> 请将此行替换为您的稿件编号（双击此处进行编辑） <

SMR-Net：基于多尺度特征与自注意力网络的机器人卡扣检测

Kuanxu Hou

摘要—**在机器人自动化装配中，卡扣装配的精度和效率直接决定整体生产质量。作为核心前提，卡扣检测与定位对后续装配的成功至关重要。传统视觉方法在处理复杂场景（如透明或低对比度卡扣）时，鲁棒性较差且定位误差较大，无法满足高精度装配需求。为此，本文设计了一种专用传感器，并提出SMR-Net，这是一种基于自注意力的多尺度目标检测算法，可协同提升检测与定位性能。SMR-Net采用一种经注意力增强的多尺度特征融合架构：通过嵌入注意力机制的特征提取器对原始传感器数据进行编码，以强化关键卡扣特征并抑制噪声；采用标准卷积和空洞卷积并行处理三种多尺度特征图，在保持分辨率的同时统一维度；通过自适应重加权网络动态分配融合特征的权重，生成融合细节与全局语义的精细表征。在A型和B型卡扣数据集上的实验结果表明，SMR-Net显著优于传统的Faster R-CNN：交并比分别提高了6.52%和5.8%，平均精度均值（mAP）分别提高了2.8%和1.5%。这充分证明了该方法在复杂卡扣检测与定位任务中的优越性。**

索引词——**卡扣装配，卡扣检测与定位，目标检测，多尺度特征融合，自注意力**

一、引言

尽管机器人学在制造业中已得到广泛应用，但产品装配过程仍给机器人系统带来了重大挑战，目前这一环节仍以人工操作为主。本文聚焦于机器人装配中的关键技术问题，尤其关注带有卡扣接头的塑料零件的装配过程（以下简称“卡扣装配”）。

在卡扣装配操作中，对按扣进行准确的识别和定位是确保装配成功的核心前提。直接采用未经优化的控制策略往往会导致机器人操作过于激进，可能造成部件灾难性滑移或结构损坏，进而导致装配失败，甚至使系统停机。该问题在涉及位姿不确定性并依赖基于接触的搜索操作的场景中尤为突出 [1][2]。

目前，基于标准相机的视觉定位方法凭借能够提供丰富外观特征信息的优势，在常规场景下的卡扣识别和定位任务中展现出了良好的性能。然而，当卡扣部件由透明材料制成，或其表面颜色与背景高度相似时，传统视觉方法受到成像原理的限制，难以有效应对这类复杂工况。本文设计的专用传感器仅依赖物体的表面纹理深度，其性能不受材料透明度或背景颜色相似性的影响。因此，该传感器能够有效解决传统视觉相机在卡扣装配定位中面临的技术瓶颈。基于所设计的硬件系统，本文进一步提出了一种融合多尺度特征和自注意力机制的目标检测网络，旨在提升卡扣装配场景中识别与定位的准确性和鲁棒性。综上所述，本文的主要贡献如下：

- 我们设计了一种专用于卡扣装配场景的新型传感器，可在复杂环境中实现对卡扣的高精度识别与定位；
- 我们提出了一种基于注意力的多尺度网络，并将空洞卷积集成到其架构中，以便在将采集到的多尺度特征输入区域提议网络（RPN）模块之前进行处理；
- 我们开发了一种自适应权重分配网络，能够自主学习不同尺度特征的重要性并动态分配权重系数，从而有效提升融合的多尺度特征的表征质量，并达到最先进（SOTA）性能；
- 我们从卡扣识别、定位和装配三个方面，将所提出的算法与主流算法进行了全面比较，以评估该方法的先进性能；
- 我们针对每个网络组件进行了消融实验，以充分展示所提出算法工作流程的紧凑性。

