中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

# 基于Transformer-CNN混合架构的跨模态融合抓取检测

王勇, 李邑灵, 苗夺谦, 安春艳, 袁鑫林

引用本文:
王勇,李邑灵,苗夺谦,安春艳,袁鑫林. 基于Transformer–CNN混合架构的跨模态融合抓取检测[J]. 控制与决策, 2024, 39(11): 3607–3616.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2023.1152

---

# 基于Transformer-CNN混合架构的跨模态融合抓取检测

王 勇[1†], 李邑灵[1], 苗夺谦[2], 安春艳[1], 袁鑫林[1]

(1. 重庆理工大学 两江人工智能学院，重庆 401135；2. 同济大学 电子与信息工程学院，上海 200092)

**摘 要:** 在机器臂抓取检测领域,RGB 图像和深度图像的处理效率仍有很大提升空间. 鉴于此,提出一种基于Transformer-CNN 混合架构的新型跨模态交互融合的机械臂抓取检测方法. 为了充分利用RGB 和深度图像的特征信息,开发一种高效的跨模态特征交互融合模块,用来校准RGB 和深度图像相对应的特征信息,并交互增强双模态的特征. 此外,设计一种 Transformer 与 CNN 并行的网络模块,结合 CNN 的局部建模能力和 Transformer 的全局建模能力,获得更好的特征表示,从而提高抓取检测性能. 实验结果表明,所提方法在Cornell 与 Jacquard 抓取数据集上分别达到了 99.1 % 和 96.2 % 的准确率. 在真实场景下的抓取检测实验验证了所提方法可以有效预测各种场景下物品的抓取位置.

**关键词:** 机械臂抓取检测；跨模态；RGB-D融合；Transformer；CNN

**中图分类号:** TP29 **文献标志码:** A

**DOI:** 10.13195/j.kzyjc.2023.1152

**引用格式:** 王勇,李邑灵,苗夺谦,等. 基于Transformer-CNN混合架构的跨模态融合抓取检测[J]. 控制与决策, 2024, 39(11): 3607-3616.

# Cross-modal interaction fusion grasping detection based on Transformer-CNN hybrid architecture

*WANG Yong*[1†], *LI Yi-ling*[1], *MIAO Duo-qian*[2], *AN Chun-yan*[1], *YUAN Xin-lin*[1]

(1. School of Artificial Intelligence，Chongqing University of Technology，Chongqing 401135，China；2. College of Electronic and Information Engineering，Tongji University，Shanghai 200092，China)

**Abstract:** In the field of robotic grasping detection, there is still great room for improvement in the processing efficiency of RGB and depth images. This article proposes a novel RGB-D cross modal interactive fusion method for robotic grasping detection based on a Transformer-CNN hybrid architecture. In order to fully utilize the feature information of RGB and depth images, an efficient cross modal feature interaction fusion module has been developed, which can calibrate the corresponding feature information of RGB and depth images and interactively enhance the bimodal features. In addition, a parallel network module between Transformer and CNN is designed to combine the local modeling ability of CNN and the global modeling ability of Transformer to obtain better feature representation and improve the performance of grab detection. The experimental results show that this method achieves an accuracy of 99.1 % and 96.2 % on the Cornell dataset and Jacquard dataset, respectively. The grasp detection experiments in real scenes verify that the proposed method can effectively predict the grasp pose of objects in various scenarios.

**Keywords:** robotic grasping detection；cross-modal；RGB-D fusion；Transformer；CNN

## 0 引 言

对于机器人,正确感知位置随机的物体并找到不同形状物体的最佳抓取位置是一个极具挑战性的问题. 不恰当的抓取位置很容易导致错误操作,因此一种准确、快速的抓取检测方法对于机器人的抓取至关重要. 许多传统的抓取检测方法[1-3]依赖于物体本身的形状和物理特性,需要在稳定的外部环境、摩擦力、刚度和精确的三维物体模型的假设下进行抓取. 由于抓取对象和环境的多样性,这些方法在实际应用中往往表现不佳.

随着深度学习技术的兴起,越来越多的深度学习方法[4-8]被应用于抓取检测领域. 为了提高检测性能并增强对抓取场景信息描述的准确性,许多研究[6-8]采用深度图像作为RGB 图像的一种空间信息补充,同时作为抓取检测网络的输入. 然而,RGB 图像和深度图像之间存在实质性的模态差异. 具体而言,RGB 图像描述的是物体的颜色、纹理和外观等信息,而深度图像则更注重物体的三维几何信息,并且场景的光

**收稿日期:** 2023-08-14；**录用日期:** 2023-12-19.

**责任编委:** 訾斌.

†通讯作者. E-mail: ywang@cqut.edu.cn.

照和色彩的变化对深度图像的影响更小[9]. 如何对这两种模态的信息进行校准,并将它们统一为有效的图像特征信息,仍然是一个待解决的问题. 此外,由于物体材料不同和测距范围有限,加上深度相机测量的不确定性,深度图像中往往伴随着噪声. 如果直接将充满噪声的深度图像与RGB图像级联作为网络输入,不仅不能增强视觉信息的描述,而且还会增加网络处理噪声数据的负担. 为了将RGB和深度图像信息统一为有效的抓取检测特征表示,并减少深度图像中噪声对抓取检测的影响,本文开发一种高效的跨模态特征交互融合模块. 该模块可以充分利用不同模态间互补特征的潜力,实现更好的双模态特征融合,提高抓取检测网络的性能.

许多深度学习的抓取检测方法采用的是基于卷积神经网络(CNN)的架构. CNN可以有效地提取图像的局部特征,但难以捕获全局表示[10-12]. 最近,随着Transformer[13]的发展,也出现了一些基于Transformer的抓取检测研究[11-12]. 这些方法试图利用自注意力机制来捕获全局上下文信息,并创建目标的远程依赖关系,以获得更强的全局表示. 然而,Transformer忽略了局部特征细节[10,12],这降低了背景与待抓取物品之间的可分辨性. 本文提出一个双流并行的Transformer-CNN混合模块,有效结合CNN与Transformer的优点,从而增强网络特征提取能力并提高抓取检测的准确率.

本文的主要研究内容和贡献可以总结如下:

1) 提出一种高效的跨模态特征交互融合模块,可以互补融合RGB图像与深度图像的特征信息,减少噪声影响,实现更好的双模态特征提取和交互.

2) 提出一种双流并行的Transformer-CNN混合网络架构,有效地结合了Transformer和CNN的优点,很好地保留了图像的局部特征和全局表示.

3) 实验表明,该方法在Cornell[6]和Jacquard[14]单目标数据集上能达到99.1%和96.2%的高抓取准确率,同时还能实现较快的检测速率(28.6帧/秒). 在多目标数据集和实际场景下的检测也能准确地检测出目标的抓取位置.

