基于权重偏置图注意网络的复杂场景中物体位置关系推理方法

左国玉 王子豪 赵 敏 于双悦

(北京工业大学信息科学技术学院 北京 100124)

(计算智能与智能系统北京市重点实验室 北京 100124)

摘 要 在复杂环境中安全抓取目标物体对于机器人技术至关重要,这要求机器人能够准确理解目标物体与周围其他物体之间的空间位置关系。尽管卷积神经网络在关系推理方面展现出一定的潜力,但由于其主要关注像素级信息提取,导致对全局信息的理解不足,并忽略了关键的物体关系,从而限制了推理的准确性。为了解决这一问题,本文提出了一种基于端到端图注意网络的关系推理模型,旨在提升推理物体位置关系的准确性。该模型首先采用 EfficientNet-B0 与双向特征金字塔网络(BiFPN)进行 RGB 特征提取。其次,在构建图结构时,通过过滤缺乏上下位置关系的物体对,使图结构更加稀疏,从而降低计算负担。随后,利用带权重偏置的图注意网络来预测物体之间的位置关系。在视觉操纵关系数据集(VMRD)上对所提模型进行了训练和评估。结果显示,该模型在关系推理的图像准确率(*IA*)指标上达到了 71.1%。此外,采用梯度加权类激活映射(Grad-CAM)进行了注意力可视化,进一步验证了模型在多物体无序堆叠场景中推断空间位置关系的有效性,使其适用于真实的机械臂抓取应用。最后,通过在实验室环境中对常见物体进行测试,成功地将模型应用于真实世界的机械臂抓取场景,证明了该模型在实际环境中的通用性和实用性。

关键词 复杂场景;关系推理;BiFPN;图注意网络;抓取顺序

中图法分类号 TP242 **DOI 号** 10.11897/SP.J.1016.2025.00572

Weight-Biased Graph Attention Network for Object Position Relationship Reasoning in Cluttered Scenes

ZUO Guo-Yu WANG Zi-Hao ZHAO Min YU Shuang-Yue

(School of Information Science and Technology, Beijing University of Technology, Beijing 100124)

(Beijing Key Laboratory of Computing Intelligence and Intelligent Systems, Beijing 100124)

Abstract To enable robots to safely grasp target objects in cluttered environments, a precise understanding of the spatial relationships between the target objects and their surrounding counterparts is essential. While convolutional neural networks (CNNs) demonstrate potential in relational reasoning, their primary focus on pixel-level feature extraction limits their ability to comprehend the global context and critical object relationships, subsequently affecting inference accuracy. To address these limitations, we propose a relationship reasoning model based on Graph Attention Networks aimed at enhancing the accuracy of spatial relationship understanding among objects. Initially, we employ EfficientNet-B0 combined with a Bidirectional Feature Pyramid Network (BiFPN) for RGB feature extraction during the detection process. To alleviate the computational burden, we filter out object pairs that lack clear contextual spatial relationships. We then utilize a sparsified Graph Attention Network that incorporates directional attention for

收稿日期:2024-06-12;在线发布日期:2024-12-19.本课题得到国家自然科学基金(62373016)、多模态人工智能系统国家重点实验室开放项目(MAIS-2023-22)资助。**左国玉**(通信作者),博士,教授,主要研究领域为机器人学习、机器人控制。E-mail:zuoguoyu@bjut.edu.cn。**王子豪**,硕士研究生,主要研究领域为机器学习、深度强化学习。**赵 敏**,博士研究生,主要研究方向为机器人抓取、计算机视觉、深度学习。**于双悦**(通信作者),博士,讲师,主要研究方向为机电一体化、可穿戴机器人、软体机器人。E-mail:yushuangyue@bjut.edu.cn。

effective relationship reasoning. The proposed model is trained and evaluated on the Visual Manipulation Relationship Dataset (VMRD), with attention visualized using Gradient-weighted Class Activation Mapping (Grad-CAM). The method proposed in this paper was evaluated on the VMRD dataset, achieving the highest precision across *mAP*, *OR*, and *IA* metrics. This reflects an improvement in both object detection and object relationship reasoning tasks. Specifically, the *mAP* metric reached 96.1%, indicating that the unique BiFPN structure in the EfficientDet network better integrates features of different scales within the image, effectively enhancing the average precision of image detection. The significant improvements in Object Recall (*OR*) and Image Accuracy (*IA*) demonstrate that our method can correctly infer a greater number of object relationship pairs during the reasoning phase. Comparative experiments against other methodologies on the same dataset reveal that our model significantly improves the accuracy of relationship reasoning, demonstrating its applicability and extensibility to real robotic arm grasping scenarios. The model achieves an image-based accuracy (*IA*) of 71.1% in relational reasoning tasks. To validate the effectiveness of the proposed model, we employed a technique called Gradient-weighted Class Activation Mapping (Grad-CAM), which is used to interpret the decision-making process of deep convolutional neural networks. Its primary aim is to visualize the attention distribution of the neural network on input images during classification tasks, aiding in the understanding of the model's predictive decision-making process. Grad-CAM visualizations further substantiate the model's capability to infer spatial relationships among multiple objects in cluttered scenes, underscoring its suitability for real-world robotic applications. Additionally, we established a visual grasping experimental platform based on the AUBO-i5 robotic arm, equipped with a two-finger electric gripper and a depth camera. To validate the practical application and generalization ability of the proposed model, we constructed a specific test set in a laboratory environment. The collected real grasping scene examples were used as a new test set to assess the model's generalization capability. The results indicate that the method described in this paper still exhibits good performance on the new dataset. While our RGB-based Graph Attention Network effectively predicts relationships among visible objects, it is validated for scenarios involving 2 to 5 objects. Future research will focus on integrating robotic operational actions and exploring methods to infer information about occluded objects based on the positional relationships of visible objects. We will also investigate strategies to enhance model performance in scenarios with more than five objects and conduct physical experiments in increasingly complex real-world operational environments to validate effectiveness and identify additional areas for improvement.

Keywords　cluttered scene; relational reasoning; BiFPN; graph attention network; grasping order

1　引　言

抓取操作是机器人执行任务时的基本技能之一，而有选择的抓取则使机器人更具智能化[1]。为了更好地与人类互动并实现语义抓取，智能机器人需要具备从多物体无序堆叠的场景中抓取指定目标物体的能力[2]。确保机器人能够安全且可靠地完成抓取任务，对于理解场景中目标物体与其他物体之间的相对位置关系至关重要[3]。

在复杂的多物体堆叠场景中，针对特定物体的抓取操作可能会对周围物体产生显著影响。图 1 展示了机器人在两种由相同物体构成的不同场景中的抓取任务，目标物体均为长方体。在场景(a)中，机器人可以直接抓取目标物体，而不会对相邻的球体或棱锥产生明显干扰。然而，在场景(b)中，若尝试直接抓取目标矩形物体，势必会影响球体和棱锥的位置，甚至可能导致它们被移出工作区域。因此，在场景(b)中安全可靠地完成抓取任务，需要首先将覆盖在矩形物体上方的其他物体移至安全位置。这

