融合全局视角校正与细粒度语义解耦的机器人指令抓取检测方法

刘 进[①] 刘志太[②] 李梓函[③] 孙彦景[①③] 缪燕子[③] 袁宪锋*[④]

[①](中国矿业大学计算机科学与技术学院/人工智能学院 徐州 221116)

[②](哈尔滨工业大学航天学院 哈尔滨 150001)

[③](中国矿业大学信息与控制工程学院 徐州 221116)

[④](山东大学低空科学与工程学院 威海 264209)

摘 要：根据语言指令准确地抓取目标物体，是服务机器人实现自然人机交互的必要前提。现有研究多采用大规模数据驱动训练或层次化特征融合策略实现视觉信息与指令文本的模态对齐，普遍忽视了底层视觉特征中目标与环境的强耦合特性，导致其在跨视角和未知背景下的组合泛化能力显著下降。针对上述挑战，本文构建了一个双视角跨场景同步抓取检测和定位数据集，用于系统评估并针对性提升现有模型的组合泛化能力。在此基础上，提出了一种指令驱动下的同步抓取检测和物体定位网络(SGL-Net)。首先，引入了跨模态全局上下文调制模块(CGCMM)，利用自然语言指令的语义先验，在视觉特征提取的早期阶段对网络视觉通道进行自适应视角校正与背景干扰抑制。其次，设计了词-像素跨模态交叉对齐模块(WPCAM)，采用扁平化注意力机制实现细粒度语义解耦，用以提升网络对动态复杂场景的语义理解能力。最后，采用统一解码器架构解耦融合后的多模态特征，实现目标物体位置与最优抓取位姿的同步输出。充足的定量与可视化实验结果表明，所提网络在多种不同场景视角的组合情况下均取得了最优的性能，并具备在真实物理世界中部署应用的能力。

关键词：抓取检测；指令驱动；多模态融合；复杂场景

中图分类号：TP242.6 **文献标识码**：A **文章编号**：1009-5896(2026)00-0001-11

DOI: 10.11999/JEIT260442 **CSTR**: 32379.14.JEIT260442

1 引言

随着机器人技术的持续发展与具身智能的快速演进[1,2]，服务机器人正逐步从结构化的工业环境走向开放、动态的室内场景，这对在复杂环境中进行连续作业与操作泛化的能力提出了极高的要求[3,4]。在此背景下，机器人不仅需要具备鲁棒的视觉感知系统，还需要深刻理解人类的自然语言指令并准确完成指定抓取操作[5]。作为连接高层语义理解与底层物理执行的关键枢纽，指令驱动的抓取检测旨在将无约束的自然语言描述精确映射为目标物体的空间位置及平面抓取位姿。一个典型的指令抓取过程通常包含指令语义解析、作业环境感知、目标视觉定位、抓取位姿预测及底层物理执行五个关键步骤。因此，相比于传统仅依赖单一视觉信息的抓取检测任务[6,7]，指令驱动的抓取检测任务在跨模态语义对齐与作业场景认知上面临着更为严峻的挑战。

早期研究多采用视觉感知与语义理解解耦的策略，通过构建级联式的检测、定位、抓取流水线来完成指令操作任务。尽管该方式中单一模块的性能已趋于完善，但这种开环的处理模式容易引发严重的误差累积效应，当检测结果出现错误后将难以保证整体抓取任务的成功率。为此，近期一些研究工作聚焦于构建视觉与语言之间的端到端映射关系。例如，Chen等人[8]提出命令抓取网络，实现了从自然语言与图像输入到抓取位姿的直接预测。Xu等人[9]进一步引入视觉-语言-动作联合建模框架，以提升复杂场景中的任务执行能力。与此同时，研究者们也在不断探索如何利用更精细的感知技术来增强模型的判别能力。Bousselham等人[10]与Yin等人[11]分别验证了跨模态特征对齐与多模态大模型在提升目标定位可靠性方面的潜力。而在语义解析方面，针对复杂背景的特征挖掘[12]以及边界细节细化[13]等技术的进步，也为端到端系统提供了更丰富的环境先验信息。随着感知维度的不断丰富，研究者开始探索更深层次的跨模态特征融合机制。Shridhar等人[14]提出CLIPort框架，通过结合预训练视觉-语言模型的知识与空间操作网络，实现了语义信息与操作空间的有效对齐融合。沿着这一思路，分层多模态融合结构[15]与掩码引导注意力机制[16]相继被提出，显著增强了系统在遮挡环境下的目标感知与消歧能力。近年来，面对真实服务场景

收稿日期：2026-04-14；改回日期：2026-07-14；网络出版：2026-07-24

*通信作者： 袁宪锋 yuanxianfeng@sdu.edu.cn

基金项目：国家重点研发计划(2025YFB4712700)，中国博士后科学基金资助项目(2025M781675)

Foundation Items: National Key R&D Program of China (Grant No. 2025YFB4712700), and the China Postdoctoral Science Foundation (Grant No. 2025M781675)

中语义理解受限及完全监督数据标注成本高昂的瓶颈，Xie等人[17]通过多源知识注入与属性感知建模，赋予了模型更深层次的常识推理能力。同时，针对多模态收敛不稳定及感知-抓取不一致的问题，其团队进一步提出基于课程调度与几何一致性的半监督框架[18]，有效提升了模型的数据利用效率与目标定位精度。尽管上述方法在引入语义信息后显著提升了模型对目标物体的理解与选择能力，但其跨模态融合过程大多依赖于后期融合策略[19]，即在特征提取完成后对全局视觉表示与语言表示进行拼接或匹配。这种粗粒度的融合方式难以在空间层面建立语言语义与局部视觉区域之间的精确对应关系，在复杂场景中容易受到视角变化及背景干扰的影响，进而导致语义对齐偏移与特征混淆问题，限制了模型在跨视角与复杂环境下的组合泛化能力。

为应对指令驱动抓取在跨视角与复杂环境下的泛化能力退化问题，本文从底层跨模态特征融合机制出发，提出了一种基于语义-视觉协同的同步抓取检测及物体定位网络(Simultaneous Grasp detection and object Localization Network, SGL-Net)。针对现有方法依赖后期融合导致的空间信息破坏与语义对齐不足问题，本文在特征交互阶段引入早期跨模态调制机制，以全局语言语义为条件对视觉特征进行通道级自适应校正，缓解视角变化与背景干扰带来的特征偏移。进一步地，通过引入词-像素跨模态交叉对齐机制，利用密集注意力实现语言片段与局部视觉区域之间的精细对应，从而提升复杂场景中的目标辨识能力。在解码预测阶段，采用共享特征与多任务分支协同设计，实现物体位置与抓取位姿的联合预测。通过上述设计，所提网络在复杂环境中显著提升了语义对齐精度，并在跨视角条件下表现出更强的组合泛化能力。

2 方法

针对指令驱动下的抓取检测算法在跨视角和复杂环境下面临的泛化退化问题，本文提出了如图1所示的同步抓取检测及物体定位网络SGL-Net，主要包含如下五个关键部分：(1)感知输入部分，用于获取RGB图像和人类指令；(2)特征提取部分，用以提取视觉和文本的隐含特征；(3)跨模态全局上下文调制模块(Cross-modal Global Context Modulation Module, CGCMM)，用以对早期视觉特征通道进行视角与背景的自适应融合；(4)词-像素跨模态交叉对齐模块(Word-Pixel Cross-modal Alignment Module, WPCAM)，用以对视觉像素进行细粒度跨模态解耦；(5)解码部分，用以同时输出物体的位置信息和最优抓取位姿。