## 1 相关工作

在深度学习方法流行之前,人们大多使用手工编写抓取规则或构建3D模型等方法实现抓取检测. 陈友东等[1]利用高斯混合模型构建物体空间位姿到机械臂关节角度的映射,可以实现快速部署,但抓取精度不高. 谢宇珅等[3]采用基于随机树分类的特征点匹配算法来识别和定位物品,不断修正机械臂实施抓取. 值得注意的是,该方法只有在物品形状比较规则时才能够有较好的抓取效果. 机械臂的操作取决于预先设计的抓取规则,在不确定的抓取场景中面对不同的抓取对象,机械臂抓取规划的设计者很难设计出合理的规则和三维模型.

与传统的抓取检测方法相比,基于深度学习的抓取检测方式更快速、更准确. Lenz等[6]提出的两步级联模型是深度学习在机器人抓取检测中的较早应用之一,但这种两阶段的检测方法实时性较差. 为了提升检测速率,Redmon等[7]试图将抓取检测视为一项回归任务,以实现端到端的训练和检测. 然而,这种基于回归的单阶段检测方法往往对环境变化特别敏感,并且多目标检测的性能较差. Morrison等[8]提出了一种与对象无关的抓取检测方法,该方法预测每个像素的抓取质量和姿态. 这种像素级的抓取检测方法被广泛应用在后续的抓取检测方法中. SE-ResUNet[15]和TDMAG-Net[16]分别使用通道注意力机制和空间注意力机制增强了抓取检测网络的性能.

许多抓取检测方法将研究重心集中在网络架构的设计,而忽略了对网络输入数据本身的处理. 如何提取和组合不同模态信息仍然值得研究. Redmon等[7]将深度图像信息替换RGB图像的蓝色通道作为图像输入. 许多方法[6,8,15-16]将三通道的RGB图像与单通道的深度图像进行组合作为网络的输入. 这些级联处理方法有效提升了抓取准确率,但是忽略了RGB图像和深度图像的本质差异. 图像的深度数据与RGB数据并不是良好对齐的[9],并且深度图像在拍摄时往往会丢失一些信息还同时伴随着噪声,这种简单的级联操作限制了不同模态特征的潜在性能增益. Kumra等[17]选择两个并行的残差网络分别提取RGB和深度特征,然后将两个特征流融合,这种并行的特征融合网络增强了特征表示,但是缺少中间过程的特征校准,细节特征仍然有待提升.

Transformer因其优秀的全局建模能力在计算机视觉领域受到广泛关注. Wang等[11]使用纯Swin-Transformer网络架构,结合模块之间的跳过连接实现抓取检测,取得了良好的抓取检测结果. 但这种全Transformer网络架构对图像的局部特征细节提取能力较弱,不利于对不规则物品的抓取检测. Dong等[12]和Niu等[18]采用了Transformer的编码器模块对图像特征信息进行编码,在CNN构建的解码器模块中对编码后的特征进行解码,以获得抓取预测. Zhang等[19]在Transformer编码器的基础上利用CNN模块提取编码器各层级的特征进行融合,在一定程度上增强了模型的特征聚合能力. 上述Transformer和CNN

结合的网络设计[12,18-19]值的借鉴,但特征提取的编码器是由全 Transformer 架构组成,没有充分利用 CNN 的局部建模能力,局部特征提取能力仍然表现不足.

## 2　Transformer-CNN 混合架构的跨模态融合机械臂抓取检测

### 2.1　基于关键点像素级的抓取检测表示

受 GG-CNN[8] 启发,设计一种基于关键点像素级的抓取表示. 对于检测图像 IMG 中的每一个像素点都与一个潜在的抓取矩形对应,用 3 个像素图来描述 IMG 中的所有潜在抓取,有

$$G=\{S,A,W\}\in\mathbb{R}^{3\times w_i\times h_i}. \tag{1}$$

其中: $S$、$A$、$W\in\mathbb{R}^{w_i\times h_i}$, $w_i$ 和 $h_i$ 是 IMG 的宽度和高度. $S$ 表示抓取分数图,每个像素点 $p(x,y)$ 的值表示 IMG 中相同位置像素点潜在抓取矩形 $G_p$ 的抓取分数 $S_p$,其范围为 $[0,1]$. $S_p$ 的值越高,则 $G_p$ 越合适抓取物体. $A$ 表示抓取角度图,像素点的值对应抓取矩形的旋转角度 $\theta$. $W$ 表示抓取宽度图,像素点的值对应抓取矩形的宽度 $w$,矩形的高设置为 $w$ 的一半.

### 2.2　算法网络架构设计

基于 Transformer-CNN 混合架构的跨模态融合抓取检测方法的网络 CMF-Grasp (cross-modal interaction fusion grasping detection network) 如图 1 所示. 整个网络架构由编码器 (Encoder)、解码器 (Decoder) 和抓取预测模块 (Detection) 组成. Encoder 用于对输入图像进行下采样编码提取抓取检测图像的特征信息, Decoder 对这些特征信息进行上采样解码分析,最后在 Detection 中得到 $S$、$A$、$W$ 三个像素图. 取 $S$ 图中值最大的点 $p(x,y)$ 作为抓取矩形的中心,由 $A$ 和 $W$ 图中与 $p$ 点对应位置的像素值得到抓取矩形的旋转角度 $\theta$ 和矩形的宽度 $w$ 和高度.