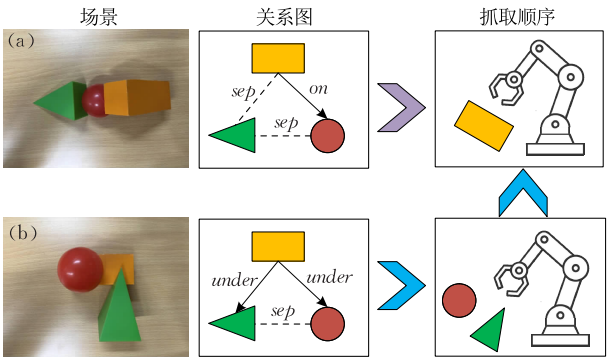

图 1 物体的抓取顺序受其位置关系的影响

一分析表明,准确理解物体间的相对位置关系对于制定机器人的操作策略至关重要。

这种通过分析目标物体在场景中的位置关系,进而推导出操作顺序的抓取方法被称为推理抓取。目前的推理抓取方法主要可分为两类:基于点云及RGB-D数据的方法,以及依赖于纯RGB图像的推理方法。然而,每种方法都有其自身的局限性。基于点云及RGB-D数据的方法主要利用深度信息来增强场景的三维表征,从而提高操作的准确性。然而,这种方法的性能在很大程度上受限于深度传感器的精度,并且在处理物体在多种姿态和位置下的点云配准时面临挑战。此外,处理庞大的点云数据集可能会影响系统的实时性能,尤其是在需要快速响应的动态环境中。另一方面,纯RGB图像的推理方法则专注于从二维图像中提取物体的局部信息,即在像素级别进行特征提取。尽管这类方法在处理图像数据时效率较高,但它们可能会忽略场景的整体上下文信息。这种局限性可能导致对环境中物体间复杂关系的全局理解不足,从而影响抓取策略的准确性和可靠性。

为了解决这一问题,本文提出了一种基于图注意力网络的关系推理模型,以提高物体在空间位置关系中的准确性。首先,在检测过程中采用EfficientNet-B0结合双向特征金字塔网络(Bidirectional Feature Pyramid Network, BiFPN)进行RGB特征提取;接着,通过筛选过程排除那些缺乏明确上下文位置关系的物体对,以减少计算负担;随后,我们使用经过稀疏化处理并包含方向注意的图注意网络来进行关系推理。最后,所提出的模型在视觉操纵关系数据集[4]上进行了训练和评估,并使用梯度加权类激活映射(Gradient-weighted Class Activation Mapping, Grad-CAM)[5]对注意力进行了可视化。这项工作的贡献概述如下:

(1)提出一种基于图注意的视觉关系操作网络,它能更准确地解释物体之间的位置关系和操作顺序。

(2)引入定向注意力机制,根据不同对象对同一对象的不同接近程度采取不同权重,并为具有不同侧面关系的特征设置不同的偏置。

(3)在公开的视觉操纵关系数据集(Visual Manipulation Relationship Dataset, VMRD)和实验室场景中的实验表明了该方法的有效性和适用性。

本文第2节介绍基于关系推理的方法在机器人抓取任务中的研究进展;第3节详述本文提出的基于图注意网络的关系推理方法;第4节详细介绍实验的过程,包括实验设置、评价指标、实验的结果和分析;最后,第5节对本文工作进行归纳总结。

2 相关工作

在机器人抓取任务中,基于关系推理的方法旨在综合考虑目标物体与周围物体之间的相互关系,以便在复杂环境中规划出合理的抓取顺序。这一规划过程至关重要,因为不当的操作顺序可能导致物体遭受不可逆转的损害。尽管该领域已有一些初步研究,仍然存在许多挑战。例如,Guo等人[6]在抓取检测过程中引入了物体的类别信息,但尚未形成一个有效的抓取操作序列。Fischinger等人[7]则通过结合物体的高度信息来增强对堆叠场景的理解能力。此外,其他研究[8-9]尝试利用物体的局部抓取置信度和高度信息来确定抓取顺序。尽管这些方法在抓取顺序的决策中有所推动,它们大多数忽略了对场景中物体之间操作关系的全局建模和深入理解,导致缺乏一套稳定且合理的操作序列。因此,针对复杂环境下的机器人抓取任务,仍需进一步研究以提升模型的全局关系推理能力,从而确保抓取行为的有效性和安全性。

为了解决堆叠场景抓取过程中的这些缺陷,研究人员逐步开发了视觉层面的操纵关系模型,以帮助机器人在堆叠场景中进行抓取操作。Lu等人[10]首次提出了一种利用区域卷积神经网络(Region-based Convolutional Neural Network, RCNN)进行物体检测并融合语言先验信息以识别物体间关系的方法,通过结合两个物体实体和一个关系谓词来预测图像中存在的多种关系。张翰博等人[11]定义了视觉操纵关系(Visual Manipulation Relationship,

VMR)，并构建了一个基于卷积神经网络(Convolutional Neural Network，CNN)的视觉操作网络架构，采用三个下采样 CNN 进行场景分类，该方法不仅能够识别目标物体，还能分析其与周围物体的相对位置关系，并通过操作关系树展示结果。Yang 等人[12]采用邻接概率矩阵构建操纵关系模型，并将条件随机场(Conditional Random Field，CRF)[13]引入关系推理过程中，以确立场景中所有操纵关系元素的确定性。Zuo 等人[14]将物体和操纵关系分别视作图的节点和边，应用图神经网络从全局视角预测操纵关系。Ding 等人[15]则引入门控循环单元(Gated Recurrent Unit，GRU)[16]作为信息传播的函数，以提升模型对物体间关系的认知能力。这些针对视觉操纵关系推理的方法大体上遵循两个阶段的流程：首先是生成对象提议的目标检测阶段，其次是对物体关系进行推理的关系推理阶段。尽管这些方法在目标检测结构方面具有一定的共性，但它们在关系推理策略的设计和实现上存在明显差异。

除了基于 RGB 图像的方法，还有一类研究基于 RGB-D 图像。例如，Xiong 等人[17]提出了一种创新的多视图映射分割策略，利用两个深度相机从不同视角(正视图和俯视图)捕获物体信息，结合正视图的深度图与三维重构图，并利用俯视图中识别的标签图进行映射分割，以区分场景中的独立物体。每个分离的物体被封装在一个三维有向边界框(Oriented Bounding Box，OBB)内。通过分析每个物体的几何特征，如法向量的垂直分量，进而推断物体间的空间关系并构建关系。然而，这类基于 RGB-D 的方法受到深度传感器性能的限制，并且由于涉及额外的维度信息处理，通常需要更多的计算资源，尤其是在实时性能要求较高的应用场景中。

最近，Tchuiev 等人[18]提出了一种创新的方法，通过结合 Adj-Net(Adjacency-Network)和 DUQIM-Net(Due-order Understanding for Quality Inference and Manipulation Network)来增强对场景中物体层次关系的评估以及物体操作决策的能力。Adj-Net 采用了基于 Transformer 的 Encoder-Decoder 架构作为物体检测框架，并引入了邻接头(Adjacency Head)技术，以概率方式评估场景中物体之间的层次堆叠关系，从而推导出加权邻接矩阵形式的对象层次结构。与此同时，DUQIM-Net 作为一种专门针对堆叠对象场景设计的决策模型，利用 Adj-Net 输出的邻接矩阵评估结果，结合现有的 Transformer 物体检测器，并通过新增的邻接头来推断场景中物体的底层层次结构。基于此信息，DUQIM-Net 能够做出决策，从而有效地协助完成物体抓取任务。在此基础上，Xu 等人[19]提出了一种基于 DETR(Detection Transformer)的模型，旨在解决物体检测中的集合预测问题，实现抓取关系的端到端检测。尽管该方法能够获得较高的预测准确率，但其模型参数较多且计算复杂，影响了实时分析的性能。此外，Wu 等人[20]提出了分层堆叠关系网络(Hierarchical Stacking Relationship Network，HSRN)，该网络旨在深入感知场景并生成堆叠关系树(Stacking Relationship Tree，SRT)，以描述场景中物体间的关系。此方法主要适用于厨房餐具的堆叠场景，例如碗、筷子和餐盘等物体的抓取。机械臂可以通过直接抓取底部物体，来同时托起上方的所有物品，从而实现一次性多物体的抓取操作。