II. 相关工作

A. 触觉传感器

触觉传感器的设计与开发已形成基于多种原理并行发展的格局，主流技术路线包括压电式、电阻式、电磁式和光学式 [3-6]。这类传感器的核心研究目标是在特定区域内准确测量接触力和接触位置。然而，传统的非光学传感器通常面临制造工艺复杂、布线繁琐等工程应用瓶颈。相比之下，基于光学原理的触觉传感器凭借布线简化、制造工艺便捷和空间分辨率高等显著优势，得到了广泛研究。Schneiter 和 Sheridan [7]以及 Begej[8]利用光纤捕捉硅胶表面的光反射特性，并通过摄像头检测光纤信号的变化模式。Tar 和 Cserey 等研究人员 [9]设计了一种传感器，采用中空半球形橡胶罩作为接触介质，并在罩顶内侧集成反射面。传感器底部设有三个接收器，通过测量变形穹顶反射的光信号来估算三轴接触力。然而，这类传感器缺乏空间测量能力，这一技术局限严重制约了按扣定位任务的准确性和可靠性。

A. 目标检测算法

按扣的外部形态差异显著。有些按扣不仅尺寸微小，而且纹理结构复杂。这些特点使基于深度学习的目标检测算法难以充分提取有效特征，从而影响检测准确率和可靠性。作为当前目标检测领域的主流优化方法之一，多尺度特征融合通过整合网络不同层的特征信息，能够有效提升细微纹理的表征能力和定位精度。例如，FPN [10] 构建了自上而下的特征融合路径，显著增强了低层特征的语义表达能力；PANet [11] 在此基础上增加了自下而上的融合分支，实现特征信息的双向流动与深度交互；YOLOF [12]设计了高效的融合单元并扩大网络感受野，在保证计算效率的同时提升了小目标检测性能。

特征增强方法可以提升特征图中小目标和精细纹理的显著性。例如，空洞卷积 [13] 可以在扩大感受野的同时捕获上下文信息，从而加强对小目标的理解；各种注意力机制（SENet [14], DANet [15], CABM [16], GAM [17]）可以引导网络关注感兴趣区域，并减少视野中噪声的干扰。

然而，小目标检测在适应极端尺度不平衡或不规则目标方面仍存在局限，而且一些增强策略往往会引入冗余信息，导致训练难以收敛。因此，在确保效率的前提下实现对小目标的稳健检测，仍然是一个重要的实际研究方向。

III. 数学

A. 传感器硬件简介

1）凝胶基底原理：本文设计的传感器旨在对目标物体的形状、外观和表面纹理进行高精度测量。传感器与目标物体接触的接触表面由透明弹性体材料制成，其表面涂覆有高反射率银粉涂层。当目标物体压在弹性体表面上时，弹性膜会发生自适应变形，从而准确复现物体表面的三维地形特征。为捕捉这些变形信息，在弹性膜背面布置了一台高精度相机，用于记录弹性膜上呈现的物体表面轮廓。基于相机采集的图像数据，可通过三维重建算法重建目标物体表面的三维模型。

2）弹性基底设计：本研究中，传感器的弹性基底尺寸设计为 40mm×40mm，厚度为 4mm。为有效捕捉物体表面细微且复杂的纹理信息，同时适应不同深度的表面纹理，弹性基底必须具有出色的柔韧性。弹性基底与目标物体接触的表面涂有精细的银粉涂层。该涂层的漫反射特性使其能够对更广范围的表面法线产生均匀响应，从而为准确捕捉物体表面的整体形貌提供有利条件。此外，传感器的最小分辨率可达 5 微米。

B. 传感器设计与制造

传感器的主体框架采用3D打印尼龙材料制造。传感器主要由法兰、高分辨率工业相机、支撑框架和镀银弹性基底组成。其中，支撑框架用于固定安装相机和弹性基底。在弹性基底与支撑框架之间增设丙烯酸导向板，以确保结构装配精度。弹性基底符合Neo-Hookean固体模型，硬度接近肖氏00-45，剪切模量为 μ 0.145 MPa。它安装在支撑框架底部的中央区域。在照明良好的条件下，当弹性基底受压变形时，光反射

其表面镀银层的方向会发生差异性变化。这使物体边缘对应区域的亮度较低，而与物体接触区域的基底表面亮度较高。

相机镜头中心与弹性基底之间的距离设为 125 mm，弹性基底厚度为 4 mm。本研究采用 BFLY-U3-23S6C-C 相机，其分辨率为1920×1200 像素，并配备索尼 ICX249 1/1.2 英寸逐行扫描 CCD图像传感器，可高分辨率采集接触表面的几何形貌。相机通过 USB接口连接至计算机，以实现高效、快速的数据传输。采集图像数据后，将原始图像裁剪为仅保留一个边长为 1000×1000 像素的中心正方形区域。