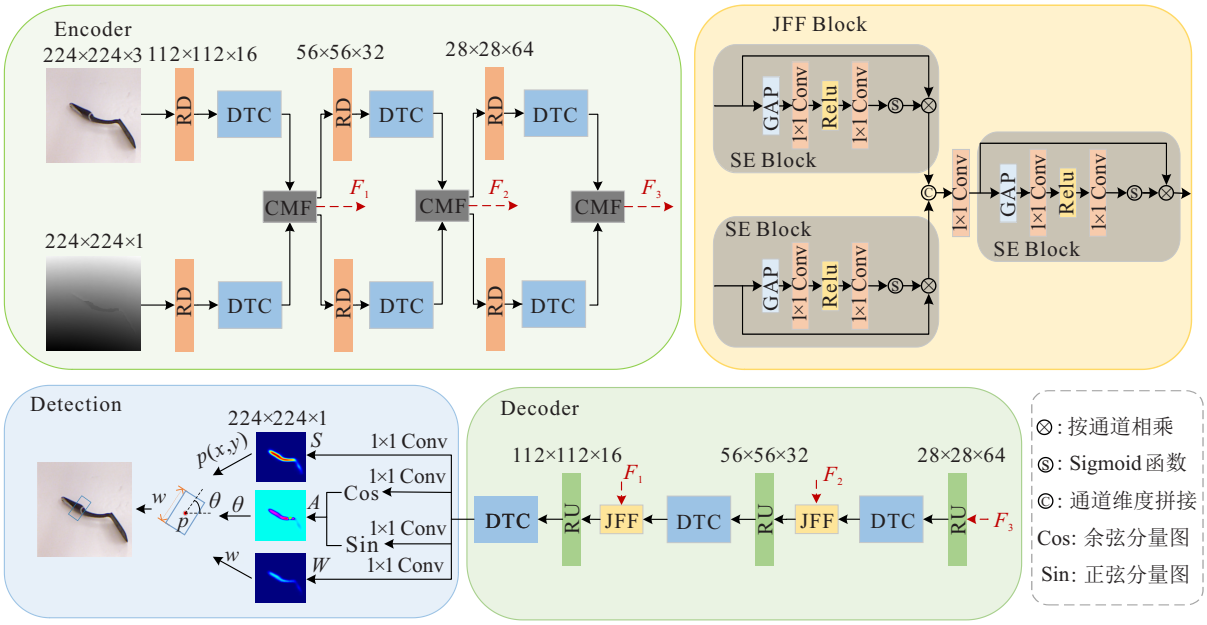


图 1　CMF-Grasp 网络结构

#### 2.2.1　残差下采样 (RD) 和残差上采样 (RU) 模块

Encoder 和 Decoder 中的下采样和上采样分别在 RD (residual downsampling) 和 RU (residual upsampling) 模块中完成,它们的模块设计如图 2 所示. 残差网络[20]在图像分类[21]、目标检测[22]和图像分割[23]等应用中非常有效. 为了利用其卓越的性能, RD 和 RU 模块也采用了残差网络的设计. 与原始残差网络不同的是, RD 和 RU 采用泛化性更强的 Leaky-Relu[24] 激活函数和更加稳定通用的 FRN (filter response normalization)[25] 归一化层替换 Relu 和 BN (batch normalization) 层. RU 中采用了 PixelShuffle[26] 操作来扩大分辨率. 在实验过程中,这

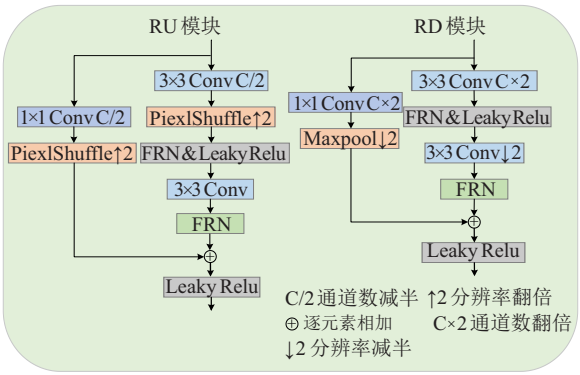


图 2　残差上采样模块 (左) 和残差下采样模块 (右)

种残差网络的设计相比于原始的残差网络模型,其收敛速度会更快,模型训练也更加稳定.

### 2.2.2 新型双流并行的Transformer-CNN混合网络架构(DTC)

CNN可以更多地关注局部特征信息,而Transformer具有非局部信息的建模能力. DTC(dual stream parallel transformer CNN hybrid network)结合了CNN组成的残差网络和Swin-Transformer(SwinT)模块[27],以利用这两种网络架构的优势.

DTC模块如图3所示. 输入的特征向量 $f^{\text{in}}$ 经过一个 $1\times1$ 卷积层(Conv( ))后被分别输入到SwinT分支(Trans( ))和残差卷积网络分支(Res( ))中进行特征提取获得 $f^{\text{in}}_{\text{res}}$, $f^{\text{in}}_{\text{Trans}}$. 在两个分支分别进行特征提取的时候互不干扰,这样局部和非局部特征可以独立并行处理,最后将Transformer模块提取的全局特征与CNN模块提取的局部特征进行融合,使提取的特征更加全面细致. 将 $f^{\text{in}}_{\text{res}}$, $f^{\text{in}}_{\text{Trans}}$ 拼接(Concat( ))后再用一个 $1\times1$ 的卷积层使其通道数目恢复一致,并与 $f^{\text{in}}$ 融合相加($\oplus$)得到最后的特征输出向量 $f^{\text{out}}$. 整个过程公式化如下:

$$f^{\text{in}}_{\text{res}},\ f^{\text{in}}_{\text{Trans}} = \text{Conv}(f^{\text{in}}), \tag{2}$$

$$f^{\text{out}}_{\text{res}},\ f^{\text{out}}_{\text{Trans}} = \text{Res}(f^{\text{in}}_{\text{res}}),\ \text{Trans}(f^{\text{in}}_{\text{Trans}}), \tag{3}$$

$$f^{\text{out}} = \text{Conv}(\text{Concat}(f^{\text{out}}_{\text{res}},\ f^{\text{out}}_{\text{Trans}})) \oplus f^{\text{in}}. \tag{4}$$

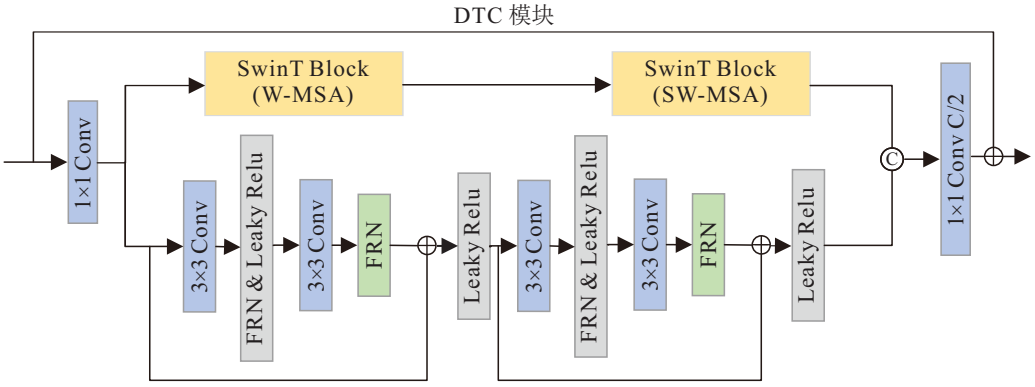


图3 双流Transformer-CNN混合模块

### 2.2.3 高效跨模态特征交互融合模块(CMF)

跨模态特征交互融合模块(cross modal feature interaction and fusion module, CMF)用于对RGB和深度图像特征流进行特征校正,融合不同模态互补特征信息形成增强的特征表示并减少噪声影响.

CMF模块设计如图4所示. 首先,使用全局平均池化(GAP)获得RGB特征图($F_{\text{RGB}}$)和深度特征图($F_{\text{Depth}}$)的全局特征向量. 与ECA-Net[28]计算通道注意力方法类似,将这两个特征向量输入一个 $3\times3$ 的卷积层和Sigmoid激活函数中,以获得通道注意力向量 $\text{Att}^{C}_{\text{RGB}}$ 和 $\text{Att}^{C}_{\text{Depth}}$,分别反映RGB特征和深度特征的重要性. 然后按通道相乘将注意力向量应用于输入特征. 通过这种方式,得到的特征图 $F^{C}_{i}$ 将明确地关注重要的信息,并抑制不必要的信息,加强对场景理解. 此过程可以定义为

$$\text{Att}^{C}_{i} = \text{Sigmoid}(\text{Conv}(\text{AvgPool}(F_i))), \tag{5}$$

$$F^{C}_{i} = \text{Att}^{C}_{i} \otimes F_i. \tag{6}$$

其中:Conv( )表示卷积操作,AvgPool( )表示全局平均池化操作,$i \in [\text{RGB}, \text{Depth}]$,$\otimes$表示按通道相乘.