2.1 感知输入部分

该部分构成了网络数据输入的基础。在视觉信息方面，机器人通过RGB相机获取分辨率为640×480的图像$\boldsymbol{I}\in\mathbb{R}^{3\times640\times480}$，本研究主要选取机器人顶部(Top)和底部(Bottom)两个观测视角的图片。在指令信息方面，机器人接收人类自然语言指令$\boldsymbol{H}=\{w_t\}_{t=1}^{L}$作为输入，其中$L$为单词总数目。为适配机器人机载运算单元性能，单条指令单词数上限设为20。

2.2 特征提取部分

该部分主要负责将视觉输入与文本指令映射为高维特征向量。具体而言，本文引入对比语言-图像预训练模型(Contrastive Language-Image Pre-training,

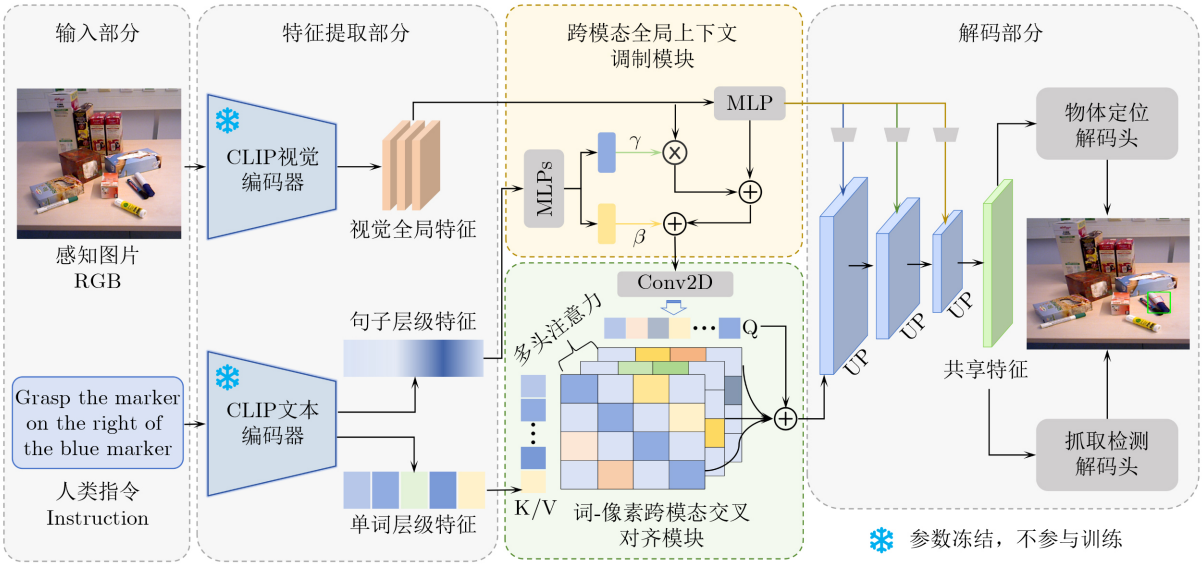

图 1 本文所提网络架构图

CLIP)[20]作为视觉特征和文本特征的编码器。依托于大规模图文对数据集的预训练先验，CLIP模型具备强大的多模态信息表征与特征对齐能力。为维持其固有的预训练知识空间，模型的所有参数在本研究的训练和测试阶段均保持冻结，不参与网络的反向传播与参数更新。在文本编码阶段，为提取细粒度的语言指令信息，本网络提取CLIP文本编码器最后一层中对应于所有输入单词的隐藏状态序列，定义为单词层级特征$\boldsymbol{T}_{\mathrm{w}} \in \mathbb{R}^{L\times 512}$，指令的句子层级特征为$\boldsymbol{T}_{\mathrm{s}} \in \mathbb{R}^{1\times 1024}$。在视觉编码阶段，提取的稠密视觉特征$\boldsymbol{V}_{\mathrm{g}} \in \mathbb{R}^{1\times 2048\times 10\times 10}$。

2.3 跨模态全局上下文调制模块

传统视觉-语言抓取检测任务多采用后期融合策略，即将提取完毕的全局图像特征与句子特征在网络末端进行简单的拼接或内积操作。如图2左侧所示，这种粗粒度的黑盒融合方式不仅破坏了图像原有的二维空间拓扑结构，且难以建立局部视觉区域与具体语言指令片段(如名词属性、空间方位词)之间的细粒度语义关联，导致其在复杂动态场景下极易产生“语义对齐偏移”与“特征混淆”。为此，本文提出采用语义引导的前置调制融合策略，并设计了跨模态全局上下文调制模块。如图2右侧所示，该模块在特征融合的早期阶段，利用指令的全局句向量作为条件输入，对深层视觉特征进行通道级的动态自适应校正。这样使得网络能够忽略与指令无关的背景噪声，增强模型对跨视角场景的宏观适应性。

具体而言，给定全局视觉特征$\boldsymbol{V}_{\mathrm{g}} \in \mathbb{R}^{16\times 2048\times 10\times 10}$与指令的句子层级特征为$\boldsymbol{T}_{\mathrm{s}} \in \mathbb{R}^{1\times 1024}$。模块首先通过一个包含ReLU激活函数的多层感知机MLP对文本特征进行非线性映射，提取出精炼后的上下文先验，并将其投影至视觉特征通道的两倍维度空间。随后，该联合参数在通道维度上被对半拆分为缩放因子γ与平移因子β，该过程表达式为：

$$\begin{aligned}\boldsymbol{P}_{\mathrm{film}} &= \mathrm{MLPs}(\boldsymbol{T}_{\mathrm{s}})\\ [\gamma,\beta] &= \mathrm{Chunk}(\boldsymbol{P}_{\mathrm{film}}, 2, \mathrm{dim}=-1)\end{aligned} \tag{1}$$

为实现与二维视觉特征图的空间维度匹配，网络对γ和β进行维度扩展。在特征调制阶段，为保证网络在训练初期的收敛稳定性并促进残差学习，本文采用带有恒等映射偏置的仿射变换策略，对视觉特征进行逐通道调制，该过程表达式为：

$$\boldsymbol{F}_{\mathrm{mod}} = (\boldsymbol{1}+\gamma)\odot \boldsymbol{V}_{\mathrm{g}} + \beta \tag{2}$$

式中，$\boldsymbol{1}$为全1张量$\odot$表示基于张量广播机制的按元素相乘操作。与传统后期融合相比，这种前期的动态调制机制使得网络能够在多模态交互的初始阶段，就自适应地强化特定视角下的目标特征响应，并从根源上抑制无关的背景噪声，从而为后续的微观对齐提供了更纯粹的特征基础。