近年来，图网络在场景图推理[21-23]和语义抓取[24]中得到了广泛应用。这些网络不仅能够表示个体信息，还能捕捉个体间的交互关系[25]。Pei 等人[22]和 Sun 等人[23]提出了用结构化图形来表示场景图的方法，将对象编码为节点，并将它们的关系编码为有向边。他们指出，在给定的三元组中，如果已知其中两个元素，就能够预测第三个元素。基于此，我们提出了一种新的基于图注意力(Graph Attention Networks，GAT)的推理方法，旨在提取对象之间的关系信息，并结合定向注意力机制来提高关系推理的准确性。

3　方　法

本文提出了一种以目标为导向的视觉关系推理方法，即基于图注意力网络(Graph Attention Network，GAT)。模型的整体结构如图 2 所示，主要由特征提取、目标检测和关系推理三个核心部分组成。首先，模型以视觉传感器采集到的 RGB 图像作为输入。通过特征提取模块，我们提取出图像的基础特征，并对这些特征进行多尺度融合，以增强对不同对象和场景细节的捕捉能力。接着，经过目标检测模块，我们获取到目标物体的边界框和分类信息。最后，将特征提取和目标检测模块的输出信息进行融合，作为关系推理模块的输入。这一模块通过分析目标物体之间的相对位置和关系，最终生成对目标物体与其他物体之间位置关系的预测信息。

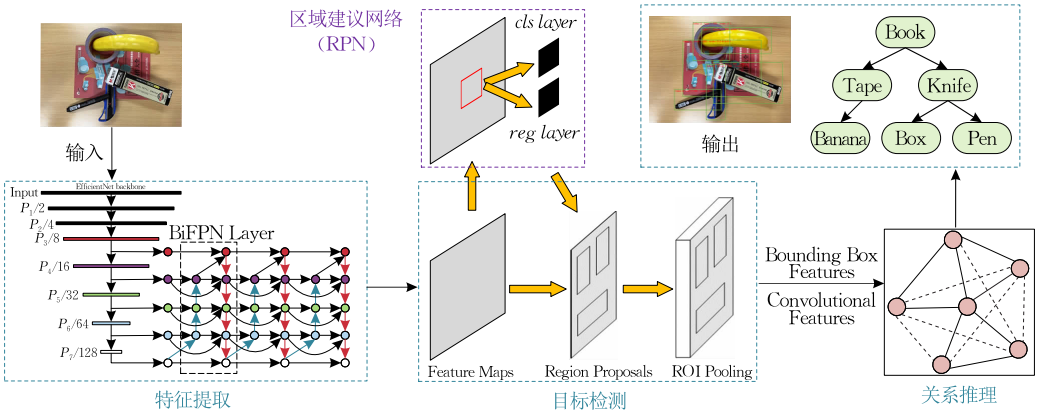

图 2 基于图注意力网络的视觉关系推理方法的整体结构

接下来,本节将对该方法进行详细阐述,包括其工作原理、各个模块的具体实现以及所采用的技术细节。

3.1 特征提取

特征提取是整个网络架构的首要环节,其主要任务是将视觉传感器获取的 RGB 图像信息转换为可供后续处理的特征表示。为了实现这一目标,输入特征需要通过一个骨干网络(backbone)进行提取。近年来,基于卷积的特征提取网络层出不穷,其中一些经典的模型包括 AlexNet、VGG 和 ResNet 等[26]。理想的骨干网络应具备在尽可能短的时间内提取丰富的图像特征的能力,以提高整个系统的效率和性能。为了解决速度与精度之间的权衡,Tan 等人提出了 EfficientNet[27],该网络在保持与其他模型相当的准确率的同时,显著提升了处理速度。随后,Tan 等人在 EfficientNet 的基础上,结合特征金字塔网络(Feature Pyramid Networks,FPNs),提出了 EfficientDet[28]。在这一模型中,FPN 被改进为双向特征融合网络(BiFPN),能够自上而下和自下而上地进行特征的有效融合,从而进一步增强了特征提取的能力。在本文中,我们选择 EfficientDet-D0 作为特征提取网络,其输入尺寸与 VGG16 网络相同,为 224×224×3。这一选择不仅确保了特征提取的高效性,还兼顾了模型的准确性。

3.2 目标检测

目标检测部分的目的是对输入图像中存在的物体进行精确的分类和边界框的定位。Faster R-CNN 以其高精度、端到端的训练方式、强大的特征提取能力、支持多种骨干网络和先进的 RPN 设计,在目标检测领域占据了重要地位。Faster R-CNN 网络模型如图 3 所示。相比 Yolo 等一阶段网络更为精准,尤其是对高精度及密集小物体的优势更为明显,并且 Faster R-CNN 更适合处理不同尺寸的目标,生成不同尺寸的候选区域。基于这些考量,本文采用 EfficientDet 作为 Faster R-CNN 的骨干网络,以增强模型的性能。其中 RPN 是区域建议网络,用于生成不同尺度的候选目标;RoI Pooling 则将 RPN 生成多样的区域转换为统一尺寸的特征,以便进行有效的分类和精确的边界框回归,最终得到某个类别的置信度和边界框的 4 个坐标。高检测率能够为后续物体间关系的预测提供较好的保障。

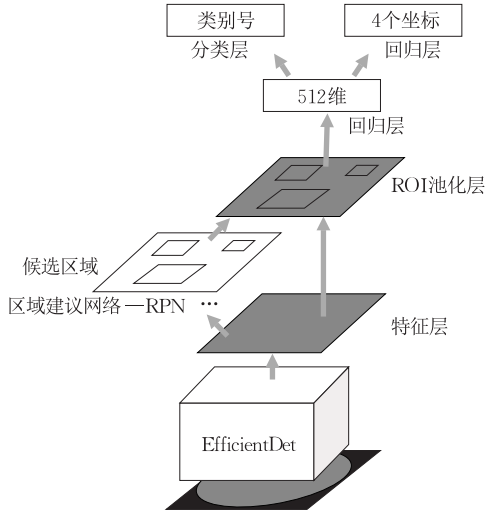

图 3 Faster R-CNN 网络模型

3.3 基于 GAT 的关系推理

关系推理部分是基于图注意力网络完成的,是整个网络中最重要的一环,也是能顺利完成物体间位置关系判断最关键的部分。基于图注意力的推理的全过程如图 4 所示。

首先,本文针对 RGB 信息中的所有物体及关系构建出一个无向图的形式,图的每个节点对应一个独立的物体,图的每条边对应一个隐式的关注关系,

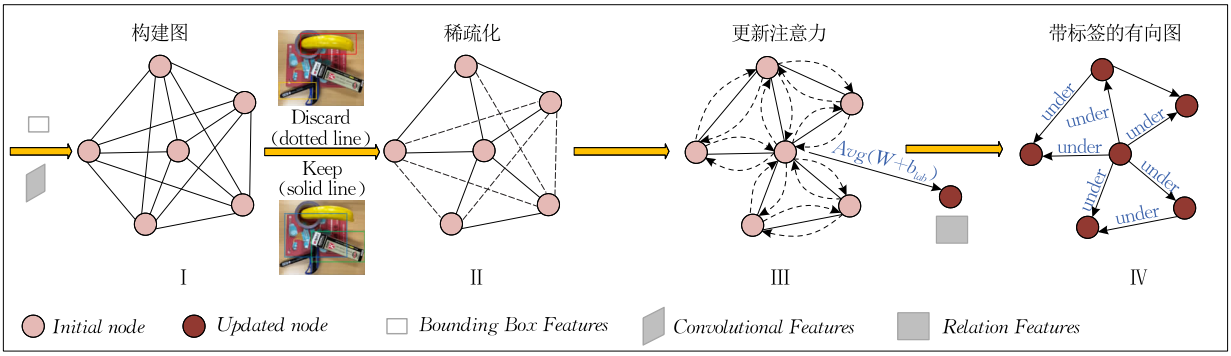

图 4　图网络的处理及更新过程

也就是目标物体与邻接物体的位置信息(当前是未知的),如图 4 中构建图步骤所示。根据图神经网络的定义 $G=(V,E)$,该图网络可以被定义为 $G=(O,\boldsymbol{R})$,其中 $O=\{o_1,o_2,\cdots,o_n\}$ 表示物体的集合,n 是物体的数量,$\boldsymbol{R}$ 表示物体-物体关系矩阵。物体间的关系可以被定义为三类,以目标物体为核心,分为在其上方(on)、在其下方($under$)以及与其分离(sep)。在关系矩阵中,这三种关系分别通过数值 1、−1 和 0 来进行标识和区分。

