络高效利用低分辨的语义信息和高、中分辨率的空间信息，提升分割图质量。多分支结构虽然为平衡精度和速度提供了一种有效的方法，但由于多分支结构每一分支具有相似的学习功能，会造成结构冗余现象和精度不佳问题。文献[23]提出的Fast-SCNN以MobileNetV2[24]作为骨干网络，采用单分支融合的方式获取分割图，但对空间特征图和语义特征图进行通道拼接融合，造成语义信息和空间信息融合不充分，最终导致分割精度不佳。

为达到预测位姿准确性和实时性的平衡，本文提出一种两阶段动态多物品定位抓取方法。第1阶段提出获取目标物品掩码区域的多尺度上下文感知的单分支融合语义分割网络，该网络由特征提取网络、特征融合网络和特征增强网络组成。特征提取网络采用改进的VGG前七层卷积结构作为骨干网络获取丰富的空间信息；上下文感知结构通过级联多尺度感知层（MSPL）聚合图像中的多尺度空间位置信息，提升网络区域关注能力；上下文嵌入层（CEL）补充全局上下文信息，与多尺度信息融合获取高层语义信息，随后特征融合网络采用双边引导特征融合模块（BGFFM），通过双向指导的方式实现深层语义信息和浅层空间信息充分融合，从而提升分割图的输出质量，获取目标物品掩码区域。为进一步提升分割精度，提出特征增强网络，通过特征辅助收敛模块（FACM）提升网络各层的收敛能力。第2阶段采用基于轮廓点检测的快速位姿估计策略根据掩码的轮廓坐标信息，预测目标物品抓取点位置坐标和姿态角度。将所提定位抓取方法应用于小型流水线实验平台进行实验。机器人抓取实验结果表明，所提方法能实时预测物品位姿并精准完成抓取。

2 基于多尺度上下文感知的单分支融合语义分割网络

基于多尺度上下文感知的单分支融合语义分割网络由3部分组成，具体如图1所示：第1部分为特征提取网络，分别提取图像中的浅层空间特征信息和深层语义特征信息；第2部分为特征融合网络，两种特征信息通过双向指导的方式进行深度融合；第3部分为特征增强网络，指导网络在浅层和深层更好收敛，进一步提高分割精度。

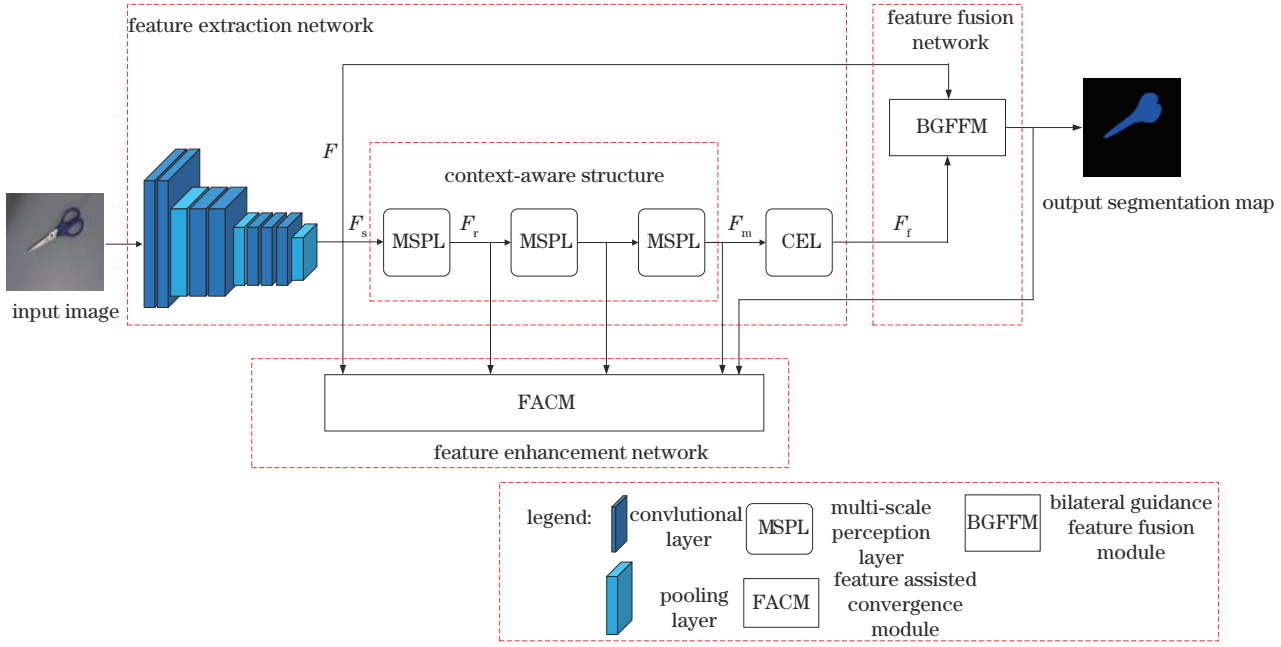

图 1　多尺度上下文感知的单通道融合网络结构

Fig. 1　Multi-scale context-aware single-channel fusion network structure

2.1　特征提取网络

特征提取网络采用单分支网络，由骨干网络、上下文感知结构和上下文嵌入层组成，其中骨干网络用于获取丰富的空间信息，单分支网络整体作为语义分支网络获取高层语义信息。

2.1.1　骨干网络

骨干网络采用改进的 VGG 前七层卷积结构获取空间特征图 F 和浅层语义特征图 F_s；单分支网络输出深层语义特征图 F_f。

$$F, F_s = \mathrm{VGG\_7}(I) \tag{1}$$

式中：I 表示输入的原图；VGG_7 表示改进的 VGG 前七层卷积结构。相比于标准的 VGG 前七层卷积层，VGG_7 在每个卷积层后添加 ReLU 层和批归一化（BN）层，引入非线性，加速网络推理速度，并将 Conv5、Conv6 和 Conv7 卷积层输出特征图通道数从 256 变为 128，减少网络模型参数。

