智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

弱监督下语言引导的图像分割与定位综述

张磊, 黄咏秋, 李欣, 王宝艳

引用本文:
张磊, 黄咏秋, 李欣, 等. 弱监督下语言引导的图像分割与定位综述[J]. 智能系统学报, 2025, 20(6): 1304-1327.
ZHANG Lei, HUANG Yongqiu, LI Xin, et al. Review of weakly supervised language-guided image segmentation and grounding[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(6): 1304-1327.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202505001

DOI: 10.11992/tis.202505001
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20251021.1311.002

弱监督下语言引导的图像分割与定位综述

张磊[1], 黄咏秋[2], 李欣[2], 王宝艳[2]
(1. 广东石油化工学院 电子信息工程学院, 广东 茂名 525000; 2. 广东石油化工学院 计算机学院, 广东 茂名 525000)

摘 要: 语言引导的图像分割(referring image segmentation, RIS)与定位(referring expression grounding, REG)旨在根据自然语言指令预测目标的掩码或边界框, 是视觉−语言理解的重要任务。完全监督方法因标注成本高受限, 促使弱监督学习成为研究热点。对此, 从统一视角梳理弱监督 RIS 与 REG 研究进展, 重点介绍仅依赖图像−文本对及无标注数据的方法, 并探讨现存问题与未来方向。介绍 RIS 与 REG 任务背景, 分析弱监督学习的价值与挑战; 归纳不同类型的弱监督信号, 分类综述代表性方法并分析其特点; 介绍主流数据集与评价指标, 并比较典型方法性能。研究表明, 引入多模态大语言模型等预训练模型可显著提升性能, 但仍受限于预训练模型的局限性与任务适配性。未来, 优化跨模态细粒度对齐、模型效率与泛化能力将是该领域的重要研究方向。

关键词: 深度学习; 计算机视觉; 弱监督学习; 无监督学习; 指代图像分割; 指代表达定位; 多模态; 大语言模型
中图分类号: TP391 **文献标志码**: A **文章编号**: 1673−4785(2025)06−1304−24

中文引用格式: 张磊, 黄咏秋, 李欣, 等. 弱监督下语言引导的图像分割与定位综述 [J]. 智能系统学报, 2025, 20(6): 1304–1327.
英文引用格式: ZHANG Lei, HUANG Yongqiu, LI Xin, et al. Review of weakly supervised language-guided image segmentation and grounding[J]. CAAI transactions on intelligent systems, 2025, 20(6): 1304–1327.

Review of weakly supervised language-guided image segmentation and grounding

ZHANG Lei[1], HUANG Yongqiu[2], LI Xin[2], WANG Baoyan[2]
(1. School of Electronic Information Engineering, Guangdong University of Petrochemical Technology, Maoming 525000, China; 2. School of Computer Science, Guangdong University of Petrochemical Technology, Maoming 525000, China)

Abstract: Language-guided image segmentation (referring image segmentation, RIS) and grounding (referring expression grounding, REG) aim to predict masks or bounding boxes for target objects based on natural language instructions, serving as key tasks in vision-language understanding. Fully supervised methods are constrained by high annotation costs, driving increasing interest in weakly supervised learning. This paper reviewed recent advances in weakly supervised RIS and REG from a unified perspective, focused on methods based on image-text pairs and unlabeled data, and discussed current challenges and future directions. It introduced the background of RIS and REG and analyzed the value and challenges of weak supervision. It summarized different types of weak supervision signals, categorized representative methods, and analyzed their characteristics. It presented mainstream datasets and evaluation metrics, and compared the performance of typical methods. Studies showed that incorporating pretrained models, such as large language models, can significantly improve performance. However, limitations due to the constraints of pretrained models and task adaptation remain. In the future, optimizing fine-grained cross-modal alignment, model efficiency, and generalization ability will be important research directions.

Keywords: deep learning; computer vision; weakly supervised learning; unsupervised learning; referring image segmentation; referring expression grounding; multimodal; large language model

语言引导的图像分割[1](referring image segmentation, RIS)和定位[2-3](referring expression grounding, REG)均属于计算机视觉与自然语言处

收稿日期: 2025−05−06. **网络出版日期**: 2025−10−21.
基金项目: 国家自然科学基金项目 (62476064); 广东省自然科学基金项目(2024A1515010455).
通信作者: 李欣. E-mail: lixin@gdupt.edu.cn.