接着,根据目标检测网络生成出的每个物体的边界框,构建出一个初步的图邻接矩阵。我们将两个物体边界框的交集面积所占小物体边界框的面积比来表示两个物体存在上下位置关系的程度,用如下公式进行表示:

$$P_w=\frac{I(o_i,o_j)}{\min(S_{o_i},S_{o_j})},\ i\neq j \tag{1}$$

其中,$P_w\in(0,1)$,$I(o_i,o_j)$ 表示两物体的边界框相交的面积大小,$\min(S_{o_i},S_{o_j})$ 表示两物体中边界框面积较小的值。对于 P_w 值为 0 的物体-物体关系设为(sep),并删除主图形式中对应的边,使图变得稀疏,如图 4 中稀疏化步骤所示,以免之后在图推理的过程中进行重复计算。

最后,对处理后的图 $G'=(O,\boldsymbol{R})$ 进行位置关系推理。将两物体边界框相交部分的图像特征用于计算图的边特征,将每个独立的边界框用于计算物体的节点特征。鉴于生成的图是有向图,且同一物体的每个相邻物体与其存在上下位置的关系程度不同,构建一个图注意力网络来对有向图的节点特征进行更新,如图 4 更新注意力步骤所示,并且对边也进行了特征表示,可以和节点特征互相进行信息传递[29]。因为目标物体只与相邻的物体存在直接的上下方位置关系,即物体 i 在物体 j 的上方和物体 i 在物体 j 的下方,所以采用 Mask Graph Attention(掩膜注意力机制)的计算方式。该图网络的更新过程如下所示:

第一步计算注意力系数。首先将节点特征通过共享参数 W 进行维度扩展,之后利用串联操作将节点 i 和节点 j 经过参数变换后的特征进行拼接,最后利用注意力函数 a 将拼接后的高维特征映射为实数,如式(2)所示:

$$e_{ij}=a([Wh_i\coprod Wh_j]),\ j\in Ni \tag{2}$$

其中,i 表示正在计算的节点,j 表示 i 的一阶相邻节点,W 表示用于执行线性变换的可训练共享参数,e_{ij} 表示节点 j 对节点 i 的原始注意力得分。

第二步对注意力系数进行归一化。对上一步得到的 e_{ij} 使用 Softmax 函数进行归一化,确保对于每个节点 i,其所有相邻节点 j 的注意力系数 α_{ij} 之和为 1,如式(3)所示:

$$\alpha_{ij}=\mathrm{Softmax}(e_{ij})=\frac{\exp(\mathrm{LeakyReLU}(e_{ij}))}{\sum_{k\in N_i}\exp(\mathrm{LeakyReLU}(e_{ik}))} \tag{3}$$

其中,LeakyReLU 是非线性激活函数。

第三步是更新节点特征。节点 i 的新特征向量 $\boldsymbol{h}'_i$ 由所有相邻节点的特征加权求和得到,如式(4)所示:

$$\boldsymbol{h}'_i=\sigma\Big(\sum_{j\in N_i}\alpha_{ij}Wh_j\Big) \tag{4}$$

其中,σ 是一个非线性激活函数。

最后,由于所构成的图应包含标签信息,且具有方向性[30],需要将式(4)改写为式(5):

$$\boldsymbol{h}'_i=\sigma\Big(\sum_{j\in N_i}\alpha_{ij}(Wh_j+b_{lab(i,j)})\Big) \tag{5}$$

在式(5)中,$b_{lab(i,j)}$ 表示基于节点 i 和 j 之间直接关系的偏置项。我们通过引入额外的偏置项,模型可以学习到更丰富的信息,从而提高对图结构数据的表达和推理能力。

最终,我们可以推导出一个明确的有向图,如图 4 中 IV 部分所示,该图通过关系三元组描绘了目标对象与所有相邻实体之间的位置关系。

3.4 损失函数

本文提出的网络模型为一个结合了 EfficientDet-D0(用于特征提取并作为 Faster R-CNN 的骨干网络)、图注意力网络(用于关系推理),以及图稀疏化处理的复合模型。模型的损失函数主要包含物体检测损失和关系推理损失。

对于物体检测的损失,由于使用了 Faster R-CNN 的结构,应包含区域建议网络(Region Proposal Network,RPN)损失和检测器的损失。所以,模型的最终损失函数应表示为

$$L=\lambda_1 L_{\text{RPN}}+\lambda_2 L_{\text{Detector}}+\lambda_3 L_{\text{Relation}} \tag{6}$$

其中,λ_1、λ_2、λ_3为损失权重,RPN 损失和检测器损失均包含分类损失和回归损失。在分类任务下,RPN 的分类损失使用二元交叉熵损失,损失函数为

$$L_{\text{cls}}^{\text{rpn}}=-\frac{1}{N_{\text{cls}}^{\text{rpn}}}\sum_i(y_i^{\text{rpn}}\log(\hat{y}_i^{\text{rpn}})+(1-y_i^{\text{rpn}})\log(1-\hat{y}_i^{\text{rpn}})) \tag{7}$$

其中,y_i^{rpn}是锚点 i 的真实标签(1 表示包含物体,0 表示背景),$\hat{y}_i^{\text{rpn}}$是预测的概率。RPN 的回归损失使用 SmoothL1 损失,这一部分专注于锚点边界框的精确调整,损失函数为

$$L_{\text{reg}}^{\text{rpn}}=\frac{1}{N_{\text{reg}}^{\text{rpn}}}\sum_i \text{SmoothL1}(t_i^{\text{rpn}},\hat{t}_i^{\text{rpn}}) \tag{8}$$

其中,t_i^{rpn}是锚点的真实边界框参数,$\hat{t}_i^{\text{rpn}}$是预测边界框参数。同样地,检测器的分类损失和回归损失,也可以分用以下两个式子进行表示:

$$L_{\text{cls}}^{\text{det}}=-\frac{1}{N_{\text{cls}}^{\text{det}}}\sum_i \log(\hat{y}_{i,c_i}^{\text{det}}) \tag{9}$$

$$L_{\text{reg}}^{\text{det}}=\frac{1}{N_{\text{reg}}^{\text{det}}}\sum_i \text{SmoothL1}(t_i^{\text{det}},\hat{t}_{i,c_i}^{\text{det}}) \tag{10}$$

其中,c_i是物体 i 的真实类别,$\hat{y}_{i,c_i}^{\text{det}}$是对应类别的预测概率,$t_i^{\text{det}}$是真实的边界框参数,$\hat{t}_{i,c_i}^{\text{det}}$是预测的边界框参数,仅对正样本(即包含物体的样本)计算。