表 1　改进的 VGG 前七层卷积结构

Table 1　First seven-layer convolution structure of improved VGG

Network layer	Input size	Kernel size	Output size
Conv1+BN+ReLU	3×640×640	3×3	64×640×640
Conv2+BN+ReLU	64×640×640	3×3	64×640×640
MaxPooling	64×640×640	2×2	64×320×320
Conv3+BN+ReLU	64×320×320	3×3	128×320×320
Conv4+BN+ReLU	128×320×320	3×3	128×320×320
MaxPooling	128×320×320	2×2	128×160×160
Conv5+BN+ReLU	128×160×160	3×3	128×160×160
Conv6+BN+ReLU	128×160×160	3×3	128×160×160
Conv7+BN+ReLU	128×160×160	3×3	128×160×160
MaxPooling	128×160×160	2×2	128×80×80

2.1.2　上下文感知结构

骨干网络在图像中具有相同的感受野，对不同尺寸的特征感知不敏感。所提上下文感知结构通过 3 层多尺度感知层进行上下文信息递进传递，强化待检测物品关注区域特征信息。每个多尺度感知层均通过 4 层级感受野提取多尺度特征，利用注意力机制引导获取上下文空间位置信息感知，提高区域关注能力。多尺度感知层如图 2 所示。

将改进的 VGG 前七层卷积结构输出的浅层语义特征图 F_s 作为原始特征图，采用 4 分支结构，通过特征金字塔池化（SPP）获取多尺度特征图 F_j，分别记为 F_1、F_2、F_3 和 F_4。

$$F_j = \mathrm{Up}\left\{\mathrm{Conv}\left[\mathrm{Adap}(F_s, j)\right], \theta_j\right\}, \tag{2}$$

式中：Adap 表示自适应平均池化，将原始特征图下采样输出指定尺寸的特征图，4 个特征图尺度大小设置为 1×1、2×2、3×3 和 6×6；Conv 表示 1×1 卷积，在实现跨通道信息交互同时，改变特征图通道数，减少网络参数；θ_j 表示各个 1×1 卷积对应的权重；Up 表示上采样，将特征图恢复到原始特征图大小。

对于多尺度特征图，对比特征可以提供每个图像区域局部尺度的重要信息[25]，本研究利用上下文对比特征图 C_j 捕捉特定空间位置特征和相邻空间特征的差异。

$$C_j = F_j - F_s。 \tag{3}$$

利用上下文对比特征图 C_j 生成对应的上下文权重图 W_j，记为 W_1、W_2、W_3 和 W_4，预测各个尺度特征在每个空间位置的关注程度。

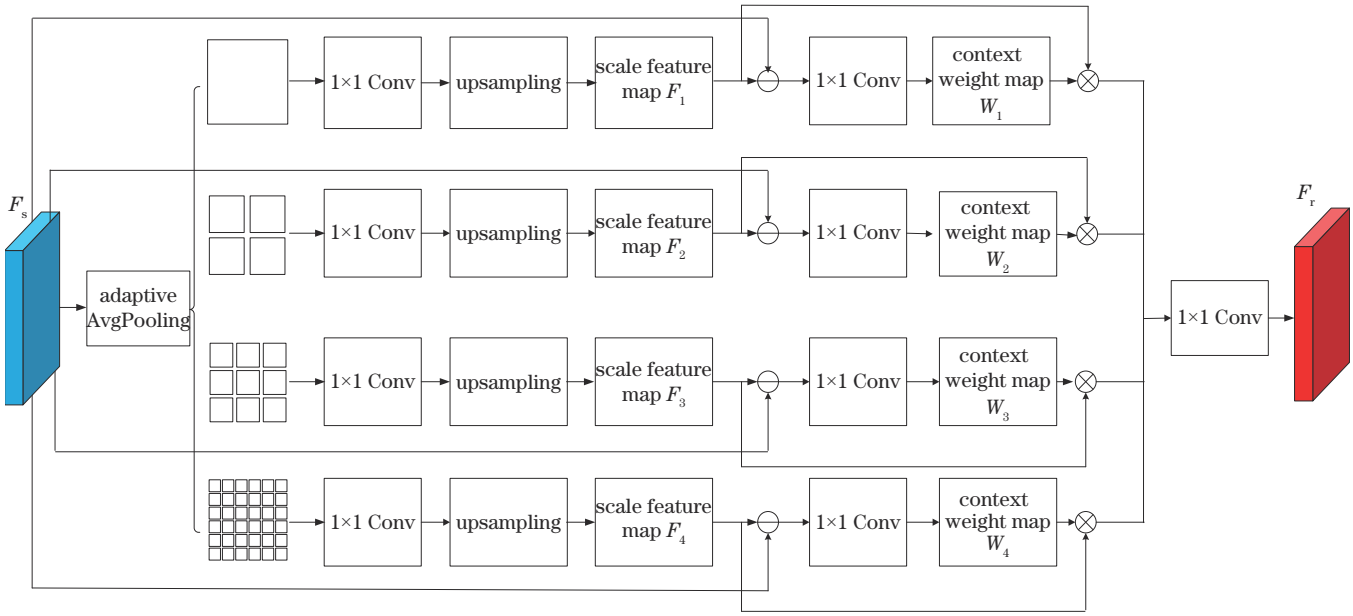

图 2　多尺度感知层

Fig. 2　Multi-scale perception layer

$$W_j=\sigma\left[\mathrm{Conv}(C_j,\theta_j)\right], \tag{4}$$

式中：Conv 表示 1×1 卷积，生成图像中每个空间位置的权重；σ 表示 Sigmoid 激活函数，将权重值映射到 0～1 之间，增强网络的非线性能力。

对于上下文权重图，利用其指导网络学习多尺度特征图中的关注区域，生成多尺度上下文感知特征图 F_r。

$$F_r=\mathrm{Conv}\left[\mathrm{multi}(W_j,F_j)\right], \tag{5}$$

式中：multi 表示各个分支的尺度特征图和它对应的上下文权重图逐像素相乘，并输出对应的特征图；Conv 表示 1×1 卷积，将 4 个分支的特征图融合获取多尺度上下文感知特征图。