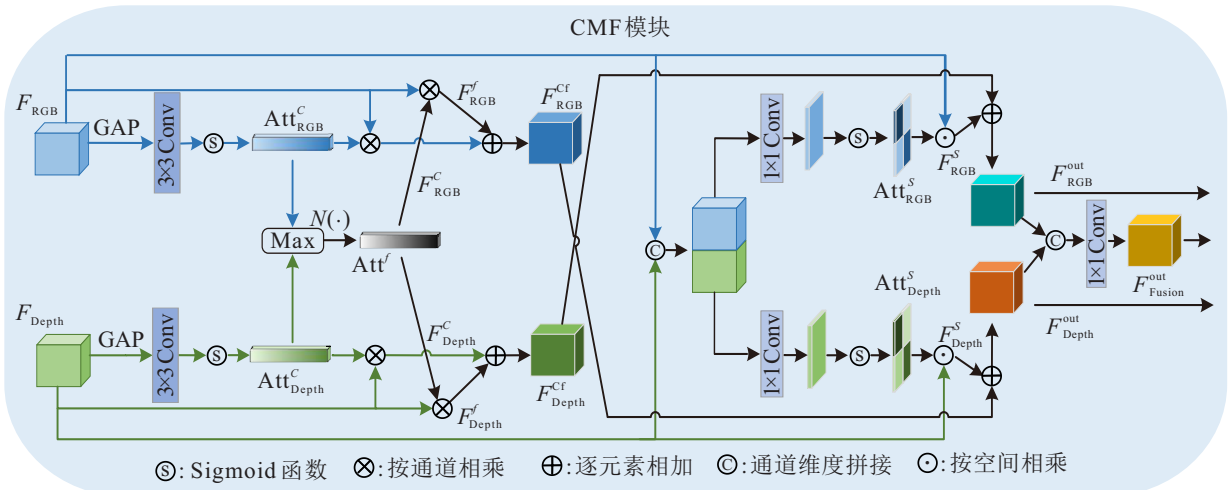


图4 跨模态混合交互融合模块

注意力向量$\text{Att}^C_{\text{RGB}}$和$\text{Att}^C_{\text{Depth}}$通过最大聚合函数(Max)来获得RGB流和深度流中的权重最大特征通道注意力向量,然后对其做归一化运算($N(\,)$)得到交互融合通道注意力向量$\text{Att}^f$. 这有效地抑制了两种模态低质量的特征响应,保留了信息量最大的视觉外观和几何特征. 利用RGB流中的高置信特征来过滤掉相同级别的异常深度特征,同时抑制了深度流中的噪声特征. $\text{Att}^f$分别与$F_{\text{RGB}}$和$F_{\text{Depth}}$进行通道相乘获得通道上相互校准的特征向量与$F^f_{\text{RGB}}$和$F^f_{\text{Depth}}$,然后分别与$F^C_{\text{RGB}}$和$F^C_{\text{Depth}}$相加获得跨模态交互的通道增强特征$F^{\text{Cf}}_{\text{RGB}}$和$F^{\text{Cf}}_{\text{Depth}}$. 此过程可以定义为

$$\text{Att}^f = N(\text{Max}(\text{Att}^C_{\text{RGB}}, \text{Att}^C_{\text{Depth}})), \tag{7}$$

$$F^f_{\text{RGB}} = \text{Att}^f \otimes F_{\text{RGB}}, \tag{8}$$

$$F^f_{\text{Depth}} = \text{Att}^f \otimes F_{\text{Depth}}, \tag{9}$$

$$F^{\text{Cf}}_{\text{RGB}} = F^C_{\text{RGB}} \oplus F^f_{\text{RGB}}, \tag{10}$$

$$F^{\text{Cf}}_{\text{Depth}} = F^C_{\text{Depth}} \oplus F^f_{\text{Depth}}. \tag{11}$$

其中: $N(\,)$和Max$(\,)$分别表示归一化和最大聚合操作,$\oplus$表示逐元素相加.

为了克服不同模态间特征差异性,同时对局部信息的空间特征进行校正,利用两种模态特征的空间相关性进行了跨模态互补聚合. 首先,将$F_{\text{RGB}}$和$F_{\text{Depth}}$拼接,分别用一个1×1卷积层将联合的特征图映射到两个空间权重图中,利用Sigmoid激活函数得到两个互补校准的空间注意力图$\text{Att}^S_{\text{RGB}}$和$\text{Att}^S_{\text{Depth}}$. 将$\text{Att}^S_{\text{RGB}}$和$\text{Att}^S_{\text{Depth}}$与输入特征相乘即得到空间互补校准的增强特征$F^S_{\text{RGB}}$和$F^S_{\text{Depth}}$. 此过程可表示为

$$F_m = \text{Concat}(F_{\text{RGB}}, F_{\text{Depth}}), \tag{12}$$

$$\text{Att}^S_{\text{RGB/Depth}} = \text{Sigmoid}(\text{Conv1} \times 1(F_m)), \tag{13}$$

$$F^S_{\text{RGB}} = \text{Att}^S_{\text{RGB}} \odot F_{\text{RGB}}, \tag{14}$$

$$F^S_{\text{Depth}} = \text{Att}^S_{\text{Depth}} \odot F_{\text{Depth}}, \tag{15}$$

其中$\odot$表示在空间维度相乘. 最后交互联合$F^{\text{Cf}}_{\text{Depth}}$和$F^{\text{Cf}}_{\text{RGB}}$得到$F^{\text{out}}_{\text{RGB}}$和$F^{\text{out}}_{\text{Depth}}$作为下一个模块的输入. 同时,$F^{\text{out}}_{\text{RGB}}$和$F^{\text{out}}_{\text{Depth}}$被进一步级联经过一个1×1卷积层生成跨模态融合特征$F^{\text{out}}_{\text{Fusion}}$作为Decoder的解码特征输入. 此过程可以表示为

$$F^{\text{out}}_{\text{RGB}} = F^{\text{Cf}}_{\text{Depth}} \oplus F^S_{\text{RGB}}, \tag{16}$$

$$F^{\text{out}}_{\text{Depth}} = F^{\text{Cf}}_{\text{RGB}} \oplus F^S_{\text{Depth}}, \tag{17}$$

$$F^{\text{out}}_{\text{Fusion}} = \text{Conv}(\text{Concat}(F^{\text{out}}_{\text{RGB}}, F^{\text{out}}_{\text{Depth}})). \tag{18}$$

在图5中可视化了CMF-Grasp中第1个CMF模块的输入和输出特征图. 采用无参考图像的估值信噪比(SNR)[29]来评估特征图中的噪声水平,使用图像局部方差的最大值和最小值之比作为图像SNR,SNR越大说明图像噪声越少. 可以发现$F^{\text{out}}_{\text{RGB}}$中待抓取对象区域的值基本在0.7以上,与$F_{\text{RGB}}$相比有着明显提升. $F^{\text{out}}_{\text{Depth}}$与$F_{\text{Depth}}$相比,细节特征更加明显,并且物品轮廓周边的噪音有着明显的减少,SNR也更大. 跨模态融合特征$F^{\text{out}}_{\text{Fusion}}$既增强了特征表示又保留了物品的细节特征.