2.4 词-像素跨模态交叉对齐模块

在经过全局视角与背景校正后，为进一步对抗物体堆叠与相互遮挡带来的特征混淆，本文随后引入词-像素跨模态交叉对齐模块。该模块利用局部词向量序列作为细粒度的语义引导，对视觉像素进行密集的语义解耦与交叉对齐。具体而言，给定经过二维卷积网络降维后的视觉特征$\boldsymbol{V}_{\boldsymbol{g}}' \in \mathbb{R}^{1\times 1024\times 10\times 10}$以及单词层级特征为$\boldsymbol{T}_{\mathrm{w}} \in \mathbb{R}^{L\times 512}$。本模块首先打破视觉特征的二维拓扑限制，将其在空间维度上进行展平与转置，构建跨模态查询矩阵。同时，通过线性投影网络将文本词特征映射至统一的视觉维度空间，构建键矩阵与值矩阵，该过程表达式为：

$$\begin{aligned}\boldsymbol{Q} &= \mathrm{Flatten}(\boldsymbol{V}_{\mathrm{g}}')^{\mathrm{T}} \in \mathbb{R}^{1\times 100\times 1024}\\ \boldsymbol{K} &= \boldsymbol{V} = \mathrm{MLPs}(\boldsymbol{T}_{\mathrm{w}}) \in \mathbb{R}^{1\times 20\times 1024}\end{aligned} \tag{3}$$

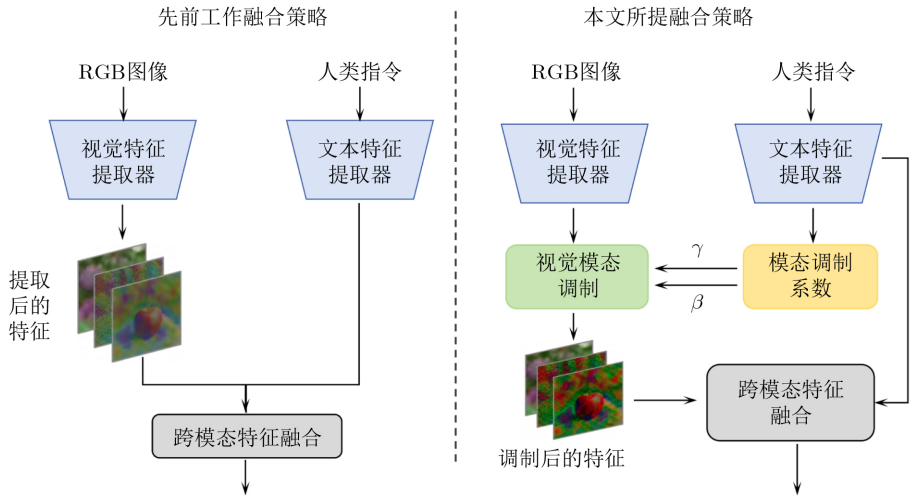

图 2 融合策略对比图

基于上述异构特征，网络进一步引入多头交叉注意力机制。该机制使得图像中的每一个扁平化像素能够独立地向语言序列发起查询，计算其与指令中各个局部词汇(如目标名词、修饰属性)的相关性权重，并以此聚合对应的文本语义，该过程表达式为：

$$\boldsymbol{F}_{\text{attn}} = \text{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \tag{4}$$

最后，为防止跨模态信息注入过程破坏原始的视觉空间细节，本文引入了残差连接与层归一化。聚合后的特征序列被重新折叠并转置，严谨地还原至二维空间拓扑结构，以便供后续解码器使用：

$$\begin{aligned}&\boldsymbol{F}_{\text{aligned}} = \text{Reshape}(\text{LayerNorm}(\boldsymbol{Q} + \boldsymbol{F}_{\text{attn}}))\\&\in \mathbb{R}^{1\times 100\times 10\times 10}\end{aligned} \tag{5}$$

通过这种密集的交叉注意力机制，网络能够在杂乱的物理空间中精准锚定与局部语言指令高度匹配的像素区域，实现了视图不变的细粒度语义定位。

2.5 解码部分

该部分基于跨模态全局上下文调制模块和词-像素跨模态交叉对齐模块得到的多模态特征，实现对指令物体的定位和抓取位姿的预测。为生成高分辨率的共享特征图以供后续多任务头部使用，本文设计了特征精炼与级联上采样模块。该模块首先利用带有残差结构的卷积网络对深层特征进行通道降维与语义聚合，随后基于连续的多层双线性插值实现对特征图空间分辨率的提升，该过程表达式为：

$$\begin{aligned}&\boldsymbol{F}_{\text{mid}} = \text{Conv2D}(\boldsymbol{F}_{\text{aligned}}) + \boldsymbol{F}_{\text{aligned}}\\&\boldsymbol{F}_{\text{s}} = \mathcal{U}_{\text{b}}^{\times 3}(\boldsymbol{F}_{\text{mid}})\end{aligned} \tag{6}$$

式中，$\boldsymbol{F}_{\text{aligned}}$表示输入的深层特征；Conv2D$(\cdot)$表示用于通道降维的常规卷积块。$\boldsymbol{F}_{\text{mid}}$为经过通道压缩后的中间特征，$\mathcal{U}_{\text{b}}^{\times 3}$表示连续三次缩放因子为2的双线性插值上采样操作。$\boldsymbol{F}_{\text{s}}$即为最终输出的共享特征。

关于物体定位解码预测部分，本文首先利用连续的两个二维卷积层对深层视觉特征进行通道降维与信息聚合，该过程表达式为：

$$\boldsymbol{F}_{\text{loc}} = \text{Conv2D}(\text{Conv2D}(\boldsymbol{F}_{\text{s}})) \tag{7}$$

式中，$\boldsymbol{F}_{\text{loc}} \in \mathbb{R}^{L\times 512}$表示提取的深层视觉特征。随后，为了从该空间特征中回归出准确的坐标参数，本文先将二维掩码特征在空间维度上进行展平处理，再将其送入多层MLP进行解码预测，最后通过激活函数将输出坐标归一化，该过程表达式为：

$$\hat{\boldsymbol{P}} = \sigma(\text{MLPs}(\text{Flatten}(\boldsymbol{F}_{\text{loc}}))) \tag{8}$$

式中，Flatten$(\cdot)$表示将二维特征重塑为一维向量的操作，$\hat{\boldsymbol{P}} \in \mathbb{R}^{1\times 4}$表示物体在图中的预测位置，$\sigma(\cdot)$为深度学习中常用的Sigmoid激活函数，用于将预测坐标值约束在(0,1)区间内。

关于抓取检测姿态预测部分，本文沿用文献[17]像素级别预测方案，该过程表达为：

$$\begin{aligned}&\{\boldsymbol{Q}, \boldsymbol{W}\} = \{\sigma(\text{Conv}(\boldsymbol{F}_{\text{s}})), \sigma(\text{Conv}(\boldsymbol{F}_{\text{s}}))\}\\&\{\cos(2\theta), \sin(2\theta)\} = \{\text{Conv}(\boldsymbol{F}_{\text{s}}), \text{Conv}(\boldsymbol{F}_{\text{s}})\}\end{aligned} \tag{9}$$

式中，$\boldsymbol{Q}$和$\boldsymbol{W}$分别表示预测抓取质量和宽度。$\cos(2\theta)$和$\sin(2\theta)$分别表示预测的抓取角度余弦、正弦数值，以此解决平行夹爪对称性带来的“角度不连续”问题[21,22]。因此，抓取角度的计算表达式为：

$$\theta = \frac{1}{2}\arctan\frac{\sin(2\theta)}{\cos(2\theta)} \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \tag{10}$$