B. 算法流程

Faster R-CNN [18] 的网络结构由特征提取模块、区域提议网络（RPN）以及目标分类和边界框回归模块组成。具体而言，特征提取模块负责提取输入图像的语义特征和细节特征，而 RPN模块则根据提取的特征生成目标候选区域。为提高候选区域的准确性，本研究设计了一种卷积特征融合网络，采用 5×5、 3×3和 1×1卷积核替代原始 RPN 中的 3×3 卷积层，从而融合特征提取模块输出的三种不同尺度的特征。同时，在适当位置嵌入一个自注意力机制模块

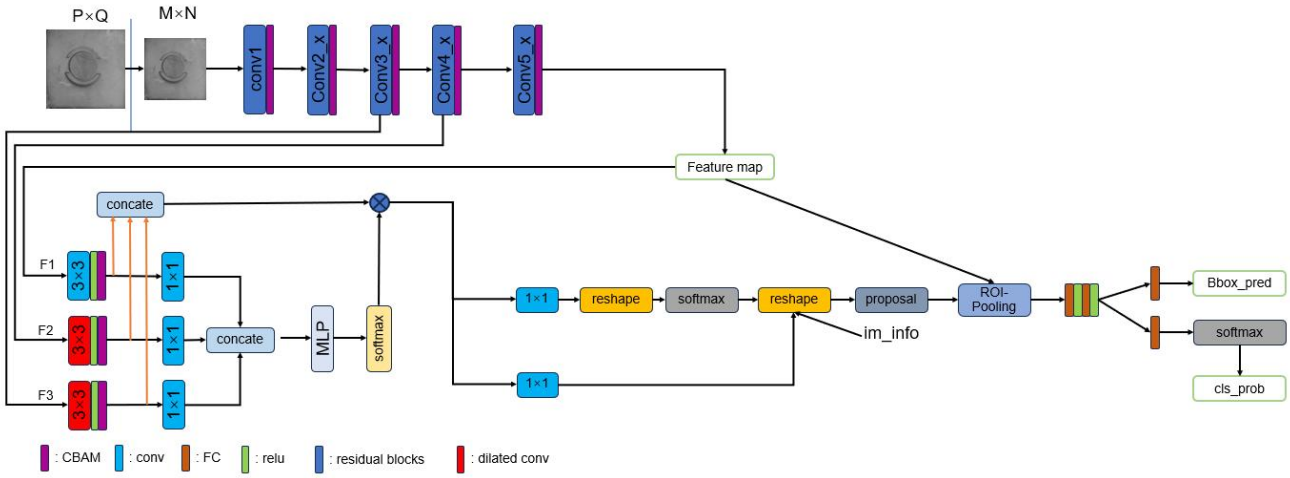

图 1. 所提出的 SMR-Net 流程概览。我们以 Faster R-CNN 为基础，集成坐标注意力块模块（CABM），以增强目标检测网络检测小目标和细微纹理的能力。具体而言，我们骨干网络中的 conv1、conv2_x、conv3_x、conv4_x 和 conv5_x 均由 ResNet34 [19] 的卷积层和残差块改造而来。

在卷积网络中的位置，以增强感兴趣区域（ROIs）的特征提取性能。总之，针对原始 Faster R-CNN，本研究分别对其特征提取模块和 RPN 模块进行了上述改进。

A. 自注意力模块

注意力机制可以有效增强卷积神经网络（CNN）的特征表征能力。因此，我们将注意力模块——坐标注意力块模块（CABM） [16]——集成到特征提取网络 ResNet34 的每个残差块之后。CABM 的具体实现过程如下：