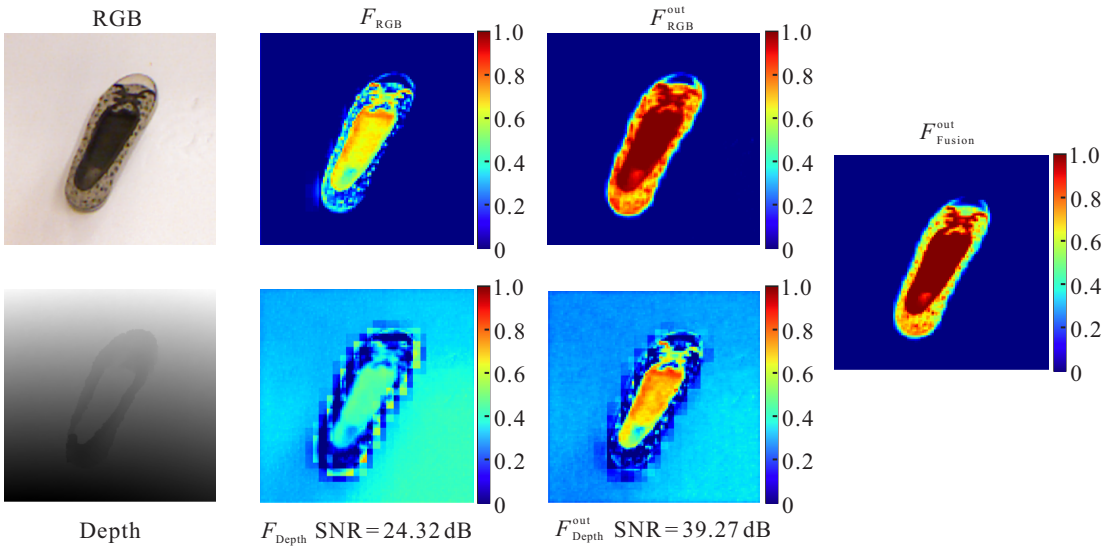


**图 5　在Cornell验证集的CMF输入输出特征图可视化**

### 2.2.4　跳跃连接特征融合模块(JFF)

远距离的跳跃连接具有很大的不确定性,如果连接的特征图之间存在较大的差异,可能会引入不必要的信息,加大模型计算负担. 为了加强跳跃连接的有效结合,采用SE-Net[30]的通道注意力的思想设计一个跳跃连接特征融合模块(jump feature fusion, JFF),如图1中JFF Block所示. 整个融合模块由3个SE Block组成. 使用这种通道注意力机制,模型可以根据

给定的输入关注需要的模态特征以及抑制不需要的特征,从而在跳跃连接中保修留有用的信息,提高模型的性能.

#### 2.2.5 抓取位姿预测模块(Detection)

Detection模块利用网络生成的特征像素图来预测抓取位姿. 首先通过4个$1\times1$卷积将Decoder输出的特征图转换为4个与检测图像分辨率一致的像素图,分别对应于分数图($S$)、角度正弦分量图(Sin)、角度余弦分量图(Cos)和宽度图($W$). 最后的抓取角度图$A$根据正弦和余弦分量图共同确定,抓取矩形的角度通过以下公式计算:

$$\theta=\frac{1}{2}\times\arctan\left[\frac{\sin2\theta}{\cos2\theta}\right]. \tag{19}$$

### 2.3 损失函数

网络损失函数的表达式如下所示:

$$L=\frac{1}{n}\sum_{i=1}^{n}(\lambda_1(S_i-S_i^G)^2+\lambda_2(A_i-A_i^G)^2+\lambda_3(W_i-W_i^G)^2). \tag{20}$$

其中:$n$是检测图像中像素点的数目,$S_i$、$A_i$、$W_i$是3个检测图中对应第$i$个像素点的值,$S^G$、$A^G$、$W^G$是与数据集中抓取矩形注释标签相对应的3个像素图标签,$\lambda_1$、$\lambda_2$、$\lambda_3$是各部分损失的权重系数. 考虑到矩形中心点和角度的预测直接影响到检测的准确率,$\lambda_1$、$\lambda_2$的值设置比$\lambda_3$更大,分别为1.5、1.5、1.

## 3 实验与分析

为了全面评估CMF-Grasp抓取检测方法,设计5个实验. 前3个实验是在单目标数据集Cornell[6]和Jacquard[14]以及多目标数据集[31]上的比较实验,旨在验证CMF-Grasp方法的有效性;第4个实验是在真实环境下的抓取检测实验,进一步测试CMF-Grasp的鲁棒性和可行性;第5个实验是消融研究,分析各模块对整体网络性能的贡献.

### 3.1 评估方法

使用与文献[7-8,11-12,15-16]相同的方法评估CMF-Grasp在给定图像上的抓取检测结果. 如果网络预测的抓取矩形$G^{\rm pre}$和数据集给定的抓取矩形$G^T$同时满足以下两个条件,则该预测被认为是正确的:

$$\begin{cases}|G_\theta^T-G_\theta^{\rm pre}|<30^\circ,\\ \text{Jaccard}=\dfrac{|G^T\bigcap G^{\rm pre}|}{|G^T\bigcup G^{\rm pre}|}>25\,\%.\end{cases} \tag{21}$$

第1个条件要求预测的抓取矩形框与注释的矩形框之间的角度差小于30°,第2个条件要求预测的矩形与注释的矩形之间的交并比(IOU)大于25%. 按图像划分方法IW (image-wise split)和按对像划分方法OW (object-wise split)两种数据集划分方法也被用于评估CMF-Grasp抓取检测的性能. IW用于测试模型在对象具有不同姿势时的抓取检测能力,OW用于测试模型的泛化能力.

### 3.2 实验细节

采用分辨率为224×224的三通道RGB图像和单通道的深度图像作为网络输入. 模型中的SwinT模块,窗口大小为7,从Encoder到Decoder六个DTC中的Transformer模块的多头数为(2, 4, 8, 8, 4, 2). 模型的训练是在Intel (R) Core (TM) i9-10900X CPU和NVIDIA A100-SXM4-80GB GPU的Windows 11系统上完成的. 初始学习率设置为0.001,每10个训练轮次学习率变为原始的0.1倍. 优化器是AdamW.

### 3.3 实验结果和分析

#### 3.3.1 Cornell数据集

Cornell抓取数据集提供了885组成对的RGB和深度图像数据. 数据集中的图像是真实环境下拍摄,并注释了多个抓取矩形. 在实验中,对图像进行旋转、缩放和随机裁剪等数据增强. 数据集中的80%用作训练集,另外20%用作测试集. 在训练集中,90%的数据用于训练,另外10%用于验证. 实验遵循先前工作[7,12,15-16]的设置,采用5折交叉验证.