为了让网络进一步学习关注区域，采用多个多尺度感知层，并在多尺度感知层间采用池化操作，进一步提升分割效果。实验结果表明，3 层多尺度上下文感知层分割性能最佳，上下文感知结构输出特征图为 F_m。

2.1.3　上下文嵌入层

因语义分支需要足够大的接收域来捕获高级语义信息，本研究采用上下文嵌入层（CEL）通过全局平均池化（GAPooing）和残差结构组合的方法，增强全局上下文信息的同时防止过拟合，如图 3 所示。将上下文感知结构输出特征图作为输入，首先通过全局平均池化和 1×1 卷积将其降维至 $C\times1\times1$；然后采用广播机制（broadcast mechanism）将其和输入特征图进行像素级相加；最后采用 3×3 卷积融合获取深层语义特征图 F_f。

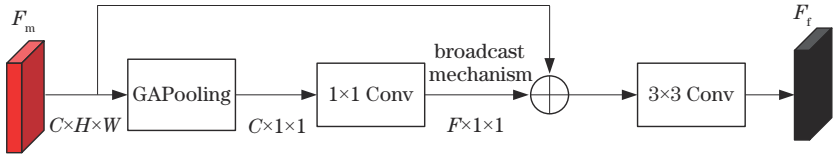

图 3　上下文嵌入层结构

Figure 3　Context embedding layer structure

2.2　特征融合网络

针对特征提取网络获取的空间特征图 F 和语义特征图 F_f，对空间特征与语义特征进行拼接或者相加的操作无法将不同维度的特征信息有效融合。本研究沿用文献[21]的思想，采用双边引导特征融合模块，通过互相指导的方式进行深度融合，如图 4 所示。在 3 个多尺度感知层间进行两次池化操作，语义特征图 F_f 首先通过上采样至空间特征图 F 的尺寸。两特征图分别采用 3×3 卷积、1×1 卷积和 Sigmoid 激活函数生成对应的特征引导注意力图，通过特征引导注意力图指导对方捕获丰富的特征组合表达，最后通过 3×3 卷积融合特征信息，获取预测分割图。

2.3　特征增强网络

随着卷积层和池化层的增加，图像降维过程中信息不断缺失，导致深层网络很难学习到新的特征信息，出现网络退化问题。所提特征增强网络将特征辅助收敛模块嵌入骨干网络、上下文感知结构中各个多尺度感知层和特征融合模块后端，以特征辅助收敛模

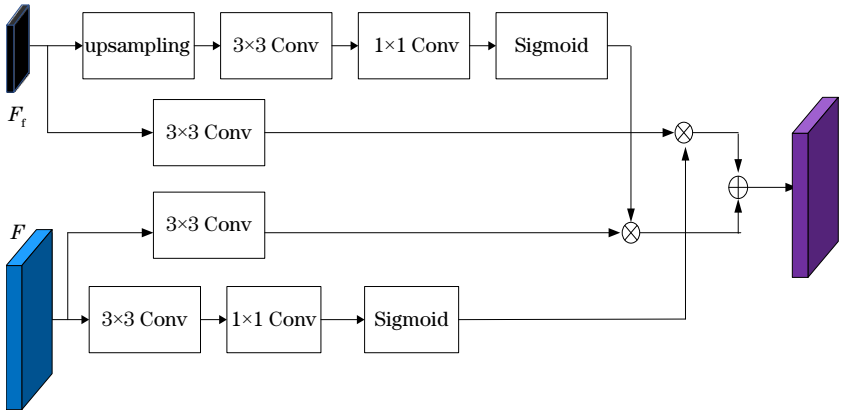

图 4　双边引导特征融合模块结构

Fig. 4　Bilateral guidance feature fusion module structure

块输出的特征图和对应的真值图之间的差异作为损失，优化器根据反向传播的梯度信息来更新网络参数、降低损失，使网络在浅层和深层都得到更好的收敛，指导更多空间和语义信息向最优解靠拢，提高网络各层的收敛能力。特征辅助收敛模块结构如图 5 所示。

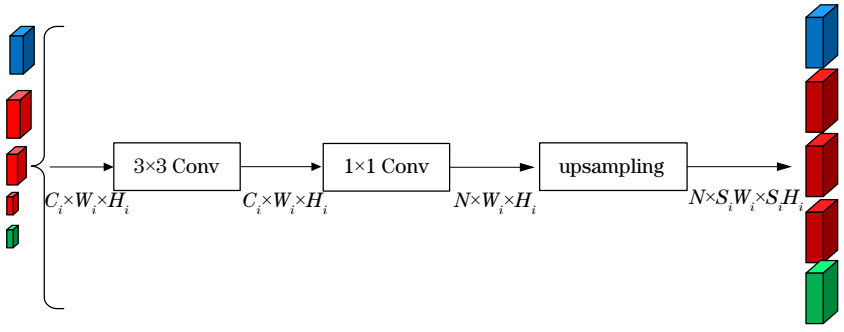

图 5　特征辅助收敛模块结构

Fig. 5　Feature-assisted convergence module structure

将骨干网络、3 层多尺度感知层和特征融合网络输出特征图作为输入，记为 $C_i\times W_i\times H_i$。通过 3×3 卷积和 1×1 卷积在不改变特征图大小的前提下，将各个特征图的通道数压缩至类别数 N，结果特征图记为 $N\times W_i\times H_i$，最后通过上采样将各个特征图以对应的上采样系数 S_i 放大至原图 I 尺寸，最终特征图记为 $N\times S_iW_i\times S_iH_i$。采用在线难分样本挖掘交叉熵损失(OHEML)[26]函数计算损失：