2.6 训练推理

为减少训练过程产生的模块级联误差，本研究采用端到端的训练范式。该过程的损失函数包括物体定位损失L_{o}和抓取检测损失L_{c}，其表达式为：

$$L = \lambda * L_{\text{c}} + L_{\text{o}} \tag{11}$$

式中，λ为平衡损失函数引入的平衡因子，根据实验结果设置为2。其中，物体定位损失用于衡量网络预测的定位和真实结果之间的空间偏差，其表达式为：

$$L_{\text{o}} = \sum_{i=1}^{M} L_1(\hat{\boldsymbol{P}}^i, \boldsymbol{P}^i) + L_{\text{g}}(\hat{\boldsymbol{P}}^i, \boldsymbol{P}^i) \tag{12}$$

式中，L_1表示为坐标回归损失，L_{g}表示为使用GIoU[23]衡量的重叠度损失，M表示为样本的数量。至于抓取检测损失，本文使用平滑L_1交叉损失来优化，其表达式为：

$$\begin{aligned}&L_{\text{o}} = \frac{1}{M}\\&\sum_{n}^{N}\sum_{i=1}^{M}\begin{cases}0.5(\boldsymbol{G}_i^n - \hat{\boldsymbol{G}}_i^n)^2, & \text{if } \left|\boldsymbol{G}_i^n - \hat{\boldsymbol{G}}_i^n\right| < 1\\ \left|\boldsymbol{G}_i^n - \hat{\boldsymbol{G}}_i^n\right| - 0.5, & \text{otherwise}\end{cases}\end{aligned} \tag{13}$$

式中，$\boldsymbol{N} \in \{\boldsymbol{Q}, \boldsymbol{W}, \cos(2\theta), \sin(2\theta)\}$，$\hat{\boldsymbol{G}}$和$\boldsymbol{G}$分别表示为真实和预测的抓取位姿。在进行推理的时候，网络直接预测最优的物体的位置和抓取位姿，并联合标定矩阵将位姿转换到机器人坐标系，并执行抓取。

3 实验

3.1 实验设置

数据集：为了充分验证所提网络的有效性，本研究在Liu等人[15]的基础上采用场景-视角互斥划分策略重新划分跨场景跨视角的数据集。数据集合包含31种可抓取目标物体，指令集合与Liu等人[15]的

保持一致。此外，数据集共包含2个关键视角案例(Bottom和Top)，6个场景，且其背景与视角的场景组合均未在训练集合中出现，相应统计数据如表1所示。其中，案例1的训练集合样本数目为35 913，测试集合样本数目为8 092，相机视角为底部；而案例2的训练集合样本数目为37 606，测试集合样本数目为9 084，相机视角为顶部。训练和测试集合样本示例如图3所示，值得注意的是，案例1场景复杂度高于案例2。

实验参数设置：本文实验所用图像均为RGB格式，并统一裁剪至320×320分辨率。所提网络基于PyTorch框架，在单张NVIDIA RTX 3090 GPU上实现。网络训练采用AdamW优化器，初始学习率设为1e-3，样本训练批次大小为16，总训练轮数设置为25。

评估指标：对于物体定位结果评估，本文采用三种标准指标来验证网络性能。(1)精度P@X，用于计算预测为正样本且与真实标签匹配的实例占所有预测实例的比例，X表示不同的IoU阈值，在本实验中，X值空间设置为{0.5，0.7}；(2)平均交并比(Mean Intersection-over-Union, MIoU)，用于计算所有预测结果与其对应真实标签之间交并比的算术平均值；(3)整体交并比(Overall Intersection-over-Union, OIoU)，计算所有目标的累积交集面积总和与累积并集面积总和的比值。而对于抓取位姿预测的评估，本研究采用抓取准确率[15,24]来衡量网络性能。该指标要求预测结果同时满足预测抓取角度小于阈值ϕ且雅可比指数[26]大于φ。在本实验中，ϕ值空间设置为{10°，20°，30°}，φ值空间设置为{0.25，0.3，0.4}。值得注意的是，当$\phi=30°$且$\varphi=0.25$，此时抓取准确率被设置为标准抓取准确率(Grasp Accuracy, GA)。

表 1　数据集介绍

场景案例	训练集合	测试集合	测试样本视角
案例1	35 913	8 092	底部(Bottom)
案例2	37 606	9 084	顶部(Top)

3.2　与主流算法对比实验结果

表2和表3展示了本文所提网络SGL-Net与当前主流算法在案例1和案例2数据集上的对比结果。从表中可以看出，(1)SGL-Net在同步抓取检测与物体定位的各项指标上均取得了最优表现。如在案例

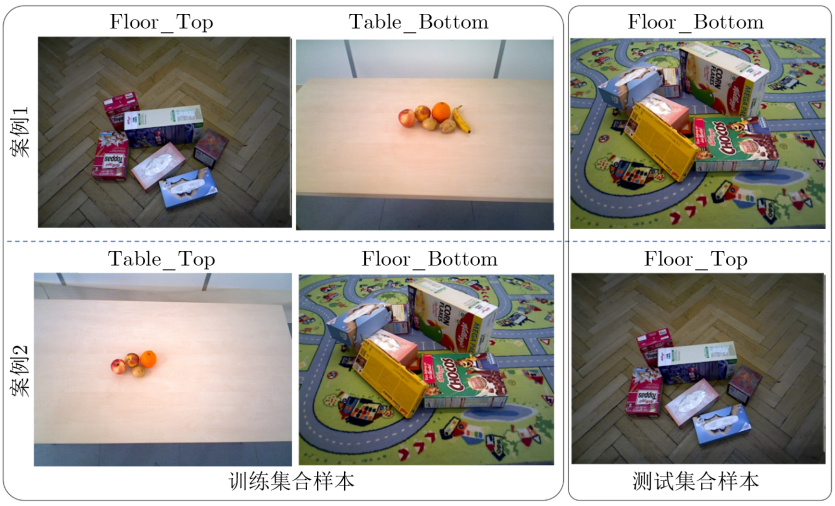

图 3 数据集可视化案例

表 2　本文算法及对比算法在案例1测试集合上的实验结果(%)

算法	抓取检测精度					物体定位精度			
	$30^\circ_{0.3}$	$30^\circ_{0.4}$	0.25_{10°	0.25_{20°	GA($30^\circ_{0.25}$)	P@0.5	P@0.7	MIoU	OIoU
GRCNN[24]	16.34	10.74	11.05	15.31	18.35	0.68	0	5.47	6.04
GGCNN[25]	3.62	1.45	2.14	4.34	5.14	1.69	0.27	5.62	6.19
GGCNN2[26]	20.74	11.70	10.52	18.24	23.22	0.87	0.14	5.66	6.62
CLIPort[14]	68.56	58.69	56.43	66.46	70.09	41.09	11.30	39.67	40.21
ELGNet[15]	65.67	55.72	52.46	63.37	67.45	15.89	6.52	31.21	30.82
本文算法	**73.23**	**61.58**	**59.82**	**71.30**	**74.63**	**56.07**	**21.77**	**47.55**	**45.23**

1设置下，SGL-Net在抓取检测表现上超过CLIPort模型4.54%，在物体定位精度MIoU上，超过CLIPort模型7.88%；(2)以GRCNN和GGCNN为代表的传统卷积神经网络模型在处理当前复杂场景时，各项性能指标均处于较低水平。相比之下，基于CLIP的方法凭借其丰富的图文预训练知识，为异构模态间的语义对齐奠定了坚实基础，能够更有效地捕捉全局特征依赖与空间几何关系，从而显著提升了模型的多模态理解能力；(3)由于案例1的场景背景杂乱程度要比案例2的难度更高，对模型的感知与抗干扰能力提出了更为严苛的挑战，但本文网络不仅在整体精度上保持最优，而且在极具挑战性的案例1中依然展现出了远超对比算法的稳定性，这进一步验证了本文算法在应对分布外场景时的组合泛化能力。