**表1 在Cornell数据集的实验结果对比**

| 算法 | 输入 | 准确率/% | | 速度/(帧/秒) |
|---|---|---|---|---|
| | | IW | OW | |
| 文献[8] | Depth | 73.0 | 69.0 | **52.60** |
| 文献[6] | RGB-D | 73.9 | 75.6 | 0.07 |
| 文献[7] | RGB-D | 88.0 | 87.1 | 3.31 |
| 文献[32] | RGB | 88.7 | — | 5.00 |
| 文献[17] | RGB-D | 89.2 | 88.9 | 16.03 |
| 文献[33] | RGB-D | 90.2 | 90.6 | 41.70 |
| 文献[34] | RGB-D | 97.7 | 96.6 | 50.00 |
| 文献[11] | RGB-D | 97.9 | 96.7 | 24.40 |
| 文献[12] | RGB | 98.1 | — | 47.60 |
| 文献[15] | RGB-D | 98.2 | 97.1 | 40.00 |
| 本文算法 | RGB-D | **99.1** | **98.5** | 28.60 |

表1展示了CMF-Grasp方法和部分现有算法在Cornell数据集上的检测性能对比. CMF-Grasp在IW和OW分割的准确率分别为99.1%和98.5%,优于其他算法. 检测速度为28.6帧/秒,满足实时检测的要求. 图6展示了CMF-Grasp与GR-CNN[34]、TF-

Grasp[11] 在各抓取数据集上的检测效果对比. 其中: GR-CNN 是基于卷积残差网络研究的抓取检测方法, TF-Grasp 是基于纯 SwinT 架构的抓取检测算法. 图 6 中各数据集第 1 行到第 3 行分别是输入的深度图 (Depth)、抓取预测结果的 RGB 图 ($RGB_{pre}$) 和检测图得分图 $S$. 从 Cornell 检测结果图可以看出, CMF-Grasp 可以准确预测出物品的抓取矩形. 与 GR-CNN 和 TF-Grasp 相比, 采用 Transformer 和 CNN 混合架构的 CMF-Grasp 能更好地区分物体和背景, 物体的可抓取区域预测也更加准确和细致.

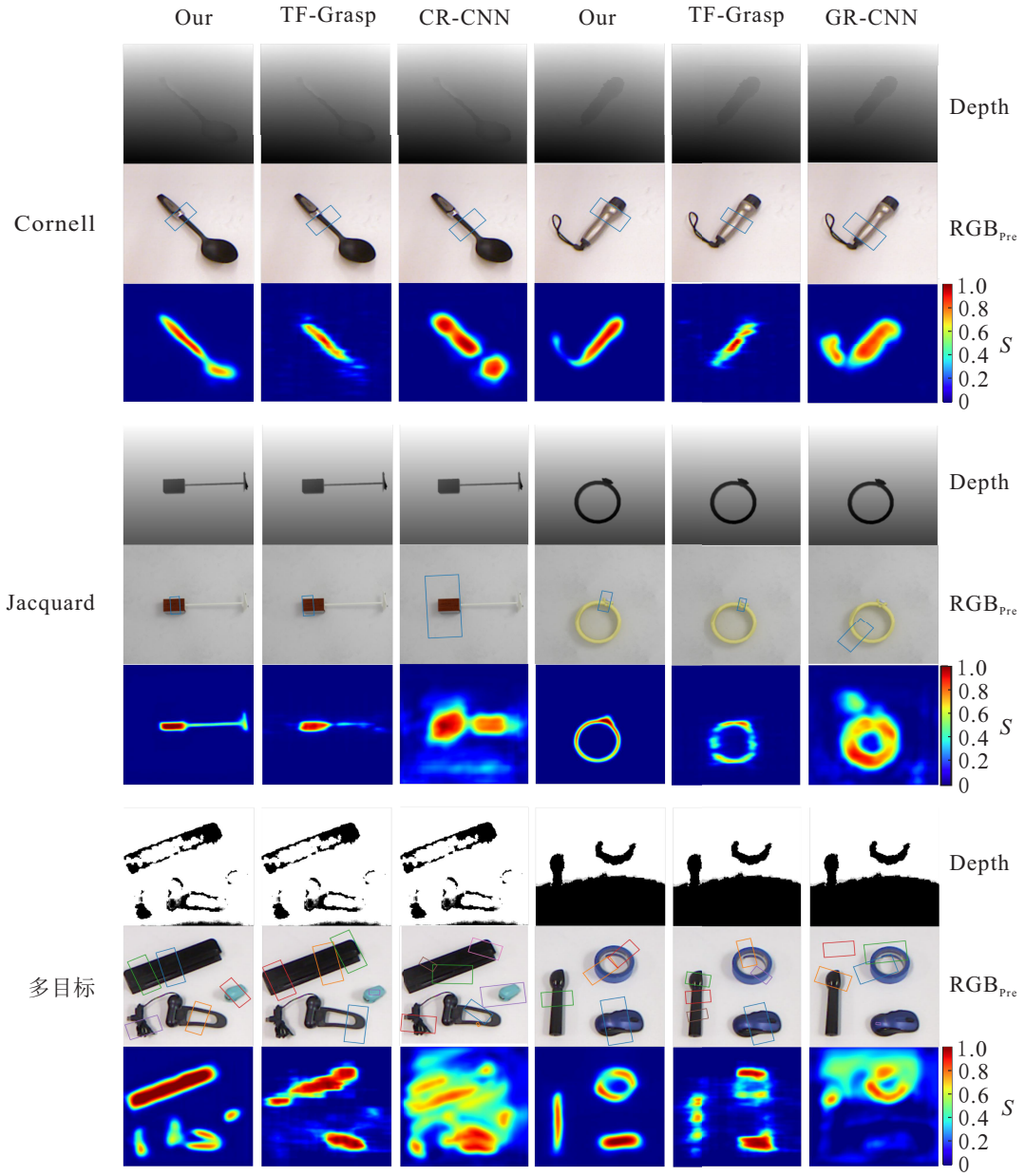


图 6　各抓取数据集的对比检测结果

### 3.3.2　Jacquard 数据集

Jacquard 抓取数据集由 54 k 个 RGB-D 图像组和在模拟环境中进行抓握测试的成功抓握矩形注释组成. 表 2 展示了在 Jacquard 数据集上的抓取检测结果, CMF-Grasp 能达到 96.2 % 的最高检测精度. 从图 6 与其他网络的对比结果也可以看出, CMF-Grasp 的抓取得分图 $S$ 中的物品可抓取区域最为清晰, 且保留了更多细节, 这有利于网络抓取检测准确率的提升.