对于关系推理的损失,是基于分类的关系推理,需要推理出每对物体间的关系的离散类别(如“上方”“下方”“无关”等),交叉熵损失可以有效地用于这种多类分类问题:

$$L_{\text{Relation}}=-\sum_i y_i\log(\hat{y}_i) \tag{11}$$

4 实 验

4.1 数据集和评价指标

本文在视觉操纵关系数据集(Visual Manipulation Relationship Dataset,VMRD)上进行了实验,并对模型的性能进行了评估。VMRD 数据集是由 Zhang 等人提出的,旨在帮助机器人在多物体无序堆叠场景中识别物体之间的空间关系,从而提高机器人学习和感知环境信息的能力,以实现安全可靠的抓取操作[31]。该数据集包含 31 类物体,共 4683 张图像,涵盖超过 43000 个操纵关系。每张图像展示了 2 至 5 个物体的堆叠场景,并且每个物体都通过矩形框进行了标注,附带相应的类别标签。在实验过程中,我们对整个数据集进行了划分,编写了一个数据划分程序(split),将 4233 张图像分配为训练集(约占数据集的 90%),其余 450 张图像作为测试集(约占 10%)。此外,考虑到数据集中存在大量的文件名标签信息与图像名称不一致的情况,我们还设计了一个校验程序(check),用于批量校正这些名称,以确保数据的一致性和准确性。通过这些步骤,我们为后续的模型训练和评估奠定了基础。

当评估检测模型时,mAP、OR、OP 和 IA 是视觉操纵关系推理任务中常用的几种评价指标,各自代表不同的性能衡量标准。

(1)mAP(Mean Average Precision)表示为平均精度均值,它是衡量目标检测模型在所有类别上平均性能的一个指标。在目标检测中,精度(*Precision*)是指模型正确检测到的物体数量占所有检测到的物体数量的比例,召回率(*Recall*)是指在所有真实目标中,正确检测到的目标所占的比例。AP(Average Precision)是通过绘制精度-召回率(*Precision-Recall*)曲线并计算曲线下面积 AUC(Area Under the Curve)得到的。mAP 是先对每个类别的 AP,然后对所有类别的 AP 取平均值,可以用以下的式子来表达:

$$mAP=\frac{1}{C}\sum_{c=1}^{c} AP_c \tag{12}$$

其中,C 表示类别,对于每个类别 C,mAP 是该类别的平均精度。

(2)OR(Object-based Recall)表示对象召回率,它是衡量模型能否正确检测出图像中所有相关对象对以及它们的关系。一个对象对的检测结果被视为正确是指两个对象都被正确检测(类别正确且预测边界框和真实边界框的 IoU 大于 0.5),并且它们之间的操作关系也被正确预测。可以具体为以下的式子:

$$OR=\frac{TP}{T} \tag{13}$$

其中,TP(True Positives)是模型正确预测的对象对数量,而 T 是测试集中所有实际存在的相关对象对的数量。

(3)OP(Object-based Precision)表示对象精

度，它用于衡量模型预测的对象对中有多少是正确的。与 *OR* 类似，正确性的标准是两个对象的检测和关系预测都必须正确。也可以具体为下式：

$$OP=\frac{TP}{P} \tag{14}$$

其中，TP 仍是模型正确预测的对象对数量，而 P 是模型预测出的所有对象对的数量，并不一定与实际存在的 T 相等。

(4) *IA*(Image-based Accuracy)表示为图像准确率，它是衡量整个图像级别上关系推理准确率的指标。这个指标检查一个图像中所有可能的相关对象对及其关系是否都被正确识别。如果一个图像上所有对象及它们的关系都检测正确，那么这张图片就被认定为检测正确，可以用以下关系式来定义：

$$IA=\frac{C_{\text{images}}}{T_{\text{images}}} \tag{15}$$

其中，C_{images} 是预测正确的图像数，T_{images} 为总的评估图像数量。

4.2　实验设置

实验的代码部署在 Ubuntu 18.04 系统的服务器上，使用一块显存为 12 GB 的 NVIDIA GTX 1080Ti GPU 进行训练，显卡驱动为 470.161.03，CUDA 版本为 10.1，pytorch 版本为 1.7.0，后续在更换网络的 backbone 后使用了 cuda 11.3 和 pytorch 1.10.1 的版本。

由于需要对 ResNet50、ResNet101、VGG16 和 EfficientDet-D0 作为 Faster R-CNN 网络结构的 backbone 训练特征提取网络，并进行目标检测，我们针对不同的网络结构编写好不同的 yml 配置文件并进行替换训练，方便参数的设置。统一输入 RGB 图像裁剪为 $W\times W\times 3$，学习率初始值设为 0.001，每个批次训练的图像数设为 4，最大训练周期数设为 30，并使用数据增强防止过拟合。

具体来说，输入的 RGB 图经压缩后提取到特征图，RPN 网络从特征提取后的结果中选取 256 个锚点，其中包括 128 个前景和 128 个背景，通过 RPN 的卷积计算和非极大抑制(NMS)处理，去除重叠的边界框，留下最佳的候选框。然后，ROI Align 从每个候选框中提取 7×7 的特征图，再将其展平并通过一系列全连接层进行处理，转换成 n 个对象和 m 个关系的 1024 维特征矩阵。最后输入到 2 层图注意网络进行训练，更新节点特征和边的特征，其中隐藏层维度设为 256，头数设为 8。其他超参数的设置如表 1 所示。

表 1　模型训练中使用的超参数

超参数	Value
学习率衰减步数	10000
学习率衰减因子 γ	0.1
每批次的图像数量	2
权重衰减率	0.0005
RPN 在每次迭代中使用的样本数量	256
RPN 中正样本的 *IoU* 阈值	0.7
整个 RCNN 模型的批次大小	256
背景 *IoU* 阈值的下限	0
应用 NMS 前 RPN 保留候选框的数量	12000
应用 NMS 后 RPN 保留候选框的数量	2000

4.3　实验结果

为了充分展示 Faster R-CNN 网络的性能优势，我们使用相同的骨干网络来对 Faster R-CNN 与 SSD 两种网络在同一数据集上进行目标检测任务的表现进行了比较。性能评估指标为平均精确率(mAP)。实验结果如表 2 中显示，从中可以观察到，Faster R-CNN在检测精度上相较于SSD显示出更优异的表现。

表 2　Faster R-CNN与SSD使用相同骨干网络时的性能比较

目标检测网络	骨干网络	*mAP*
SSD	VGG16	94.2
	Resnet101	91.3
	EfficientDet	94.9
Faster R-CNN	VGG16	95.2
	Resnet101	95.4
	EfficientDet	96.1

本文提出的方法在VMRD数据集上与Multi-task CNN、VMRN 模型和 GVMRN 模型的不同结构进行了相同条件下的测试比较。表 3 和表 4 分别展示了各模型在不同评价指标下的量化结果，各指标的最大值加粗进行表示。

表 3　不同网络模型的比较结果

网络模型	*mAP*	*OR*	*OP*	*IA*
Multi-task CNN-Res101[31]	—	86.0	88.8	67.1
VMRN-Res101[11]	95.4	85.4	85.5	65.8
GVMRN-Res101[14]	94.5	86.3	87.1	68.0
VMRN-VGG16[11]	94.2	86.3	88.8	68.4
GVMRN-VGG16[14]	95.4	87.3	89.6	69.7
VMRN-Res50	93.5	79.9	79.3	51.3
GAT-EfficientDetD0(ours)	**96.1**	**88.8**	**89.4**	**71.1**