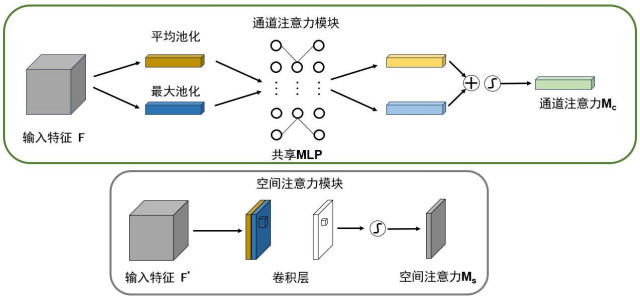

图 2. CABM [16] 算法流程示意图。

如图 1 所示，CBAM（卷积块注意力模块） [16]由两个子模块组成：CAM（通道注意力模块）和 SAM（空间注意力模块），分别执行通道维度和空间维度上的注意力计算。它可以无缝集成到现有网络框架中。通道注意力模块在压缩空间

维度的同时保留通道维度，聚焦于输入图像中的有意义信息。相比之下，空间注意力模块在压缩通道维度的同时保留空间维度，重点关注目标的位置信息。这两个模块（通道注意力和空间注意力）可以串联组合，从而更有效地提取图像信息。

B. 带自注意力的特征提取网络（*SAFE-Net*）

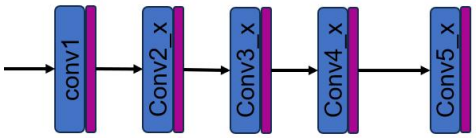

图3. SAFE-Net特征提取网络示意图。

特征提取模块是整个 Faster R-CNN 框架的基础。原始网络采用 VGG-16 [20] 作为特征提取骨干网络；然而，该模型的参数规模过大，限制了多尺度网络结构的进一步扩展。为解决这一局限，如图3所示，本研究采用 ResNet-34 替代 VGG-16 执行特征提取任务。同时，该网络为 RPN模块输出三个不同尺度的特征图，从而提高算法对不同尺寸目标的检测准确率。

具体而言，对 ResNet-34 网络进行结构分解，得到多个卷积层和残差块单元，包括 conv1、conv2_x、conv3_x、conv4_x 和 conv5_x。在每个卷积层和残差块之后均添加一个注意力层。通过自适应地关注关键特征并抑制冗余信息，有效提升了提取特征的质量。最终，该骨干网络输出的三个多尺度特征图为后续两阶段检测流程提供了更具判别力的特征支持。

C. 多尺度特征融合网络（*MSFF-Net*）

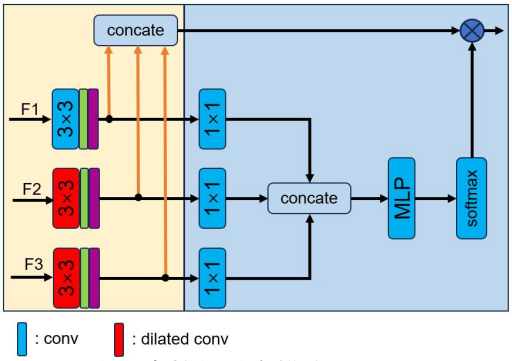

图3. 多尺度特征融合模块。

鉴于工业场景中的一些按扣具有尺寸较小、纹理细微等特点，传统目标检测算法难以充分提取其特征，导致检测准确率较低，难以满足实际应用需求。为解决这一问题，本研究提出了一种多尺度特征融合模块，其具体结构如图 4 所示。如图 4 左侧所示，特征提取网络输出的三个不同尺度特征图（F1、F2、F3）首先分别进行卷积操作：F1 后接一个标准 3×3 卷积，而 F2 和 F3 则各自后接 3×3 空洞卷积。利用空洞卷积在保留更多细节信息的同时扩大感受野这一特性，并结合多尺度特征融合策略，可以有效提高 RPN模块的定位回归精度。同时，每次卷积操作后都集成了一个注意力模块，通过强化目标的关键特征并抑制背景干扰，进一步提高目标检测的分类精度。

多尺度特征融合之所以具有独特优势，核心原因在于不同层级的特征图包含不同的信息：小目标的细节特征主要蕴含在低层特征图中，而高层特征图则富含语义信息。原始 Faster R-CNN 仅依赖高层特征图进行目标检测。输入图像经过多次下采样操作后，尽管语义特征较为丰富，但细节特征却严重丢失。这会导致 RPN 生成的候选区域置信度较低，最终造成小目标检测性能不理想。