3.3 所提网络模块消融实验结果

为验证所提网络中CGCMM与WPCAM模块的有效性，本文设计了逐模块移除式消融实验，在案例1和案例2上的对比结果分别如表4和表5所示。实验结果表明，移除WPCAM会使网络丧失细粒度特征语义对齐能力，导致其性能显著下降(如案例1中P@0.5指标数值由56.07%降至45.71%)。当进一步移除CGCMM时，受场景语义解耦能力缺失的影响，模型表现继续下降(如案例2中MIoU指标数值从55.88%跌至47.33%)。这些结果表明，各个模块的设计均对整体有正向作用。此外，也可以发现案例1上消融泛化实验效果弱于案例2，这与数据集本身的难度设置相一致。但整体而言，所提网络仍能维持较为稳定的泛化表现，进一步印证了其良好的泛化能力。

3.4 可视化对比实验结果

为了更加直观展示所提网络性能，本文进一步开展了可视化实验，在案例1和案例2上与当前主流算法的可视化对比结果如图4所示。

从图中可以发现，本文所提的SGL-Net能够准确地根据指令定位物体，并生成合理的抓取位姿。而如实验图5中的结果，GRCNN和GGCNN2模型不仅难以定位物体，而且无法给出准确的抓取位姿。这也表明在当前实验场景下，仅仅依靠卷积网络难以实现视觉和文本两种模态的对齐。相比较之下，基于CLIP的架构的CLIPort、ELGNet及SGL-Net能够利用丰富的预训练知识显著提升视觉和文本的模态对齐能力。然而，从图中也可以发现，由于难以建立局部视觉区域与具体语言指令片段之间的细粒度语义关联，CLIPort和ELGNet也出现了定位错误和抓取不准的问题。如图中第一行的ELGNet模型，给定的目标物体是“Pear”，但其给出的物体定位却是“Orange”。此外，图中第四行的CLIPort给出“Ball”的抓取位姿无法执行，并非最优位姿。相比较之下，SGL-Net不仅可以准确的定位指令物体，而且可以给出适合的抓取位姿，进一步验证了所提网络的有效性。

为了探究网络在与测试案例相同视角且位于相似背景时的泛化能力，本实验针对案例1(蓝色线上方)和案例2(蓝色线下方)额外引入了部分测试样本。值得注意的是，测试样本仅仅是背景和训练样本类似，但其视角仍保持为Bottom和Top。从图中可以发现，所对比的算法由于无法充分将背景和物体进行解耦，难以充分理解语义和视觉信息，产生了一定的“语义对齐偏移”与“特征混淆”，导致

表 3 本文算法及对比算法在案例2测试集合上的实验结果(%)

算法	抓取检测精度					物体定位精度			
	$30^{\circ}_{0.3}$	$30^{\circ}_{0.4}$	$0.25_{10^{\circ}}$	$0.25_{20^{\circ}}$	GA($30^{\circ}_{0.25}$)	P@0.5	P@0.7	MIoU	OIoU
GRCNN[24]	16.29	12.88	10.76	13.78	17.73	0.85	0.31	5.72	5.87
GGCNN[25]	2.20	1.06	1.19	2.50	2.96	1.23	0.00	6.72	7.28
GGCNN2[26]	24.04	20.15	17.81	22.62	25.28	1.01	0.20	6.28	7.11
CLIPort[14]	73.80	70.15	60.34	71.04	74.31	71.47	39.02	57.37	54.21
ELGNet[15]	73.08	65.20	59.0	69.97	74.21	70.66	33.40	55.21	50.63
本文算法	**76.57**	**71.75**	**63.72**	**73.49**	**77.34**	**76.20**	**45.79**	**59.73**	**56.66**