表 4　不同物体个数下的 *IA* 指标

网络模型	*IA*	每张图片中的物体个数			
		2	3	4	5
Multi-task CNN-Res101[31]	67.1	87.7	64.1	56.6	72.9
VMRN-Res101[11]	65.8	90.8	63.2	55.7	65.7
GVMRN-Res101[14]	68.0	90.0	68.8	60.3	56.2
VMRN-VGG16[11]	68.4	90.8	66.5	60.4	67.1
GVMRN-VGG16[14]	69.7	91.4	69.9	62.9	58.9
VMRN-Res50	51.3	86.2	46.4	35.8	57.1
GAT-EfficientDetD0(ours)	**71.1**	**92.3**	**68.9**	**64.2**	**68.6**

从表 3 的结果中可以看出，本文提出的方法在 VMRD 数据集上进行评估，*mAP*、*OR* 和 *IA* 指标都达到了最高的精度。这反映了本方法在目标检测和物体关系推理任务上的性能均有所提升。具体来说 *mAP* 指标达到了 96.1%，这表明 EfficientDet 网络中独特的 BiFPN 结构，能够更好地对图像不同尺度的特征进行融合，有效地提高了图像检测的平均精度。对象召回率(*OR*)和图像准确率(*IA*)的显著提升，表明了本方法在推理阶段能够正确地推理出更多的物体关系对。虽然对象精度(*OP*)与 GVMRN-VGG16 相比略有下降，但是根据上述评价指标中的式(3)～(13)和式(3)～(14)可知 *OR* 越大 *TP* 越大，而 *OP* 却变小，说明 *P* 的值越大，从另一方面反映出本方法在总体对象对的预测方面更为积极。综上所述，本文提出的基于图注意力网络的视觉关系推理方法能够更准确地识别出对象的信息，从而推断出更多正确的关系对。

表 4 表示的是不同物体数量下的 *IA* 指标以及预测正确的图片占测试集图片总数的比率。从中可以看出本文提出的方法能够正确地预测出最多的图片数，正确比例为 71.1%，整体正确率最高。分开看不同物体个数的 *IA* 指标，随着每张图片中物体个数的增加，*IA* 在绝大多数模型中呈现出下降的趋势。这是因为物体数量的增加，场景变得复杂，导致模型更难准确地识别和推理物体之间的关系。然而当物体个数为 5 时相比物体个数为 4 时的 *IA* 指标却有所上升，原因可能是测试的数据集中物体个数为 4 的图片中出现了较高的平行遮挡情况，也就是将视觉上没有上下位置关系的物体对进行了错误识别。本方法在物体个数为 2 和 4 的情况下取得了最高精度，在物体个数 3 和 5 的情况下取得了次高精度。总的来说，不同模型在处理不同物体个数的图像时表现出一定的差异，但本方法在多物体场景下的整体表现更为出色，具有最高的图像准确率(*IA*)。

在图 5 中，呈现了本方法在 VMRD 数据集上的部分可视化测试结果，我们分别选取了 2～5 个物体不等的图片各两张，分为上下两组，从左到右每一列的物体个数依次为 2、3、4、5。第一、三行是图像识别输出的结果，包括物体的索引、类别、置信度以及抓取的顺序(1～2 表示先抓 1 再抓 2)；第二、四行是物体间的位置关系示意图，箭头表示从下方的物体指向其上方物体，而打叉的箭头表示错误的关系推理。通过结合模型测试的可视化成果，分析发现物体数量为 5 时 *IA* 值异常上升的原因主要有两个：一是物体个数为 5 的图片整体样本量要远小于其他物体个数的图片；二是物体个数为 5 的图片中，大多

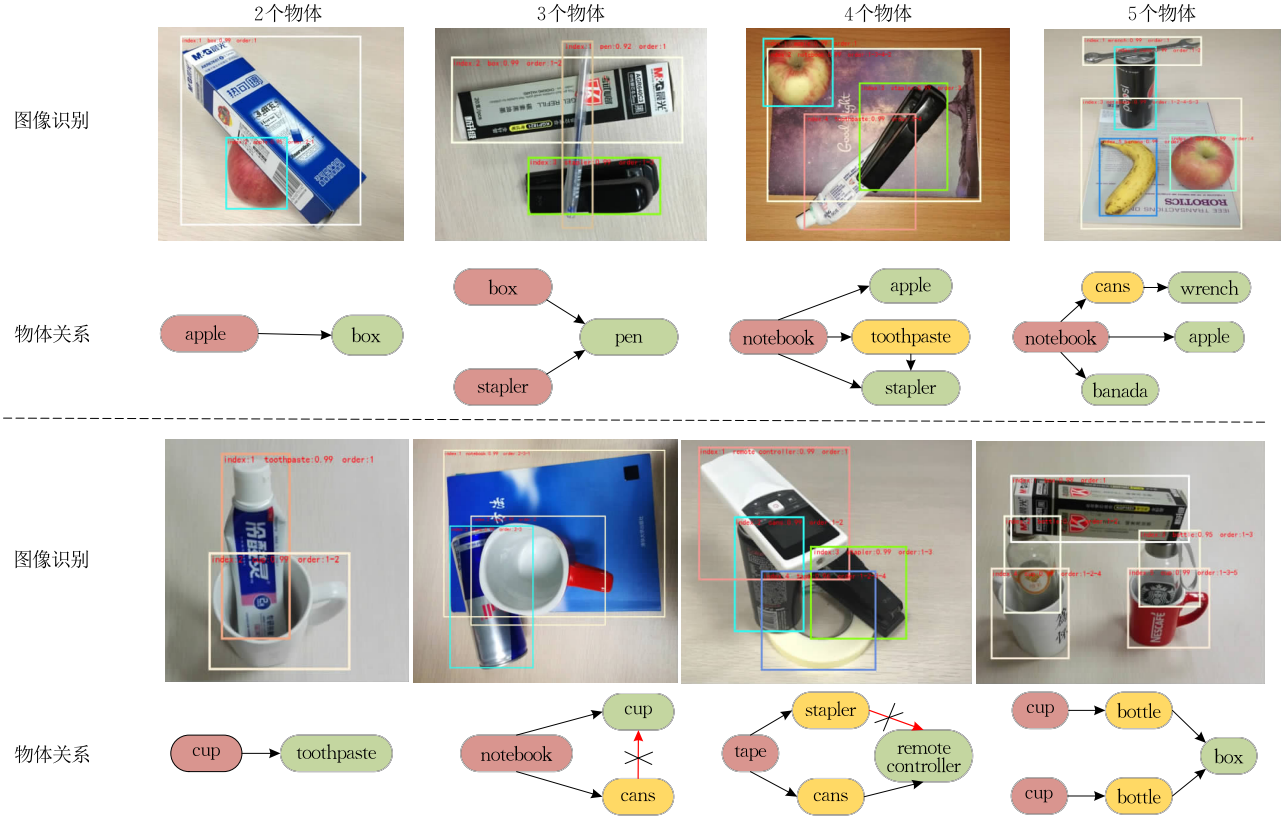

图 5 网络模型在 VMRD 数据集上的测试结果

数物体摆放得较为分散，出现因视觉上遮挡而实际并不存在上下位置关系的物体对较物体个数为 3 或为 4 图片情况要少。

4.4　模型分析

为对本文提出模型的有效性进行验证，我们采用了一种名为梯度加权类激活映射的方法，简称为 Grad-CAM，它是一种用于解释深度卷积神经网络决策过程的技术，其主要目的是可视化神经网络在图像分类任务中对输入图像的注意力分布，以协助理解模型的预测决策过程。我们将已经在测试中表现良好的训练模型应用于 Grad-CAM，通过该方法可视化特征图。对于测试图中，我们随机选择了一组物体关系对，并展示了对应的热力图，如图 6 所示。图中的第一行和第三行展示了两组物体个数分别为 2、3、4、5 的原始测试图片；第二行和第四行则展示了经过 Grad-CAM 方法处理过的相应物体关系对的热力图。这种方法有助于更全面地理解模型对不同物体关系的关注程度，从而进一步证明了模型的有效性。