C. 重加权网络（*RW-Net*）

不同尺度特征对检测结果的贡献存在显著差异。为更高效地利用多尺度信息，我们设计了一个自适应权重学习模块，用于自主学习各层特征的最优权重。如图3所示，该模块位于网络结构的右侧，首先通过 1×1卷积对多尺度特征进行通道维度压缩和信息整合。随后，将处理后的特征输入多层感知机（MLP），进行特征映射和权重计算。最后，通过 softmax函数归一化，得到各特征层的权重系数。将这些权重系数与对应的尺度特征图进行加权融合，即可得到最终的融合特征图，该特征图结合了各层特征的优势，进一步增强了网络对不同尺寸目标的特征表征能力。

IV. 实验

开展实验以评估所提出的软硬件系统在卡扣检测和安装过程中的有效性。

A. 与主流方法的对比实验

为评估所提方法的定位性能，实验采用交并比

（IoU）作为核心指标，具体操作流程如下：计算网络输出的最优候选边界框与人工标注的真实边界框之间的交并比，以量化模型的定位准确率。对于数据集，针对两种不同类型的按扣各采集 1000 张传感器图像，构建实验样本集；通过计算所有样本的交并比均值，得出最终的定位性能结果。

作为衡量两个区域重叠程度的经典指标，交并比（IoU）广泛应用于目标检测和语义分割等计算机视觉任务。其计算方法如下：通过计算预测候选边界框与真实边界框的重叠区域面积占两个区域并集面积的比例，可以直观地反映模型定位结果与实际情况的一致性。该比例越接近 1，定位性能越好。其计算公式如下：

$$IoU = \frac{P \cap Q}{P \cup Q} \qquad (1)$$

表 I 不同算法在 DIVERSESNAP 数据集上的交并比实验结果

方法	快照	交并比（%）
Yolo V8[21]	A	80.19
	B	82.66
Fast R-CNN[22]	A	80.58
	B	83.23
Faster R-CNN[18]	A	85.26
	B	86.32
Ours	A	**91.78**
	B	**92.12**

为验证所提算法的定位准确率，我们以交并比（IoU）为评估指标进行了定量分析实验。针对两种按扣（A型和B型），分别采集了1000张传感器图像，构建实验数据集。计算每张图像中按扣目标的预测边界框与真实边界框之间的交并比，并将所有样本的交并比结果求和取平均，得到各算法最终的定位性能指标。如表1所示，所提算法的平均交并比在对比方法中最高。这充分表明，所提算法具有更优的定位准确率，能够更精确地预测按扣目标的位置。

表 II 不同算法卡扣识别平均精度均值对比

方法	快照	平均精度均值 (%)
Yolo V8[21]	A	95.7
	B	95.1
Fast R-CNN[22]	A	95.5
	B	96.2
Faster R-CNN[18]	A	96.5
	B	97.9
Ours	A	99.3
	B	99.4

由于不同类型按扣的安装方式存在差异，因此需要准确识别按扣类型。针对按扣类型识别任务，我们分别采集了1000张A型和B型按扣图像，构建实验数据集。采用平均精度均值（mAP）作为评估指标：计算每张图像的检测精度，再求和取平均，作为各算法最终的类型识别性能结果。如实验数据表所示，所提算法的检测精度显著优于其他对比算法。具体而言，与Faster R-CNN相比，A型和B型按扣的识别准确率分别提高了2.8%和1.5%。这充分验证了所提改进策略能够有效增强区分按扣类型的能力。

为验证所提出的检测算法在实际按扣安装场景中的应用效果，本研究设计并开展了按扣安装实验。研究选取了两种不同类型的按扣，并对每种类型进行了50次安装测试。采用安装成功率作为评估算法性能的核心指标。如表1所示，实验结果表明，所提出的算法在按扣安装成功率方面取得了最佳表现。借助所提出的检测算法，A型和B型按扣的安装成功率均达到98%。这充分表明，该算法能够满足实际工业安装场景对准确性和稳定性的要求，并具有出色的工程应用价值。