$$L_{\text{Total}}=\sum_{i=1}^{5}\text{OHEML}_i(X_i;\ W),\tag{6}$$

式中：OHEML_i 为各个特征辅助收敛模块的损失；X_i 表示各个特征辅助收敛模块输出特征图；W 表示特征图对应的真值图。

2.4　实验分析与讨论

在自建的数据集上采用所提方法和多种方法进行对比实验。实验均在 Ubuntu 系统下进行。训练时深度环境配置如下：显存为 11 GB 的 GeForce_RTX_2080_TiGPU、Python 3.7、CUDA 10.1 和 Pytorch 1.8。测试时深度环境配置如下：Intel(R) Core(TM) i5-9400 CPU、Python 3.7 和 Pytorch 1.8。初始学习率设置为 5×10^{-6}，权重衰减率为 1×10^{-7}，训练次数为 1000。为防止网络出现过拟合现象，对样本图像进行裁剪、翻转和放缩等操作，增加数据集图片数量，提高网络的泛化能力。

2.4.1　图像采集与样本标注

常用的公开机器人抓取数据集 Cornell Grasping Dataset 和 JACQUARD 的图片背景过于简单，易出现网络模型泛化能力较差的问题，很难满足复杂场景抓取的需求；同时，该数据集图像只有一种物品，对多物品实时检测和预测抓取位姿的效果不佳；此外，该数据集标签信息只有抓取框 4 个顶点坐标信息，没有物品类别信息和物品轮廓信息，无法预测物品类别信息，所提基于轮廓点检测的快速位姿估计策略也无法使用；而且该数据集未包含本实验抓取的物品种类。

鉴于此，本研究自建数据集，通过调节工业相机的焦距、对比度参数和曝光度参数，采集不同背景和亮度下的物品图片，同时设置不同拍摄高度和拍摄角度，增加数据集的多样性。采集的图片数量共 1527 张，包括 6 类不同大小、形状的物品，分别为 234 张空调遥控器、245 张裁纸刀、269 张中性笔、238 张剪刀、281 张螺丝和 260 张螺母。

采用 Labelme 软件中的多边形标注工具对图像中

的物品轮廓进行标注，并对标注物品分别命名，其中空调遥控器记为ykq，裁纸刀记为dao，中性笔记为bi，螺丝记为luosi，螺母记为luomu，剪刀记为jianzi。将所有的标注文件保存为json格式，作为标签文件，标签文件包含标注图像的文件名、文件地址、宽高和物品轮廓点坐标信息。

2.4.2 评价指标

选用语义分割中常见的平均交并比（mIoU）指标和执行时间实时性指标来评价网络的性能。执行时间指单张图像的预测时间；mIoU指真实值和预测值两个集合的交集和并集之比：

$$R_{\mathrm{mIoU}}=\frac{1}{K+1}\sum_{i=0}^{K}\frac{P_{ii}}{\sum_{j=0}^{K}P_{ij}+\sum_{j=0}^{K}P_{ji}-P_{ii}},\quad(7)$$

式中：P_{ij}表示属于类i但被预测为类j的像素数量；P_{ji}表示属于类j但被预测为类i的像素数量；P_{ii}表示属于类i被预测为类i的像素数量；$K+1$表示类别数。

2.4.3 网络预测速度和精度实验分析

从网络模型参数量大小、模型预测时间和网络分割精度等3个方面将所提网络和另外6种语义分割网络进行对比，结果如表2所示。

表2 模型参数量大小、模型预测时间和分割精度对比

Table 2 Comparison of model parameters, model prediction time, and segmentation accuracy

Method	Size /MB	Average running time of test image /s	mIoU /%
UNet[18]	94.97	2.23	95.6
PSPNet[19]	9.31	0.36	83.5
BiSeNet[20]	52.52	0.68	91.5
ICNet[22]	31.52	0.96	95.3
Fast-SCNN[23]	4.60	0.32	88.6
BiSeNetV2[21]	20.47	0.51	95.1
Proposed method	19.08	0.41	96.8

表2数据表明，在分割精度方面，所提网络的mIoU指标最高，比UNet高1.2个百分点，比PSPNet提升12.1个百分点，比PSPNet提升5.3个百分点，比ICNet提升1.5个百分点，比Fast-SCNN提升8.2个百分点，比BiSeNetV2提升1.7个百分点。所提方法具有较高分割精度，主要原因有2点：1）所提方法将特征辅助收敛模块嵌入浅层网络和深层网络中，指导空间和语义信息向最优方向靠拢，提高各层网络的收敛能力；2）利用级联多尺度感知层在提取多尺度特征的同时通过注意力机制引导获取对空间位置的感知能力，强化对目标区域的关注程度，提高分割能力。

从模型参数量大小、模型预测时间两个方面分析：UNet结构最大，模型预测时间也最慢；Fast-SCNN结构最小，模型预测时间最快；BiSeNet、BiSeNetV2和ICNet均采用多分支结构，模型网络结构较大，预测速度较慢。所提模型预测速度和Fast-SCNN和PSPNet相差很小，满足实时性要求。所提网络预测速度相比其他对比网络预测速度快的原因主要有3点：1）采用单分支网络提取空间特征信息和语义特征信息；2）改进VGG前七层卷积，减小卷积输出特征图通道数；3）在多尺度感知层中采用1×1卷积实现跨通道信息交互的同时，减小网络参数量。综上所述，所提网络的总体性能优于其他网络。

2.4.4 多尺度感知层层数精度对比实验分析

本小节从多尺度感知层层数对分割精度和预测速度的影响进行叙述和分析，实验结果如表3所示。

表3 多尺度感知层层数精度和预测速度对比

Table 3 Accuracy and prediction speed comparison of multi-scale perception layers

Method	mIoU /%	Average running time of test image /s
Proposed method(one MSPL)	91.5	0.32
Proposed method(two MSPL)	95.1	0.36
Proposed method(three MSPL)	96.8	0.41
Proposed method(four MSPL)	97.2	0.48