表 4 案例1消融实验结果

CGCMM	WPCAM	GA	P@0.5	MIoU	OIoU
√	√	74.63	56.07	47.55	45.23
√		71.18	45.71	41.28	42.11
		69.64	41.40	40.29	37.92

表 5 案例2消融实验结果

CGCMM	WPCAM	GA	P@0.5	MIoU	OIoU
√	√	77.34	76.20	59.73	56.66
√		74.08	68.97	55.88	53.07
		71.95	60.34	47.33	45.42

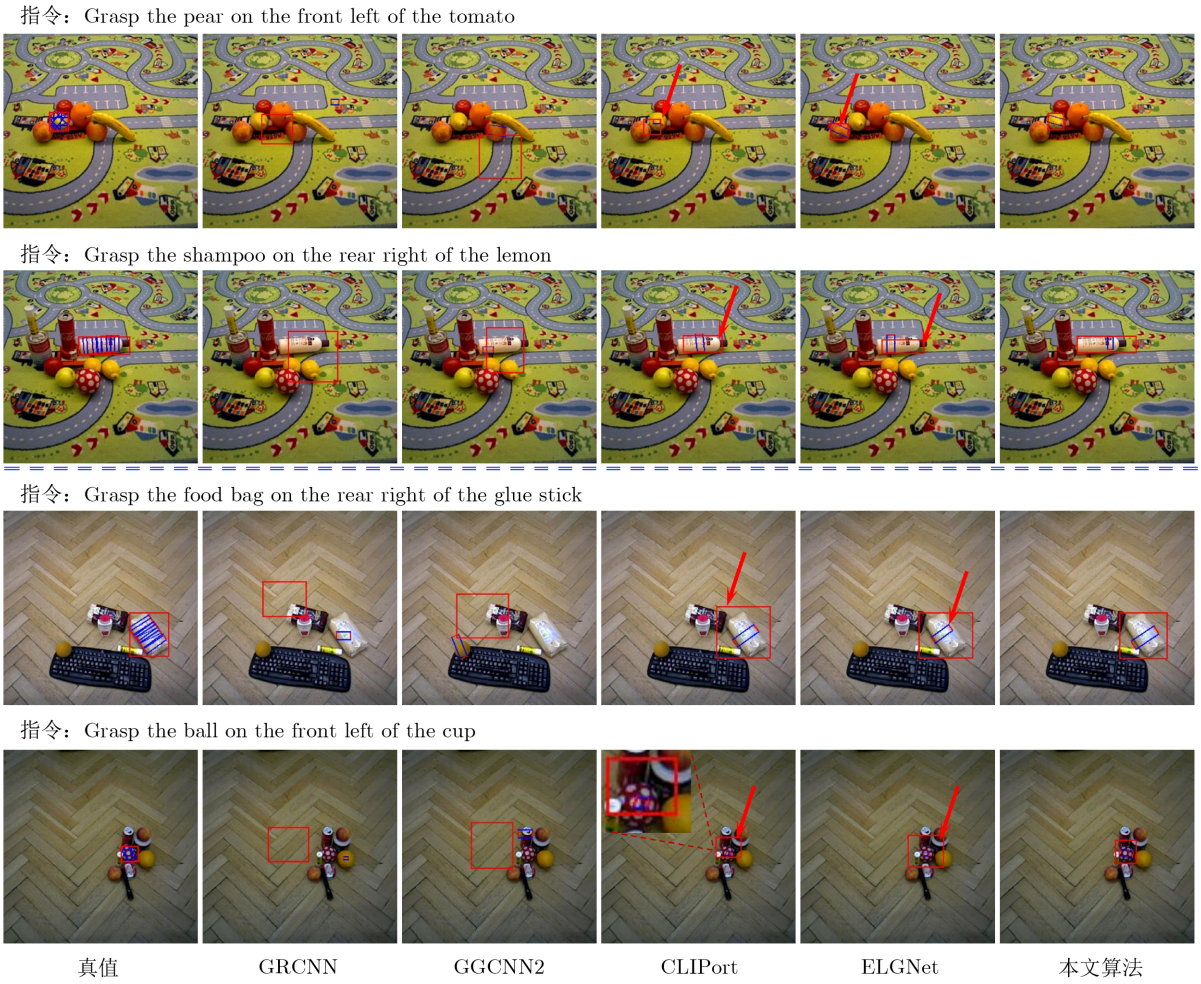

图 4 不同案例抓取检测及定位可视化结果。箭头所示为不精确预测，红色框为物体预测位置，蓝色线上方为案例1，下方为案例2。

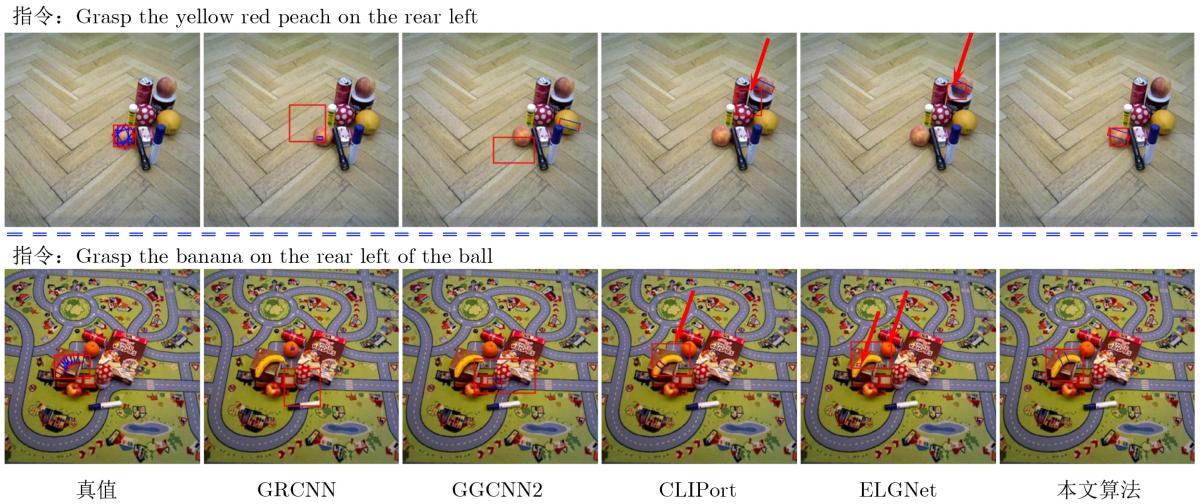

图 5 位于相似背景时抓取检测及定位可视化结果。箭头所示为不精确预测，红色框为物体预测位置，蓝色线上方为案例1，下方为案例2。

其定位就出现了一定的偏差，而且预测的抓取位姿也不足以支撑抓取，这进一步表明本文所提SGL-Net的有效性。

分布外场景的泛化性能是确保网络在实际场景中应用的重要能力，为此，本实验进一步从Liu等人的数据集中提取视角与案例1和案例2中测试样本相同，但背景不同的样本进行可视化直观验证。并与表现最优的CLIPort模型进行对比，结果如图6所示。从图中可以发现，相比于CLIPort模型，SGL-Net不仅能够准确定位物体，而且可以输出最

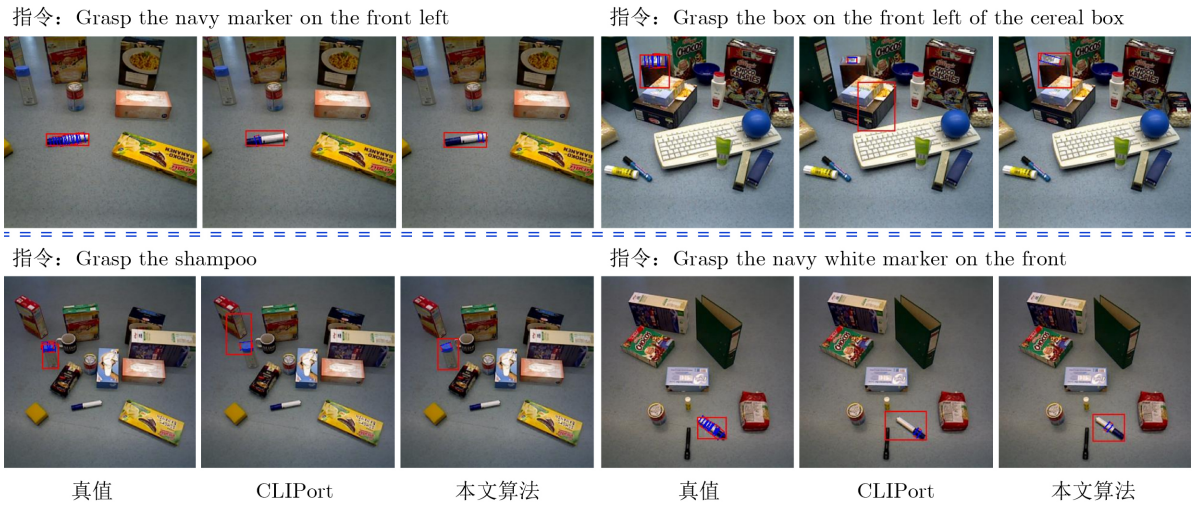

图 6 其他不同背景下的泛化实验可视化结果。红色框为预测的物体位置，蓝色线上方为案例1，下方为案例2。

优的抓取位姿。此外，即便CLIPort模型对物体提供较为准确的抓取位姿，但是对物体的定位却并不准确，甚至出现一定的偏差。这可能是由于其虽然采用分层特征残差融合结构，但对于场景的理解和模态的解耦能力仍存在较大不足，导致其性能仍差于本文所提的SGL-Net。

3.5 真实实验结果

为验证SGL-Net在真实场景中的抓取能力，本文基于自主搭建的机器人抓取平台开展实物指令抓取验证实验。该平台硬件系统由AuBo i5协作机械臂、Robotiq140夹爪及RealsenseD435相机构成。与CLIPort和ELGNet的抓取结果如表6所示。可以发现本文所提算法相比于先前算法能够取得更为稳定的抓取性能。此外，不同视角下的抓取结果如图7所示。实验结果表明，机械臂能够依据自然语言操作指令，在不同视角约束下精准完成目标物体的抓取任务，充分验证了本文所提网络在真实场景中具备可靠的落地应用能力与场景泛化性。

表 6 机器人抓取实验

算法	Top视角	Bottom视角
CLIPort	70%(35/50)	80%(40/50)
ELGNet	72%(36/50)	74%(37/50)
本文算法	84%(42/50)	92%(46/50)