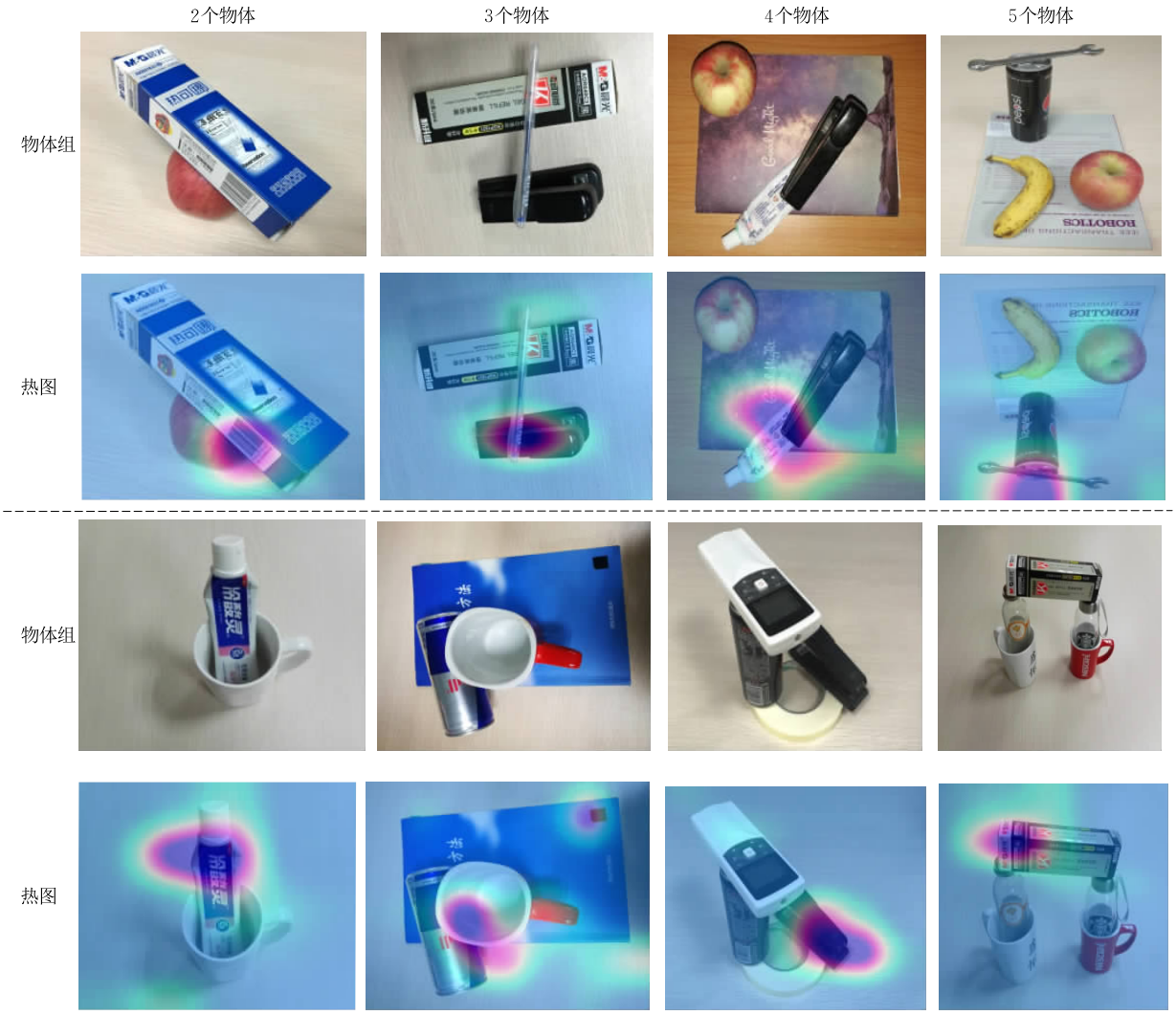

图 6　模型基于视觉关系推理的可视化激活特征

根据图 6 中的结果显示，热力图能够有效地集中在所选的物体关系对上。热力图的颜色深浅反映了输入图像在该区域对最终分类的贡献程度，颜色越深表示该区域对最终分类的影响越大。尽管在图中某些物体对的热力图最深的区域并未显示在两物体的准确中心，然而在图 5 中，模型仍然成功进行了物体识别、分类和位置关系的推理。此外，我们观察到模型的关注点主要集中在两个物体之间的接触位置，而对两个物体的主体区域关注较少。这意味着在检测过程中，如果出现了没见过的未知物体，我们需要的目标物体（已知物体）与未知物体间的关系也会被识别出来，这为预测已知物体与未知物体之间的位置关系提供了可能性。该注意力图定性揭示了本文所述模型更关注物体重叠部分信息的特性，从而更好地推理出物体之间的位置关系。通过注意力热力图来可视化本文所述方法在视觉关系推理中更

关注的位置,从而有助于提高图像准确率,具备更强的推理能力。

4.5 机器人抓取实验

本文基于 AUBO-i5 机械臂结合二指电动夹爪和深度相机,建立视觉抓取实验平台。为了验证所提出模型的实际应用能力及其泛化性,我们在实验室环境下构建了一个特定的测试集。该测试集基于 VMRD 数据集,选取了其中的一部分物品进行实验。这些物体虽然属于相同类别,但彼此并不完全相同。根据 VMRD 数据集中 450 个测试样本的组成比例,新组成的测试集共包含 100 组,每组由 2 至 5 个物体构成。具体地,组成的测试集包括 14 组两物体场景、46 组三物体场景、24 组四物体场景和 16 组五物体场景,部分示例图像如图 7 所示。

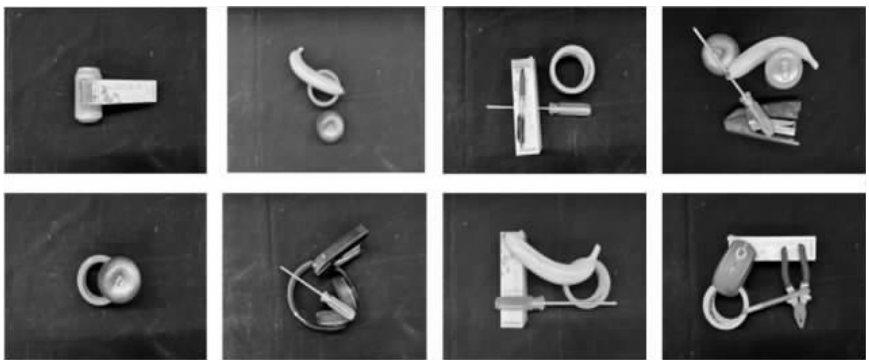

图 7 不同物体个数基于图像精度的比较

我们将收集到的真实抓取场景示例作为新的测试集来测试模型的泛化能力,并将测试结果与之前的 VMRD 测试集进行比较。测试结果如表 5 所示,其中第 5 至 8 列的数据表示了模型推理正确的场景数量与总图片数量的比率。表 5 中的数据可以清楚地表明,该模型在实际抓取环境中展示了较好的推理能力、实用性和泛化性,但是在 *OR*、*OP* 和 *IA* 的指标上的总体表现低于先前针对 VMRD 数据集的测试结果。这一差异可能由以下三个因素导致:首先,新测试集中的对象属于 VMRD 数据集中相同的类别,但实际对象并不完全相同,这一差异可能影响了模型识别的准确性。其次,数据采集环境不同,新采集数据的背景是黑色的。最后,在"眼在手外"的视觉抓取系统中,工作区的物体只能观察到俯视图,这与 VMRD 数据集中多视角的数据构成存在明显区别。因此,这些因素共同导致新测试集的总体性能低于 VMRD 测试集的表现。