由表3的实验数据可知，随着MSPL层数的增加，mIoU指标越高。即多尺度感知层有助于网络学习关注区域，提高分割精度。1～3层MSPL的网络分割精度增长幅度较大，4层MSPL的网络分割精度增长幅度很小，同时预测速度有较大的下降。综合精度和实时性指标考虑，3层MSPL组成的上下文感知结构对于动态物品抓取任务更具有优势。

3 物品抓取点位姿估计

3.1 基于轮廓点检测的快速位姿估计策略

主成分分析法（PCA）是一种常用的数据降维方法，在数据空间中重新构建新坐标空间，使数据在新坐标空间的各个维度方向上方差最大。借鉴主成分分析法思想，本研究提出一种基于轮廓点检测的快速位姿估计策略，该策略抓取点位姿时预测准确、速度快。具体预测目标物品抓取点位姿流程如下：首先对分割结果图进行预处理，采用图像二值化和降噪处理去除分割图中可能存在的噪声；随后对分割图中的目标物品进行轮廓检测，保存轮廓点坐标；最后根据轮廓点坐标信息采取轮廓点位姿估计方法确定机器人抓取位置坐标和姿态角度。

1）用N个二维轮廓点像素坐标组成$N\times2$矩阵，记为$\boldsymbol{M}_0$。对矩阵沿列方向求均值，将得到的1×2矩阵中的元素作为机器人抓取点像素坐标位置P。

$$\boldsymbol{M}_{\mathrm{O}}=\begin{bmatrix} X_1 & Y_1 \\ X_2 & Y_2 \\ X_3 & Y_3 \\ \vdots & \vdots \\ X_N & X_N \end{bmatrix}, \tag{8}$$

$$P=(X,Y)=\frac{1}{N}\sum_{i=1}^{N}(X_i,Y_i)。 \tag{9}$$

2)对 $N\times 2$ 矩阵去中心化，即沿列方向分别减去 P 的 X 轴坐标值和 Y 轴坐标值，得到新的 $N\times 2$ 矩阵，记为 $\boldsymbol{M}_{\mathrm{R}}$。

$$\boldsymbol{M}_{\mathrm{R}}=\begin{bmatrix} X_1-X & Y_1-Y \\ X_2-X & Y_2-Y \\ X_3-X & Y_3-Y \\ \vdots & \vdots \\ X_N-X & X_N-Y \end{bmatrix}。 \tag{10}$$

3)将 $\boldsymbol{M}_{\mathrm{R}}$ 的协方差矩阵，记为 $\boldsymbol{M}_{\mathrm{C}}$，$\boldsymbol{M}_{\mathrm{C}}$ 对应的特征值记为 λ_1、λ_2，特征向量记为 $\boldsymbol{V}_1$、$\boldsymbol{V}_2$。

$$\boldsymbol{M}_{\mathrm{C}}=\frac{1}{N-1}\boldsymbol{M}_{\mathrm{R}}\boldsymbol{M}_{\mathrm{R}}{}^{\mathrm{T}}, \tag{11}$$

$$\lambda_1,\lambda_2=\mathrm{eigenvalues}(\boldsymbol{M}_{\mathrm{C}}), \tag{12}$$

$$\boldsymbol{V}_1,\boldsymbol{V}_2=\mathrm{eigenvectors}(\boldsymbol{M}_{\mathrm{C}})。 \tag{13}$$

4)选取最大的特征值对应的特征向量，记为 $\boldsymbol{V}=(x_1,y_1)$。对 $\boldsymbol{V}$ 中的两个元素 x_1、y_1，通过反正切得出机器人抓取点姿态弧度，采用 degrees 函数将弧度转为角度，作为抓取点姿态角度，记为 θ。

$$\theta=\mathrm{degrees}\left[\arctan(x_1,y_1)\right]。 \tag{14}$$

3.2　位姿估计实验与分析

采用中性笔、裁纸刀、遥控器、剪刀、螺丝和螺母进行实验，实验结果如图 6 所示。

图 6(b)中物品主方向轴线和水平轴线组成的夹角为抓取点姿态角度：若主方向轴线在水平轴线上方，抓取角度为负值；若主方向轴线在水平轴线下方，抓取角度为正值，两轴线的交点为抓取点位置坐标。表 4 为图 6 物品抓取点位置坐标、姿态角度和预测时间具体数据。结合图 6 和表 4 的数据，提出一种基于轮廓点检测的快速位姿估计策略，由该策略预测的位姿结果和物品实际位姿基本一致，同时预测时间短，满足抓取位姿精度和速度要求。