4 结论

本文针对机器人抓取操作过程中变视角跨域泛化能力不足的难题，构建了相应的验证数据集和同步抓取检测和定位网络。通过跨模态全局上下文调制策略和词-像素跨模态交叉对齐策略，实现了对指令要求物体的准确定位和适合抓取位姿的预测。多种验证案例场景数据集和真实世界的抓取实验的结果表明，本文所提网络表现优异，具备在实际场景中应用部署的潜力。为了进一步增强零样本场景下的机器人抓取能力，未来的工作将聚焦于引入多模态大模型，并通过微调等技术手段实现网络适配。

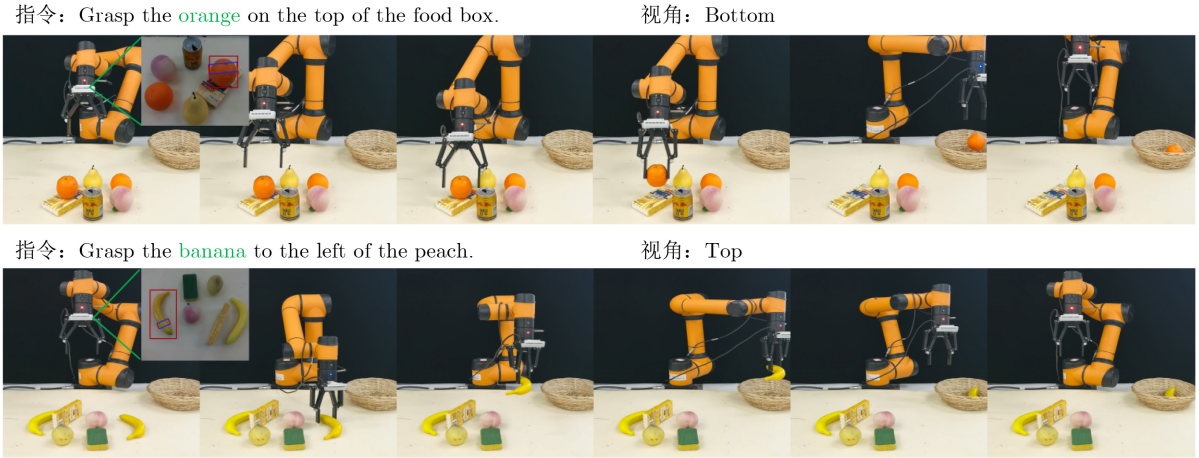

图 7 真机抓取实验可视化结果。红色框为预测的物体位置，蓝红框为预测的抓取位姿

参 考 文 献

[1] 张春云, 孟昕瞳, 陶陶, 等. 面向机器人螺栓装配的视觉感知与力控协同方法[J]. 电子与信息学报, 2026, 48(5): 2053–2065. doi: 10.11999/JEIT251193.
ZHANG Chunyun, MENG Xintong, TAO Tao, *et al*. Vision-guided and force-controlled method for robotic screw assembly[J]. *Journal of Electronics & Information Technology*, 2026, 48(5): 2053–2065. doi: 10.11999/JEIT251193.

[2] 余浩扬, 李艳生, 肖凌励, 等. 面向动态环境的巡检机器人轻量级语义视觉SLAM框架[J]. 电子与信息学报, 2025, 47(10): 3979–3992. doi: 10.11999/JEIT250301.
YU Haoyang, LI Yansheng, XIAO Lingli, *et al*. A lightweight semantic visual simultaneous localization and mapping framework for inspection robots in dynamic environments[J]. *Journal of Electronics & Information Technology*, 2025, 47(10): 3979–3992. doi: 10.11999/JEIT250301.

[3] LI Renjie, DONG Wei, SUN Jiarui, *et al*. A fast integrated gait, footstep, and motion planning framework for wheeled-legged robots[J]. *Journal of Bionic Engineering*, 2026, 23(2): 607–621. doi: 10.1007/s42235-025-00830-5.

[4] LIU Zhitai, ZHONG Hanhai, LIN Xiaotian, *et al*. Integrated motion control framework of wheel-legged biped robot on rugged terrain[J]. *IEEE/ASME Transactions on Mechatronics*, 2025, 30(6): 7383–7394. doi: 10.1109/TMECH.2025.3627438.

[5] 崔永成, 田国会, 周昭旭, 等. 智能空间下面向动作序列生成的服务机器人指令解析方法[J]. 机器人, 2024, 46(1): 1–15. doi: 10.13973/j.cnki.robot.230074.
CUI Yongcheng, TIAN Guohui, ZHOU Zhaoxu, *et al*. A service robot instruction parsing method for action sequence generation in intelligent space[J]. *Robot*, 2024, 46(1): 1–15. doi: 10.13973/j.cnki.robot.230074.

[6] LENZ I, LEE H, and SAXENA A. Deep learning for detecting robotic grasps[J]. *The International Journal of Robotics Research*, 2015, 34(4/5): 705–724. doi: 10.1177/0278364914549607.

[7] KUMRA S and KANAN C. Robotic grasp detection using deep convolutional neural networks[C]. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, Vancouver, Canada, 2017: 769–776. doi: 10.1109/IROS.2017.8202237.

[8] LAILI Yuanjun, CHEN Zelin, REN Lei, *et al*. Custom grasping: A region-based robotic grasping detection method in industrial cyber-physical systems[J]. *IEEE Transactions on Automation Science and Engineering*, 2023, 20(1): 88–100. doi: 10.1109/TASE.2021.3139610.

[9] CHU F J, XU Ruinian, and VELA P A. Real-world multiobject, multigrasp detection[J]. *IEEE Robotics and Automation Letters*, 2018, 3(4): 3355–3362. doi: 10.1109/LRA.2018.2852777.

[10] BOUSSELHAM W, PETERSEN F, FERRARI V, *et al*. Grounding everything: Emerging localization properties in vision-language transformers[C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 2024: 3828–3837. doi: 10.1109/CVPR52733.2024.00367.

[11] YIN Heng, REN Yuqiang, YAN Ke, *et al*. ROD-MLLM: Towards more reliable object detection in multimodal large language models[C]. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, USA, 2025: 14358–14368. doi: 10.1109/CVPR52734.2025.01339.

[12] ZHOU Yi, SU Hang, WANG Tian, *et al*. Onet: Twin U-Net architecture for unsupervised binary semantic segmentation in radar and remote sensing images[J]. *IEEE Transactions on Image Processing*, 2025, 34: 2161–2172. doi: 10.1109/TIP.2025.3530816.

[13] WANG Hao, HU Keyan, GUO Xin, *et al*. A gift from the integration of discriminative and diffusion-based generative learning: Boundary refinement remote sensing semantic segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026, 48(5): 5892–5909. doi: 10.1109/TPAMI.2026.3654243.

[14] SHRIDHAR M, MANUELLI L, and FOX D. CLIPort: What and where pathways for robotic manipulation[C]. Proceedings of the 5th Conference on Robot Learning, London, UK, 2021: 894–906.

[15] LIU Jin, XIE Jialong, and XIAO Leibing. Hierarchical multi-modal fusion for language-conditioned robotic grasping detection in clutter[J]. *IEEE Robotics and Automation Letters*, 2024, 9(10): 8762–8769. doi: 10.1109/LRA.2024.3440833.

[16] VAN VO T, VU M N, HUANG Baoru, *et al*. Language-driven grasp detection with mask-guided attention[C]. 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems, Abu Dhabi, UAE, 2024: 7492–7498. doi: 10.1109/IROS58592.2024.10802256.