表III 不同算法定位条件下的按扣安装成功率比较

方法	快照	成功率
Yolo V8[21]	A	88%
	B	88%
Fast R-CNN[22]	A	88%
	B	88%
Faster R-CNN[18]	A	90%
	B	90%
Ours	A	**98%**
	B	**98%**

B. 消融研究

为系统验证所提出算法的有效性，本研究针对论文设计的核心改进模块开展消融实验，即 SA-FE 特征提取网络、多尺度特征融合模块和重加权网络。实验采用收集到的 1000 张传感器图像作为数据集。通过构建不同的算法结构（分别移除或保留各个核心模块），进行独立训练和测试

。采用交并比（IoU）作为性能评估指标，对比分析各模块对检测准确率的贡献，从而验证所提出的各算法结构的有效性和必要性。

表 IV
比较 OF IoU 数值 不同 之间 算法结构

消融实验	快照	交并比 (%)
不含 SAFE-Net	A	85.22
	B	89.34
不含 MSFF-Net	A	86.32
	B	88.37
不含 RW-Net	A	90.11
	B	90.23
Ours	A	**91.78**
	B	**92.12**

表 V
比较 OF IoU 数值 不同 之间 算法结构

消融实验	快照	mAP
不含 SAFE-Net	A	97.2%
	B	97.1%
不使用 MSFF-Net	A	98.5%
	B	98.3%
不使用 RW-Net	A	98.9%
	B	99.1%
Ours	A	**99.3%**
	B	**99.4%**

1）特征提取网络（*SAFE-Net*）的影响：表 IV 和表 V 中的实验结果表明，将注意力模块引入网络后，模型的识别准确率和定位精度均显著提升。这一结果验证了结合空间与通道维度的注意力模块能够有效聚焦图像中的关键有效信息，并抑制冗余的背景干扰。通过自适应地增强和筛选特征，该模块为后续目标识别和定位任务提供了更具判别性的特征支持，从而提升整体检测性能。

2）多尺度特征融合模块（*MSFF-Net*）的影响：表 IV 和表 V 中的实验结果表明，无论是否采用带有注意力机制的特征提取网络，所提出的多尺度网络都表现出优于单尺度网络的检测性能。这是因为不同尺度的特征图关注图像中不同区域和层次的信息：低层特征包含丰富的细节信息，而高层特征则包含更丰富的语义信息。仅依赖单尺度特征（无论是中层还是高层）进行检测，由于信息表征不完整，往往会限制性能。相比之下，多尺度特征融合能够整合各层特征的优势，从而实现对目标信息的全面捕获，进而显著提升检测性能。此外，在网络架构的适当位置嵌入注意力层后，模型性能会进一步提升。这表明，注意力机制通过自适应地增强关键目标特征并抑制无关背景信息，与多尺度融合形成了有效的性能互补，从而进一步增强特征的判别能力。

3）重加权网络（*RW-Net*）的影响：本实验通过移除重加权网络并采用直接特征拼接方法，构建了一个对比模型。实验结果表明，在所有测试场景中，引入重加权网络的模型性能均优于直接特征拼接方案。这一结果验证了所提出的重加权网络能够自主学习为每个特征图分配最优权重，有效挖掘并充分利用不同尺度特征中的重要信息。该网络避免了简单拼接中各特征信息权重不均衡的问题，最终使融合的多尺度特征具备更稳健、更具判别性的表征能力，为提升检测性能提供了关键支持。

V. 结论

本文分析了当前机器人卡扣安装过程中检测与定位任务面临的挑战性场景。在此基础上，本文设计了一种新型接触传感器硬件，并提出了一种基于注意力的多尺度网络架构（命名为 SMR-Net），以提升挑战性场景下的卡扣检测与定位性能。SMR-Net 的核心设计包含两个方面：首先，构建集成自注意力机制的特征提取网络，分别提取三个不同尺度的特征；其次，结合空洞卷积与重加权网络实现特征融合，其有效性已通过消融实验得到验证。与现有目标检测算法相比，所提出的 SMR-Net 在定位性能和卡扣识别能力方面表现更优。未来的研究将着重简化并加速该算法，以便将其部署到边缘处理器上并实现实时运行。