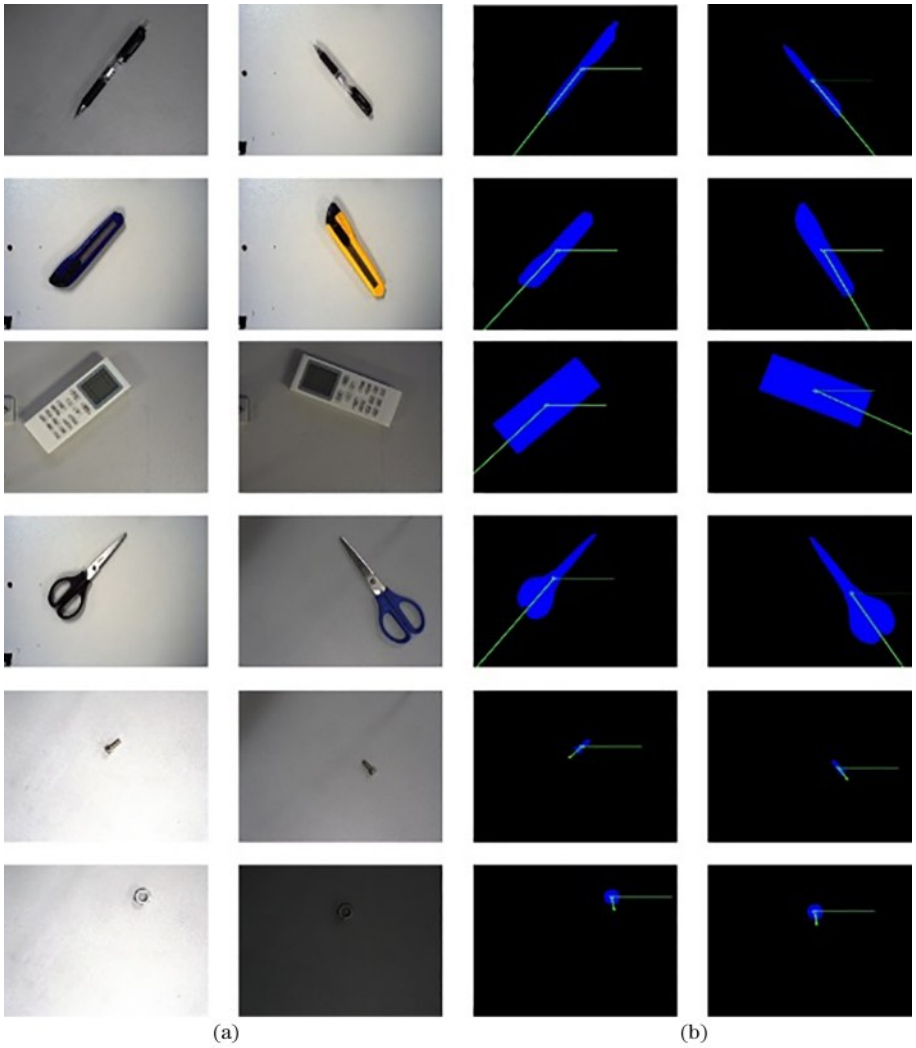

图 6　物品位姿估计实验结果。(a)原图；(b)位姿估计结果图

Fig. 6　Experimental results of object pose estimation. (a) Original images; (b) pose estimation result images

表 4　物品抓取点位姿数据表

Table 4　Pose data table of item grabbing point

Thing	Posture		Predict time /s
	Coordinate	Angle /(°)	
Gel pen	(545,433)	120.23	0.015
	(526,517)	58.34	
Paper knife	(422,485)	124.55	0.014
	(463,491)	−29.07	
Remote control	(368,432)	128.56	0.015
	(538,331)	31.20	
Scissors	(407,420)	122.96	0.016
	(544,399)	−34.17	
Screw	(544,379)	127.37	0.014
	(654,522)	59.88	
Nut	(695,217)	82.24	0.016
	(537,318)	84.25	

4　机器人物品抓取实验

搭建一个小型流水线实验平台，模拟工厂流水线环境，对流水线上的物品进行抓取实验。实验平台如图 7 所示，由 ABB 公司制造的双臂机器人、传送带、打包盒、大恒水星二代工业相机和不同种类的物品组成。双臂机器人负责物品抓取；工业相机负责物品图像采集；传送带负责运载物品；打包盒负责物品包装。具体的抓取实验流程如下：

1）实时物品图像采集

利用安装在机器人本体上的工业相机对流水线上的物品实时图像采集，保证相机成像平面和传送带平面平行。

2）位姿预测

将实时采集的图像输入所提两阶段动态多物品定位抓取方法中预测目标物品抓取点位姿，记为 $(X_{\mathrm{cam}}, Y_{\mathrm{cam}}, \theta)$。

3）坐标变换

所提定位抓取方法预测的位姿为 2D 图像坐标系下的抓取点位姿，通过各个坐标系间的关系获取机器人基坐标系下抓取位姿。图像坐标系与机器人坐标系之间的转换关系为

$$\begin{bmatrix} X_{\mathrm{robot}} \\ Y_{\mathrm{robot}} \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}_{\mathrm{cam2robot}} & \boldsymbol{T}_{\mathrm{cam2robot}} \\ \boldsymbol{0} & \boldsymbol{1} \end{bmatrix} \begin{bmatrix} X_{\mathrm{cam}} \\ Y_{\mathrm{cam}} \\ 1 \end{bmatrix}, \tag{15}$$

式中：$(X_{\mathrm{cam}}, Y_{\mathrm{cam}})$为图像坐标系下物品的抓取点坐标；$(X_{\mathrm{robot}}, Y_{\mathrm{robot}})$为机器人坐标系下的物品抓取点坐标；$\boldsymbol{R}_{\mathrm{cam2robot}}$、$\boldsymbol{T}_{\mathrm{cam2robot}}$为图像坐标系和机器人坐标系间的旋转矩阵和平移矩阵，通过九点标定法[27]获取。

4）物品抓取搬运

设机械臂在机器人坐标下的初始位姿为 $(X_0, Y_0, Z_0, a_0, b_0, c_0)$。首先控制机器人末端执行器到达物品抓取点正上方，位姿记为 $(X_{\mathrm{robot}}, Y_{\mathrm{robot}}, Z_1, a_0, b_0, c_0)$；随后末端执行器旋转预测的角度，位姿记为 $(X_{\mathrm{robot}}, Y_{\mathrm{robot}}, Z_1, a_0+\theta, b_0, c_0)$；最后末端执行器下移夹具闭合抓取物品，将其放置于打包盒中，机械臂返回初始位姿，准备下一次抓取。

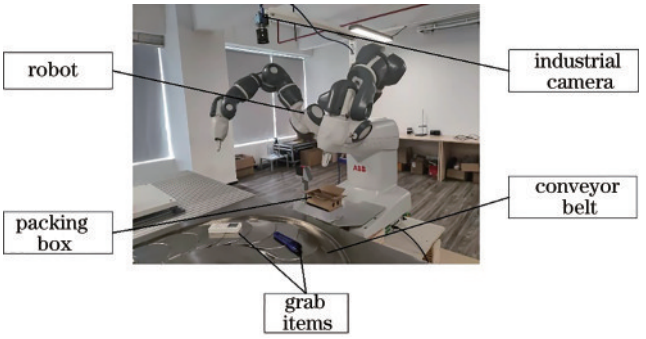

图 7　流水线实验平台

Fig. 7　Pipeline experiment platform

设计 3 组实验，分别记为第 1 组、第 2 组、第 3 组，对传送带上的 6 类物品进行抓取实验。第 1 组采用 BiSeNet 和基于轮廓点检测的位姿估计策略结合的方法；第 2 组采用 BiSeNetV2 和基于轮廓点检测的位姿估计策略结合的方法；第 3 组采用所提定位抓取方法。采用抓取成功率作为机器人抓取系统的性能评价指标：

$$R_{\mathrm{success}} = N_{\mathrm{successful\ grabs}} / N_{\mathrm{grabbing\ tests}}。 \tag{16}$$