表 2　在 Jacquard 数据集的实验结果对比

| 算法 | 输入 | 算法名称 | 准确率 /% |
|---|---|---|---|
| 文献 [14] | RGB-D | Jacquard | 74.20 |
| 文献 [8] | Depth | GG-CNN | 84.00 |
| 文献 [35] | RGB | ROI-GD | 90.40 |
| 文献 [36] | RGB | Det Seg | 92.59 |
| 文献 [34] | RGB-D | GR-CNN | 94.60 |
| 文献 [11] | RGB-D | TF-Grasp | 94.60 |
| 文献 [15] | RGB-D | SE-ResUNet | 95.70 |
| 本文算法 | RGB-D | CMF-Grasp | **96.20** |

### 3.3.3 多目标数据集

考虑到机器人在抓取检测过程中通常有多个物品出现, 也在多目标数据集[31] 上进行实验. 多目标环境下, 抓取矩形由图中局部像素值最大的点决定, 单次检测矩形最大数量设置为8. 从图6中$RGB_{pre}$图可以发现, 相比于其他两个方法的检测结果, CMF-Grasp的检测矩形更加适合物品的抓取, 矩形的宽度和角度更加准确. TF-Grasp的检测图尽管可以将物品与背景分离, 但忽略了待抓取物品的细节, 使得可抓取区域不明显, 这不利于不规则物品的抓取检测; 而GR-CNN不能准确地区分物体所在区域, 有部分背景也被当作可抓取区域. 相比之下, CMF-Grasp能够划分出各个抓取对象, 并且物品可抓取区域更加细致准确. 在CMF-Grasp的$RGB_{pre}$图中, 物体的可抓取区域以接近1的高分被强调, 而物体的边缘以小分被标记, 背景区域几乎全为0. 这说明CMF-Grasp在准确区分背景和物品的同时还能有效检测高可能性的可抓握区域.

### 3.3.4 真实场景下的抓取检测

本实验旨在验证CMF-Grasp在真实场景下的有效性和可行性. RGB和深度图像由Intel RealSense D435i相机在固定位置垂直水平面向下拍摄获得. 为了评估抓取检测算法的准确性, 对检测的RGB图像注释了适合抓取的矩形抓取标签, 图7展示了实验场景、实验所用物品和部分实验用图(深度图像、RGB图像、含矩形标签注释的RGB图像$RGB_{GT}$).


图 7 实验场景、实验物品和实验用图

对于单目标检测, 若预测的抓取矩形与其中一个注释的标签满足3.1节中的评估条件, 则认为该预测是正确的. 单目标检测实验中每次物品随机放置4次进行检测, 总共进行200次预测, 预测成功的次数为192次, 成功率为96%. 对于多目标检测, 预测的抓取矩形的最大数量设置为: 物品数+3. 当预测的矩形与每个物品的注释矩形之一都满足3.1节的评估条件时, 则认为该次预测是正确的. 多目标检测总共进行了100次, 预测成功次数为90次, 成功率为90%. 实验结果表明, 在真实场景下CMF-Grasp的抓取预测仍能够有效地提供适合物品抓取的抓取矩形. 部分检测结果如图8所示, 前4列是成功的预测, 后两列是失败的预测.

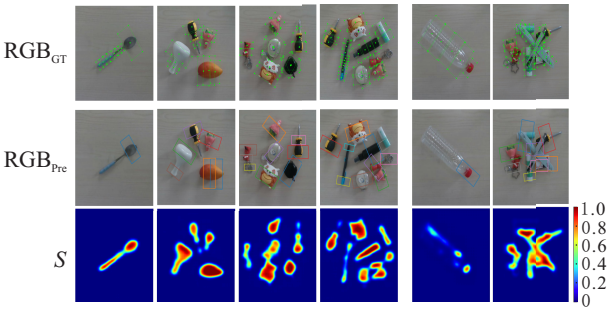


图 8 真实场景下检测效果

检测失败分析: CMF-Grasp在大部分情况下都能准确预测出抓取矩形, 但仍存在一些不足之处. 对于一些透明或半透明物体, 由于相机不能收集到有效的物体信息, CMF-Grasp不能很好地检测出可抓取区域. 在检测物品较多时, 一些小物品的可抓取区域不明显, 容易导致预测失败. 若有多个物体堆叠或者紧靠在一起, CMF-Grasp无法区分每个物体, 最终导致抓取预测失败. 这些问题可以通过采用更加精密的相机以及尝试对各个物品进行分割来解决, 这将留给后续的研究工作.

## 3.4 消融实验

### 3.4.1 与基于CNN和Transformer模型比较

为了验证DTC模块对整个抓取检测模型的贡献, 分别去除DTC中的SwinT部分和CNN部分在单目标Cornell和Jacquard数据集上进行实验. 实验结果如表3所示, 结果表明DTC的混合结构网络要比纯的CNN和纯的SwinT结构要好.

表 3 DTC消融实验结果

| | 准确率/% | |
|---|---|---|
| | Cornell | Jacquard |
| SwinT | 96.9 | 94.1 |
| CNN | 98.2 | 95.3 |
| CNN + SwinT | 99.1 | 96.2 |

### 3.4.2 跨模态混合交互融合模块(CMF)和跳跃特征融合模块(JFF)

为了验证CMF和JFF模块的有效性,分别去除这两个模块构建Net-a、Net-b和Net-c三个新的网络在单目标抓取数据集上进行实验. Net-a是保留双流编码器去除CMF模块的架构, Net-b是将RGB和深度图像级联输入单个编码器并且去除CMF模块的架构, Net-c是在解码器中去除JFF模块以逐元素相加进行跳跃连接的网络架构. 实验结果如表4所示,结果表明, CMF-Grasp这种双流编码器的特征提取网络架构比多模态级联的单编码器架构要好,同时CMF和JFF模块能有效地增强抓取网络的性能. 图5中的增强特征图也展示了CMF模块的有效性.

**表 4　DTC消融实验结果**

| | 准确率/% | | | |
|---|---|---|---|---|
| | TCFNet | Net-a | Net-b | Net-c |
| Cornell | 99.1 | 96.1 | 94.8 | 98.3 |
| Jacquard | 96.2 | 94.7 | 93.3 | 95.9 |

## 4 结　论

本文提出一种基于Transformer与CNN混合架构的RGB-D跨模态交互融合的机械臂抓取检测方法. 结合了CNN的局部建模能力和Transformer的全局建模能力,同时还利用不同模态的RGB和深度图像的特征相关性实现了跨模态的特征校准和交互融合,有效地增强了模型的特征提取能力. 在单目标和多目标数据集上的实验表明,该方法相比于其他方法具有更高的抓取准确率,同时预测的抓取物品可抓取区域更加细致明显. 真实场景下的抓取实验验证了该方法的有效性和鲁棒性. 但该方法对于透明物体以及堆叠物品的抓取预测略显不足,未来的工作将针对这类物品的抓取检测进行研究,同时尝试开发一种更加通用的抓取表示方法,使其能够适用于不同的机械臂抓取器,如三爪机械手和五指灵巧手.

## 参考文献(References)


[1] 陈友东, 刘嘉蕾, 胡澜晓. 一种基于高斯过程混合模型的机械臂抓取方法[J]. 机器人, 2019, 41(3): 343-352.
(Chen Y D, Liu J L, Hu L X. A manipulator grasping method based on mixture of Gaussian processes model[J]. Robot, 2019, 41(3): 343-352.)