参考文献

[1] S. R. Chhatpar 和 M. S. Branicky，“具有位置不确定性的销孔装配搜索策略”，载于*2001*年*IEEE/RSJ*智能机器人与系统国际会议论文集：拓展机器人在下一个千年的社会作用（目录号：01CH37180），第 1465–1470 页。[2] S. Jin、X. Zhu、C. Wang 和 M. Tomizuka，“不确定性条件下销孔装配的接触位姿识别”，载于美国控制会议（*ACC*），2021年5月，第 48–53 页。[3] R. S. Dahiya、G. Metta、M. Valle 和 G. Sandini，“触觉感知——从人类到类人机器人”，*IEEE*机器人学汇刊，第 26 卷，第 1 期，2010年2月，第 1–20 页。

[4] H. Yousef、M. Boukallel 和 K. Althoefer，“机器人灵巧手内操作的触觉感知——综述”，《传感器与执行器*A*：物理》，第 167 卷，第 2 期，2011年6月，第 171–187 页。[5] M. R. Cutkosky、W. Provancher，“力与触觉感知”，载于《*Springer*机器人学手册》，2016年，第 717–736 页。[6] Z. Kappassov、J.-A. Corrales 和 V. Perdereau，“灵巧机器人手中的触觉感知——综述”，《机器人学与自主系统》，第 74 卷，2015年12月，第 195–220 页。[7] J. L. Schneiter、T. B. Sheridan，“面向机械臂的光学触觉传感器”，《机器人学与计算机集成制造》，第 1 卷，第 1 期，1984年，第 65–71 页。[8] S. Begej，“用于机器人应用的平面式和指形光学触觉传感器”，《*IEEE*机器人学与自动化期刊》，第 4 卷，第 5 期，1988年，第 472–484 页。[9] A. Tar 和 G. Cserey，“低成本三维光学柔顺触觉力传感器的开发”，载于*2011*年*IEEE/ASME*国际先进智能机电一体化会议（*AIM*），2011年7月，第 236–240 页。[10]T. Y. Lin、P. Doll、R. Girshick 等，“用于目标检测的特征金字塔网络”，载于*IEEE*计算机视觉与模式识别会议论文集，2017年，第 2117–2125 页。[11]S. Liu、L. Qi、H. Qin、J. Shi 和 J. Jia，“用于实例分割的路径聚合网络”，载于*2018*年*IEEE/CVF*计算机视觉与模式识别会议，2018年6月，第 8759–8768 页。[12]Q. Chen、Y. Wang、T. Yang、X. Zhang、J. Cheng 和 J. Sun，“你只看单层特征”，载于*2021*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），2021年6月，第 13034–13043 页。[13]F. Yu、L. Shen、V. Koltun，“通过空洞卷积进行多尺度上下文聚合”，arXiv预印本*arXiv:1511.07122, 2015*。[14]J. Hu、L. Shen、G. Sun，“挤压激励网络”，载于*IEEE*计算机视觉与模式识别会议论文集，2018年，第 7132–7141 页。[15]J. Fu、J. Liu、H. Tian 等，“用于场景分割的双重注意力网络”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2019年，第 3146–3154 页。[16]S. Woo、J. Park、J. Y. Lee 等，“CBAM：卷积块注意力模块”，载于欧洲计算机视觉会议论文集（*ECCV*），2018年，第 3–19 页。[17]Y. Liu、Z. Shao、N. Hoffmann，“全局注意力机制：保留信息以增强通道—空间交互”，2021年，*arXiv:2112.05561*。[18]S. Ren、K. He、R. Girshick 和 J. Sun，“Faster R-CNN：利用区域提议网络实现实时目标检测”，《*IEEE*模式分析与机器智能汇刊》，第 39 卷，第 6 期，2017年6月，第 1137–1149 页。[19]K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，载于*2016*年*IEEE*计算机视觉与模式识别会议（*CVPR*），2016年6月，第 770–778 页。[20]K. Simonyan、A. Zisserman，“用于大规模图像识别的超深卷积网络”，2014年，*arXiv:1409.1556*。[21]M. Yaseen，“什么是YOLOv8：深入探索下一代目标检测器的内部特征”，2024年，*arXiv:2408.15857*。