部分成功抓取的结果如图 8 所示，实验成功率如表 5、表 6 和表 7 所示，所提定位抓取方法在遥控器、裁纸刀、剪刀、螺丝、螺母和中性笔抓取实验中，总体优于

表 5　第 1 组抓取试验结果

Table 5　First group of grasping test results

Thing	Number of experiments	Number of successful crawls	Success rate /%
Remote control	30	29	96.7
Paper knife	30	27	90
Scissors	30	28	93.3
Screw	30	24	80
Nut	30	26	86.7
Gel pen	30	28	93.3

表 6　第 2 组抓取试验结果

Table 6　Second group of grasping test results

Thing	Number of experiments	Number of successful crawls	Success rate /%
Remote control	30	29	96.7
Paper knife	30	28	93.3
Scissors	30	28	93.3
Screw	30	26	86.7
Nut	30	26	86.7
Gel pen	30	28	93.3

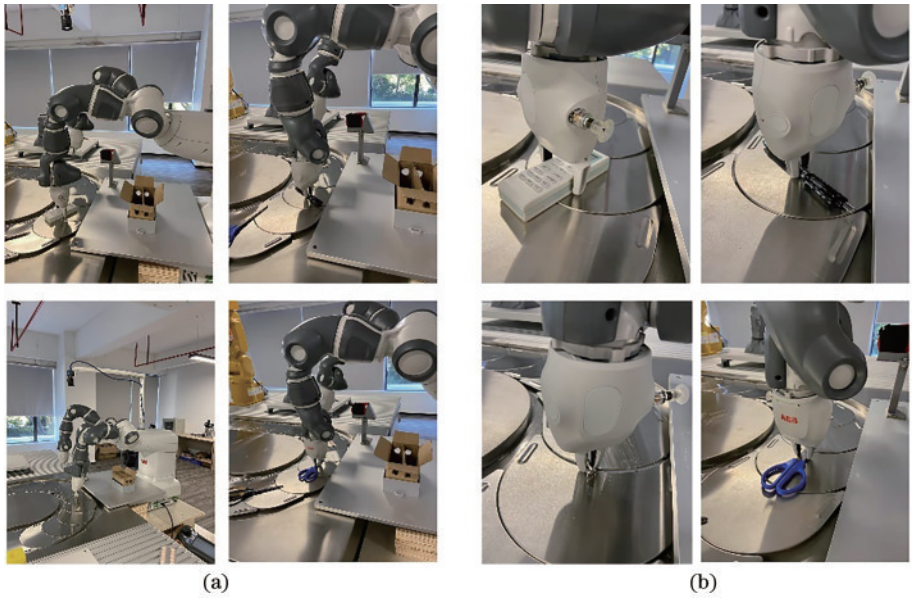

图 8　机器人物品抓取图。(a)整体图;(b)局部图

Fig. 8　Robot object grabbing diagrams. (a) Overall pictures; (b) partial pictures

表 7　第 3 组抓取试验结果

Table 7　Third group of grasping test results

Thing	Number of experiments	Number of successful crawls	Success rate /%
Remote control	30	30	100
Paper knife	30	28	93.3
Scissors	30	28	93.3
Screw	30	28	93.3
Nut	30	28	96.7
Gel pen	30	29	96.7

其他两种方法,在遥控器抓取实验中甚至达到 100% 的抓取成功率。即所提定位抓取方法能实时检测和预测动态多类物品抓取位姿,精准完成物品抓取。

5　结　　论

设计一种两阶段动态多物品定位抓取方法。第 1 阶段提出多尺度上下文感知的单分支融合语义分割网络:首先特征提取网络采用单分支网络获取空间和语义信息;其次特征融合网络通过双边引导特征融合模块以空间和语义信息相互指导方式充分融合,提升特征组合表达能力;最后为进一步提升分割精度,提出特征增强网络,将特征辅助收敛模块嵌入单分支网络中,加速浅层网络和深层网络收敛。第 2 阶段,在掩码区域采用基于轮廓点检测的快速位姿估计策略预测抓取位姿,预测速度快,位姿信息准确。

为验证所提定位抓取方法效果,在自建的物品数据集上进行对比实验,所提网络相比于其他网络,性能更好。同时搭建小型流水线实验平台,将定位抓取算法部署到双臂机器人上,对流水线上多类物品进行抓取实验。实验结果表明,所提定位抓取方法能实时检测和预测物品抓取位姿,精准完成物品抓取。因工业 RGB 相机无法获取相机到抓取点间的距离,本研究的抓取高度是固定的。后期将采用深度相机,研究 3D 测距算法,实时获取机器人抓取高度。

参　考　文　献

[1] 王吉岱, 王明鹏. 基于视觉引导的自动码放生产线设计[J]. 包装工程, 2017, 38(11): 148-152.
Wang J D, Wang M P. Design of automatic stacking of production line based on visual guidance[J]. Packaging Engineering, 2017, 38(11): 148-152.

[2] 宋薇, 仇楠楠, 沈林勇, 等. 面向工业零件的机器人单目立体匹配与抓取[J]. 机器人, 2018, 40(6): 950-957.
Song W, Qiu N N, Shen L Y, et al. The monocular stereo matching and grasping of robot for industrial parts [J]. Robot, 2018, 40(6): 950-957.