[2] Kleeberger K, Bormann R, Kraus W, et al. A survey on learning-based robotic grasping[J]. Current Robotics Reports, 2020, 1(4): 239-249.

[3] 谢宇珅, 吴青聪, 陈柏, 等. 基于单目视觉的移动机械臂抓取作业方法研究[J]. 机电工程, 2019, 36(1): 71-76.
(Xie Y S, Wu Q C, Chen B, et al. Grasping operation method of mobile manipulator based on monocular vision[J]. Journal of Mechanical & Electrical Engineering, 2019, 36(1): 71-76.)

[4] 李明, 鹿朋, 朱龙, 等. 基于RGB-D融合的密集遮挡抓取检测[J]. 控制与决策, 2023, 38(10): 2867-2874.
(Li M, Lu P, Zhu L, et al. Densely occluded grasping objects detection based on RGB-D fusion[J]. Control and Decision, 2023, 38(10): 2867-2874.)

[5] 楚红雨, 冷齐齐, 张晓强, 等. 融入注意力机制的多模特征机械臂抓取位姿检测[J]. 控制与决策, 2024, 39(3): 777-785.
(Chu H Y, Leng Q Q, Zhang X Q, et al. Multi-modal feature robotic arm grasping pose detection with attention mechanism[J]. Control and Decision, 2024, 39(3): 777-785.)

[6] 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.

[7] Redmon J, Angelova A. Real-time grasp detection using convolutional neural networks[C]. 2015 IEEE International Conference on Robotics and Automation. Seattle, 2015: 1316-1322.

[8] 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.

[9] Loghmani M R, Planamente M, Caputo B, et al. Recurrent convolutional fusion for RGB-D object recognition[J]. IEEE Robotics and Automation Letters, 2019, 4(3): 2878-2885.

[10] Peng Z L, Huang W, Gu S Z, et al. Conformer: local features coupling global representations for visual recognition[C]. 2021 IEEE/CVF International Conference on Computer Vision. Montreal, 2021: 357-366.

[11] Wang S C, Zhou Z L, Kan Z. When Transformer meets robotic grasping: Exploits context for efficient grasp detection[J]. IEEE Robotics and Automation Letters, 2022, 7(3): 8170-8177.

[12] Dong M S, Bai Y X, Wei S M, et al. Robotic grasp detection based on Transformer[C]. International Conference on Intelligent Robotics and Applications. Cham: Springer, 2022: 437-448.

[13] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, 2017: 6000-6010.

[14] Depierre A, Dellandréa E, Chen L M. Jacquard: A large scale dataset for robotic grasp detection[C]. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid, 2018: 3511-3516.

[15] Yu S, Zhai D H, Xia Y Q, et al. SE-ResUNet: A novel robotic grasp detection method[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 5238-5245.

[16] Ren G L, Geng W J, Guan P Y, 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.

[17] Kumra S, Kanan C. Robotic grasp detection using deep convolutional neural networks[C]. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vancouver, 2017: 769-776.

[18] Niu J X, Liu S, Li H B, et al. Grasp detection combining self-attention with CNN in complex scenes[J]. Applied Sciences, 2023, 13(17): 9655.

[19] Zhang Q, Zhu J W, Sun X Y, et al. HTC-grasp: A hybrid Transformer-CNN architecture for robotic grasp detection[J]. Electronics, 2023, 12(6): 1505.

[20] 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. Las Vegas, 2016: 770-778.

[21] 李维刚, 甘平, 谢璐, 等. 基于样本对元学习的小样本图像分类方法[J]. 电子学报, 2022, 50(2): 295-304.
(Li W G, Gan P, Xie L, et al. A few-shot image classification method by pairwise-based meta learning[J]. Acta Electronica Sinica, 2022, 50(2): 295-304.)

[22] 李科岑, 王晓强, 林浩, 等. 深度学习中的单阶段小目标检测方法综述[J]. 计算机科学与探索, 2022, 16(1): 41-58.
(Li K C, Wang X Q, Lin H, et al. Survey of one-stage small object detection methods in deep learning[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(1): 41-58.)

[23] 张欢, 仇大伟, 冯毅博, 等. U-Net模型改进及其在医学图像分割上的研究综述[J]. 激光与光电子学进展, 2022, 59(2): 55-71.
(Zhang H, Qiu D W, Feng Y B, et al. Improved U-Net models and its applications in medical image segmentation: A review[J]. Laser & Optoelectronics Progress, 2022, 59(2): 55-71.)

[24] Maas A L, Hannun A Y, Ng A Y. Rectifier nonlinearities improve neural network acoustic models[C]. International Conference on Machine Learning. Atlanta, 2013: 3.

[25] Singh S, Krishnan S. Filter response normalization layer: Eliminating batch dependence in the training of deep neural networks[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 11234-11243.

[26] Shi W Z, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 1874-1883.

[27] Liu Z, Lin Y T, Cao Y, et al. Swin Transformer: Hierarchical vision Transformer using shifted windows[C]. 2021 IEEE/CVF International Conference on Computer Vision. Montreal, 2021: 9992-10002.

[28] Wang Q L, Wu B G, Zhu P F, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 11531-11539.

[29] 程德强, 许超, 李岩, 等. 基于时域加权处理的MPEG视频质量评价方法[J]. 激光与光电子学进展, 2018, 55(8): 268-273.
(Cheng D C, Xu C, Li Y, et al. MPEG video evaluation method with time-domain weighting[J]. Laser & Optoelectronics Progress, 2018, 55(8): 268-273.)

[30] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 7132-7141.

[31] 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.

[32] Karaoguz H, Jensfelt P. Object detection approach for robot grasp detection[C]. 2019 International Conference on Robotics and Automation. Montreal, 2019: 4953-4959.

[33] Asif U, Tang J B, Harrer S. GraspNet: An efficient convolutional neural network for real-time grasp detection for low-powered devices[C]. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm, 2018: 4875-4882.

[34] 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. Las Vegas, 2020: 9626-9633.

[35] Zhang H B, Lan X G, Bai S T, et al. ROI-based robotic grasp detection for object overlapping scenes[C]. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. Macau, 2019: 4768-4775.

[36] Ainetter S, Fraundorfer F. End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from RGB[C]. 2021 IEEE International Conference on Robotics and Automation. Xi'an, 2021: 13452-13458.

## 作者简介

王勇 (1974–), 男, 副教授, 博士, 硕士生导师, 从事计算机视觉、自然语言处理等研究, E-mail: ywang@cqut.edu.cn;

李邑灵 (1997–), 男, 硕士生, 从事机器人抓取位姿检测的研究, E-mail: yilingli010@163.com;

苗夺谦 (1964–), 男, 教授, 博士, 从事人工智能、机器学习、大数据分析等研究, E-mail: dqmiao@tongji.edu.cn;

安春艳 (1998–), 女, 硕士生, 从事机器人抓取位姿检测的研究, E-mail: ancy2623@163.com;

袁鑫林 (1996–), 男, 硕士生, 从事机器人抓取位姿检测的研究, E-mail: xinlinsmileyuan@163.com.