表 5 不同对象数量的 *IA* 指标比较结果

数据集	*OR*/%	*OP*/%	平均 *IA*/%	推理正确场景数/总场景数			
				IA-2	*IA*-3	*IA*-4	*IA*-5
实验室采集数据集	85.7	87.2	67.0	12/14	31/46	15/24	9/16
VMRD 测试集	**88.8**	**89.4**	**71.1**	**60/65**	**144/209**	**68/106**	**48/70**

在实际的机械臂抓取过程中,仅仅识别目标物体所在的位置关系并推导出操作顺序并不足以完成有效的抓取任务,还必须结合一定的抓取方法。因此,本文采纳了二维平面抓取策略,即目标物体被置于平面工作空间内,并且抓取动作被限制在垂直方向上进行。具体来说,抓取过程中需要考虑的关键参数包括抓取点的坐标(x、y、z)、机械臂末端的旋转角度 θ 以及双指夹具的开口宽度 w。我们采用了一种较新的基于抓取矩形的评估方法[32]对 VMRD 数据集进行训练,旨在为深度摄像头所捕获的真实抓取场景生成最佳抓取矩形。图 8 展示了在视觉指导下,机械臂完成整个抓取过程的步骤。

图 8 展示了两种机械臂抓取策略的对比分析,均以螺丝刀作为目标抓取物体。A 部分表示直接抓取目标物体的过程,而部分 B 则演示了采用推理抓取策略的过程。图中第一列是场景图,第二列为抓取顺序,第三列是物体的识别检测,第四列为待抓取物体的最佳抓取矩形的生成,第五列是机械臂执行抓握动作。在 A 部分直接抓取的过程中,由于螺丝刀在盒子下方,直接抓取会将盒子打翻,甚至可能离开工作区域。而在 B 部分推理抓取的操作中,机械臂会首先移开位于螺丝刀上方的盒子,之后才执行

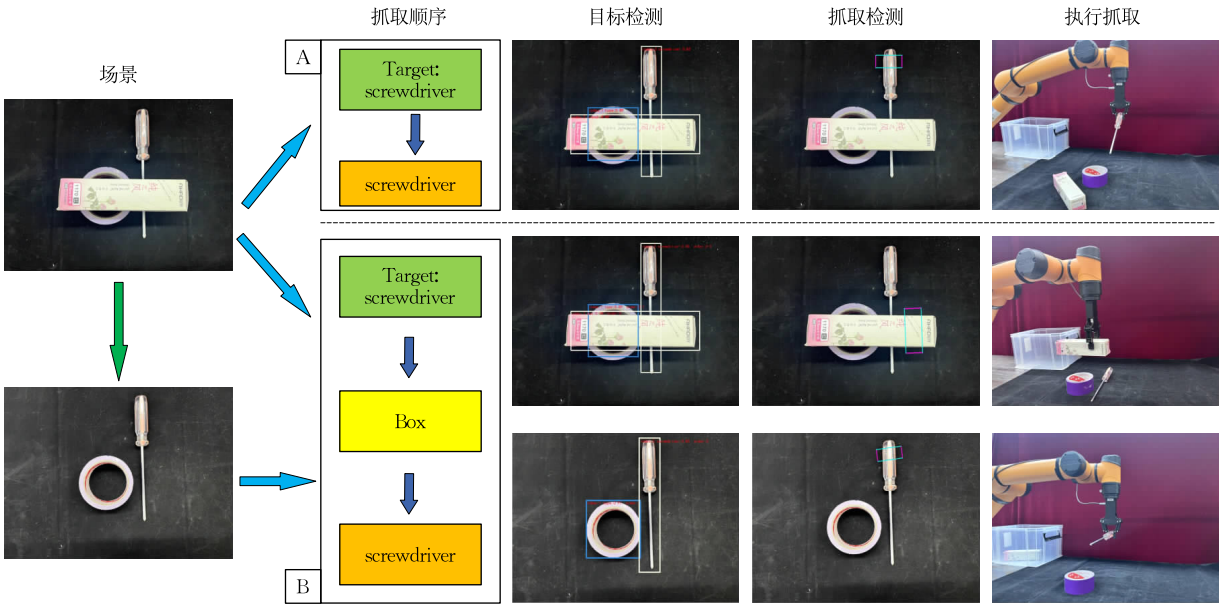

图 8　直接抓取与推理抓取策略过程的比较

对螺丝刀的抓取。虽然这种策略增加了操作步骤，但它确保了抓取过程的安全性和可靠性。上述实验结果表明，使用推理抓取算法不仅能提高抓取成功的概率，还能减少对其他物体的影响。

5　总　结

本文研究了一种将图注意网络应用于视觉运算关系推理的解决方案，可以更准确地获取多物体堆叠场景中物体之间的位置关系。我们以 Efficient-Net-B0 为骨干，结合 BiFPN 作为特征提取模型，通过双向特征融合提高了识别精度。在关系运算推理部分，我们利用图注意网络，根据不同物体与同一物体的接近程度设置不同的权重，从而获取物体之间的位置关系。此外，我们还对图进行了稀疏化处理。我们在相同的数据集上与其他方法进行了对比实验，结果表明我们的模型提高了关系推理的准确性，并且可以在真实的机械臂抓取场景中应用和推广。不过，我们提出的基于 RGB 的图注意网络仅适用于预测空间中可见物体之间的关系，我们的研究验证了当图像包含 2 到 5 个物体时的应用场景。我们未来的研究计划将整合机器人的操作动作，并考虑如何通过可见物体的位置关系获取遮挡物体的信息。此外，我们还将研究如何提高模型在包含更多物体（超过五个）的场景中的性能，以及在更复杂的实际操作场景中进行物理实验，以进一步验证有效性与发现可供改进的新问题。

参　考　文　献

[1] Duan S, Tian G, Wang Z, et al. A semantic robotic grasping framework based on multi-task learning in stacking scenes. Engineering Applications of Artificial Intelligence, 2023, 121: 106059

[2] Mohammed M Q, Kwek L C, Chua S C, et al. Review of learning-based robotic manipulation in cluttered environments. Sensors, 2022, 22(20): 7938

[3] Zuo G, Tong J, Liu H, et al. Graph-based visual manipulation relationship reasoning network for robotic grasping. Frontiers in Neurorobotics, 2021, 15: 719731

[4] Zhang H, Lan X, Zhou X, et al. Visual manipulation relationship recognition in object-stacking scenes. Pattern Recognition Letters, 2020, 140: 34-42

[5] Selvaraju R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization //Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 618-626

[6] Guo D, Kong T, Sun F, et al. Object discovery and grasp detection with a shared convolutional neural network//Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA). Stockholm, Sweden, 2016: 2038-2043

[7] Fischinger D, Vincze M, Jiang Y. Learning grasps for unknown objects in cluttered scenes//Proceedings of the 2013 IEEE International Conference on Robotics and Automation. Karlsruhe, Germany, 2013: 609-616

[8] Gualtieri M, Ten Pas A, Saenko K, et al. High precision grasp pose detection in dense clutter//Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Daejeon, Republic of Korea, 2016: 598-605