[3] 张震, 张照崎, 朱留存, 等. 一种基于 Shi-Tomasi 和改进 LBP 的特征匹配及目标定位快速算法[J]. 吉林大学学报(理学版), 2021, 59(5): 1171-1178.
Zhang Z, Zhang Z Q, Zhu L C, et al. A fast algorithm for feature matching and target location based on Shi-Tomasi and improved LBP[J]. Journal of Jilin University (Science Edition), 2021, 59(5): 1171-1178.

[4] Shi J B, Tomasi. Good features to track[C]//1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 21-23, 1994, Seattle, WA, USA. New York: IEEE Press, 1994: 593-600.

[5] 蔡雨, 黄学功, 张志安, 等. 基于特征融合的实时语义分割算法[J]. 激光与光电子学进展, 2020, 57(2): 021011.
Cai Y, Huang X G, Zhang Z A, et al. Real-time semantic segmentation algorithm based on feature fusion technology[J]. Laser & Optoelectronics Progress, 2020, 57(2): 021011.

[6] 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.

[7] 乔婷, 苏寒松, 刘高华, 等. 基于改进的特征提取网络的目标检测算法[J]. 激光与光电子学进展, 2019, 56(23): 231008.
Qiao T, Su H S, Liu G H, et al. Object detection algorithm based on improved feature extraction network[J]. Laser & Optoelectronics Progress, 2019, 56(23): 231008.

[8] 冯佳萌, 裴东, 邹勇, 等. 基于机器人激光定位的一种改进 AMCL 算法[J]. 激光与光电子学进展, 2021, 58(20): 2028003.
Feng J M, Pei D, Zou Y, et al. An improved AMCL algorithm based on robot laser localization[J]. Laser & Optoelectronics Progress, 2021, 58(20): 2028003.

[9] 朱江, 杜瑞, 李建奇, 等. 基于注意力机制的曲轴瓦盖上料机器人视觉定位和检测方法[J]. 仪器仪表学报, 2021, 42(5): 140-150.
Zhu J, Du R, Li J Q, et al. Visual location and detection method of crankshaft bearing cap feeding robot based on attention mechanism[J]. Chinese Journal of Scientific Instrument, 2021, 42(5): 140-150.

[10] 孙雄峰, 林浒, 王诗宇, 等. 基于改进 Faster RCNN 的工业机器人分拣系统[J]. 计算机系统应用, 2019, 28(9): 258-263.
Sun X F, Lin H, Wang S Y, et al. Industrial robots sorting system based on improved Faster RCNN[J]. Computer Systems & Applications, 2019, 28(9): 258-263.

[11] 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.

[12] 陈丹, 林清泉. 基于级联式 Faster RCNN 的三维目标最优抓取方法研究[J]. 仪器仪表学报, 2019, 40(4): 229-237.
Chen D, Lin Q Q. Research on 3D object optimal grasping method based on cascaded Faster RCNN[J]. Chinese Journal of Scientific Instrument, 2019, 40(4): 229-237.

[13] 李秀智, 李家豪, 张祥银, 等. 基于深度学习的机器人最优抓取姿态检测方法[J]. 仪器仪表学报, 2020, 41(5): 108-117.
Li X Z, Li J H, Zhang X Y, et al. Detection method of robot optimal grasp posture based on deep learning[J]. Chinese Journal of Scientific Instrument, 2020, 41(5): 108-117.

[14] 夏晶, 钱堃, 马旭东, 等. 基于级联卷积神经网络的机器人平面抓取位姿快速检测[J]. 机器人, 2018, 40(6): 794-802.
Xia J, Qian K, Ma X D, et al. Fast planar grasp pose detection for robot based on cascaded deep convolutional neural networks[J]. Robot, 2018, 40(6): 794-802.

[15] Dai J, Li Y, He K, et al. R-FCN: object detection via region-based fully convolutional networks[C]//The 30th Conference on Neural Information Processing Systems, December 5-10, 2016, Barcelona, Spain. Canada: NIPS, 2016: 379-387.

[16] 陈奎烨, 史旭华, 徐铭泽. 基于 GPR 和 KRR 组合模型的机械臂抓取研究[J]. 传感器与微系统, 2021, 40(1): 34-38.
Chen K Y, Shi X H, Xu M Z. Research on manipulator grasping based on GPR and KRR combined model[J]. Transducer and Microsystem Technologies, 2021, 40(1): 34-38.

[17] 王德明, 颜熠, 周光亮, 等. 基于实例分割网络与迭代优化方法的 3D 视觉分拣系统[J]. 机器人, 2019, 41(5): 637-648.
Wang D M, Yan Y, Zhou G L, et al. 3D vision-based picking system with instance segmentation network and iterative optimization method[J]. Robot, 2019, 41(5): 637-648.

[18] 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.

[19] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6230-6239.

[20] Yu C Q, Wang J B, Peng C, et al. BiSeNet: bilateral segmentation network for real-time semantic segmentation [M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11217: 334-349.

[21] Yu C Q, Gao C X, Wang J B, et al. BiSeNetV2: bilateral network with guided aggregation for real-time semantic segmentation[J]. International Journal of Computer Vision, 2021, 129(11): 3051-3068.

[22] Zhao H S, Qi X J, Shen X Y, et al. ICNet for real-time semantic segmentation on high-resolution images[M]// Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11207: 418-434.

[23] Poudel R P, Liwicki S. Fast-SCNN: fast semantic segmentation network[C]//The 30th British Machine Vision Conference, September 9-12, Wales, UK. London: BMCV, 2019: 308-315.

[24] Sandler M, Howard A, Zhu M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4510-4520.

[25] Liu W Z, Salzmann M, Fua P. Context-aware crowd counting[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 5094-5103.

[26] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 761-769.

[27] Tsai R Y, Lenz R K. A new technique for fully autonomous and efficient 3D robotics hand/eye calibration [J]. IEEE Transactions on Robotics and Automation, 1989, 5(3): 345-358.