[17] XIE Jialong, LIU Jin, ZHU Zhenwei, *et al*. Infusing multisource heterogeneous knowledge for language-conditioned segmentation and grasping[J]. *IEEE Transactions on Instrumentation and Measurement*, 2024, 73: 5029611. doi: 10.1109/TIM.2024.3446625.

[18] XIE Jialong, ZHOU Fengyu, LIU Jin, *et al*. Semi-supervised language-conditioned grasping with curriculum-scheduled augmentation and geometric consistency[J]. *IEEE Robotics and Automation Letters*, 2025, 10(4): 4021–4028. doi: 10.

1109/LRA.2025.3547619.

[19] 张梅, 金叶, 朱金辉, 等. 特征级语义感知引导的多模态图像融合算法[J]. 电子与信息学报, 2025, 47(8): 2909–2918. doi: 10.11999/JEIT250042.
ZHANG Mei, JIN Ye, ZHU Jinhui, *et al*. FSG: Feature-level semantic-aware guidance for multi-modal image fusion algorithm[J]. *Journal of Electronics & Information Technology*, 2025, 47(8): 2909–2918. doi: 10.11999/JEIT250042.

[20] LI Jindong, LI Yongguang, FU Yali, *et al*. CLIP-powered domain generalization and domain adaptation: A comprehensive survey[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026, 48(5): 5405–5424. doi: 10.1109/TPAMI.2026.3651700.

[21] ZHOU Zhenning, ZHU Xiaoxiao, and CAO Qixin. AAGDN: Attention-augmented grasp detection network based on coordinate attention and effective feature fusion method[J]. *IEEE Robotics and Automation Letters*, 2023, 8(6): 3462–3469. doi: 10.1109/LRA.2023.3268596.

[22] WANG Dexin, LIU Chunsheng, CHANG Faliang, *et al*. High-performance pixel-level grasp detection based on adaptive grasping and grasp-aware network[J]. *IEEE Transactions on Industrial Electronics*, 2021, 69(11): 11611–11621. doi: 10.1109/TIE.2021.3120474.

[23] REZATOFIGHI H, TSOI N, GWAK J, *et al*. Generalized intersection over union: A metric and a loss for bounding box regression[C]. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 658–666. doi: 10.1109/CVPR.2019.00075.

[24] KUMRA S, JOSHI S, and SAHIN F. Antipodal robotic grasping using generative residual convolutional neural network[C]. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems, Las Vegas, USA, 2020: 9626–9633. doi: 10.1109/IROS45743.2020.9340777.

[25] MORRISON D, CORKE P, and LEITNER J. Learning robust, real-time, reactive robotic grasping[J]. *The International Journal of Robotics Research*, 2020, 39(2/3): 183–201. doi: 10.1177/0278364919859066.

[26] XIE Jialong, LIU Jin, HUANG Saike, *et al*. Listen, perceive, grasp: CLIP-driven attribute-aware network for language-conditioned visual segmentation and grasping[J]. *IEEE Transactions on Automation Science and Engineering*, 2025, 22: 9729–9740. doi: 10.1109/TASE.2024.3510777.

刘 进：男，副教授，研究方向为机器人具身操作.
刘志太：男，副研究员，研究方向为机器人智能系统及智能显微操作.
李梓函：女，无，研究生，研究方向为六自由度物体姿态估计.
孙彦景：男，教授，研究方向为智能安全与应急通信.
缪燕子：女，教授，研究方向为机器视觉与自动驾驶.
袁宪锋：男，副教授，研究方向为智能机器人、具身智能及视觉语言导航.

责任编辑：余 蓉

Fusing Global Perspective Rectification and Fine-Grained Semantic Decoupling for Language-Conditioned Robotic Grasping Detection

LIU Jin[①] LIU Zhitai[②] LI Zihan[③] SUN Yanjing[①③]
MIAO Yanzi[③] YUAN Xianfeng[④]

[①](*School of Computer Science and Technology/School of Artificial Intelligence, China University of Mining and Technology, Xuzhou* 221116, *China*)

[②](*School of Astronautics, Harbin Institute of Technology, Harbin*, 150001, *China*)

[③](*School of Information and Control Engineering, China University of Mining and Technology, Xuzhou* 221116, *China*)

[④](*School of Airspace Science and Engineering, Shandong University, Weihai*, 264209, *China*)

Abstract:

Objective Accurate grasping of target objects from language instructions is an essential prerequisite for service robots to achieve natural human-robot interaction. Existing research primarily relies on massive data-driven training or hierarchical feature fusion schemes for the cross-modal alignment between visual perception and textual instructions. However, these methods commonly overlook the heavy entanglement between target objects and background environments within low-level features, leading to a significant degradation in compositional generalization ability under cross-view and unseen background scenarios. To address these issues, this paper constructs a dual-view cross-scenario synchronous grasp detection and localization dataset, which is

used to systematically evaluate and specifically improve the compositional generalization ability of existing models. Building upon this benchmark, we propose a simultaneous detection and localization network to simultaneously output the target location and the optimal grasping pose. Consequently, service robots equipped with the proposed network can perform object manipulation in real-world dynamic environments according to human instructions, providing theoretical foundations and technical support for embodied intelligence.

Methods The structure of the proposed Simultaneous Grasp and Localization Network (SGL-Net) is illustrated (Fig. 1). Firstly, a cross-modal global context modulation module (CGCMM) is proposed. The module utilizes instruction priors to perform adaptive viewpoint correction and background suppression on visual channels during the early stages of feature extraction. Secondly, a word-pixel cross-modal alignment module (WPCAM) is designed to achieve fine-grained semantic decoupling via a flattened attention mechanism, further enhancing the model's comprehension in dynamic scenes. Finally, a unified decoder is employed to simultaneously output the target location and the optimal grasping pose.

Results and Discussions Extensive quantitative and qualitative experiments are conducted on a restructured dual-view dataset, including bottom view and top view, and a real-world physical robotic platform. Comparative results demonstrate that SGL-Net significantly outperforms mainstream CNN-based and CLIP-based baseline methods in both grasp detection and object localization metrics (Table 2 and Table 3). Ablation studies explicitly validate the critical contributions of two designed modules in achieving fine-grained semantic alignment and scene decoupling (Table 4 and Table 5). Furthermore, qualitative visualizations (Fig. 4, Fig. 5, and Fig. 6) and real-world physical experiments (Fig. 7) indicate that SGL-Net exhibits strong viability for direct deployment in real-world physical environments. Overall, the network exhibits superior robust generalization capabilities and reliable practical deployment potential when confronting complex physical environments.

Conclusions To address the challenge of insufficient cross-view and cross-domain generalization capability in robotic grasping operations, this paper constructs a corresponding validation dataset and proposes a simultaneous grasp detection and object localization network. By integrating a cross-modal global context modulation module and a word-pixel cross-modal alignment module, the proposed network achieves accurate localization of instruction-specified objects and prediction of optimal grasp poses. Experimental results from diverse testing datasets and real-world grasping experiments demonstrate the superior performance of the proposed network, highlighting its potential for deployment in practical scenarios. To further enhance robotic grasping capabilities in zero-shot scenarios, future work will focus on integrating Large Multimodal Models and realizing network adaptation through fine-tuning techniques.

Key words: Grasp detection; Language-conditioned; Multi-modal fusion; Complex scenarios