理交叉领域的重要任务, 其核心目标是根据参考文本(referring expression)的指示在图像中分割及定位目标。其中, RIS 的目标是生成描述物体的像素级掩码, 而 REG 则以边界框的形式标定描述目标位置, 无需精确勾勒物体轮廓。相比传统的实例分割[4] 或目标检测[5] 任务, RIS 和 REG 不仅要求模型解析图像内容, 还需理解文本信息, 并在两者之间建立准确的映射关系, 包括识别目标对象的外观特征、空间位置及其与其他物体的关系等。因此, RIS 和 REG 在语言引导的人机交互[6]、视觉导航[7] 等任务中具有重要应用。

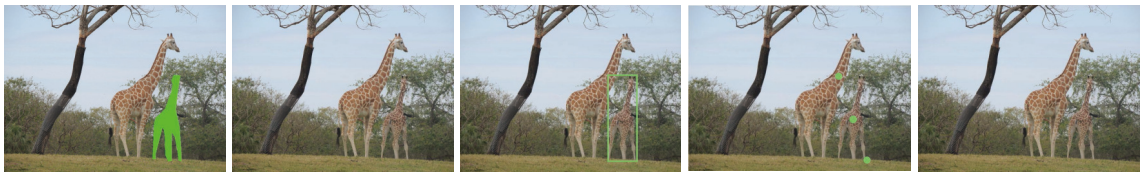
深度学习模型凭借其强大的学习与推理能力, 已成为解决 RIS 和 REG 任务的主流方法, 在这两项任务中都取得了较高的准确度[8-10]。当前, 大多数方法采用完全监督学习范式, 即利用像素级掩码或边界框作为监督信号进行训练。其优势在于, 明确的标注可直接指导模型优化, 使其逐步收敛至更准确的目标预测。通过对比模型预测与真实掩码或边界框, 模型的参数得以高效调整, 从而提升任务性能。然而, 这种方法高度依赖像素级或边界框级的精细标注, 尤其在大规模数据集上, 标注成本极为高昂。

近年来, 弱监督学习方法因其能够显著降低标注成本, 逐渐成为 RIS 和 REG 任务的研究热点。不同于完全监督学习依赖完整的“图像–目标掩码(或边界框)–文本”标注, 弱监督学习利用更少的监督信息进行训练, 例如仅使用“图像–文本”对, 而无需像素级掩码或边界框标注。相较于完整的监督信号, 这种方式仅提供图像与文本的匹配信息, 而不包含目标的精确空间位置。因此, 模型需要依赖视觉与语言的关联性, 推测目标的像素级或边界框级表达。此外, 弱监督学习还可以利用其他形式的监督信号, 如关键点标注甚至无标注数据, 辅助模型的学习。

目前, 已存在一些关于 RIS 及弱监督语义分割和目标检测的研究综述。例如, 北京交通大学信息科学研究所的邱爽团队[11] 围绕完全监督的 RIS 方法, 从多模态信息编码方式对算法进行分类; 南京邮电大学的项伟康团队[12] 介绍了基于图像级标注、其他弱标注及大模型辅助的弱监督语义分割, 并探讨了该领域的挑战及研究方向; 南京理工大学的陈震元团队[13] 针对基于图像级标注的弱监督目标检测, 从网络架构角度分类相关算法, 并在不同指标上评估其性能; 北京工业大学的李文生团队[14] 针对图像/视频任务, 从模型结构出发, 详细介绍了基于 Transformer 架构的视觉分割技术; 计算机软件新技术国家重点实验室的祁磊团队[15] 将无监督和半监督纳入弱监督场景中, 依据技术和场景类型, 分别对弱监督下的行人重识别算法进行了划分; 南京理工大学的蒋弘毅[16] 团队重点归纳了基于卷积神经网络的目标检测框架, 并根据模型不同模块的优化方法进行了归纳。这些研究重点聚焦于单模态任务, 深入分析和总结了相应领域的方法。而在多模态场景下, 针对弱监督 RIS 和 REG 的研究仍处于探索阶段, 系统性的综述工作较为缺乏, 限制了研究者对最新进展和核心挑战的快速理解。为此, 本文从统一视角出发, 系统梳理弱监督下 RIS 与 REG 的研究进展, 对各类方法进行详细介绍, 并分析不同方法的特点: 从任务共性出发, 分析 RIS 与 REG 在弱监督场景下的联系、挑战与潜在解决方案; 从监督信息的来源与形式出发, 系统归纳不同类型的弱监督信号; 综述典型方法与性能表现, 总结关键技术与现存问题, 并展望未来研究方向。

1　弱监督信号类型

基于现有方法, 结合 Shen 等[17] 的分类方式, 在 RIS 和 REG 任务中, 弱监督信号主要包括图像–文本对、边界框级信号(仅对 RIS)、关键点信号(仅对 RIS)、部分标注信号、无标注。这些不同的监督信号见图 1。



(a) 像素级掩码标注　(b) 图像–文本对标注　(c) 边界框标注　(d) 关键点标注　(e) 无标注

图 1　不同监督信号示意

Fig. 1　Illustration of different supervision signals

图像−文本对：如图 1(b) 所示，这类监督信号通常提供的是图像和对应的文本描述数据。监督信息仅表明输入的图像和文本是否为配对，而不包含图中对象更具体的信息。若文本描述与图像中的某个对象语义相关，则该文本被视为正类；反之，则文本被视为负类[18]。这种学习方式大幅降低了标注成本，但也对模型的学习能力提出了更高要求，尤其在面对复杂场景或长文本描述时，模型可能难以准确推断。

边界框信号：如图 1(c) 所示，在 RIS 任务中，边界框信号提供了目标对象的位置信息，相比完全监督的掩码信号，边界框信号仅提供较粗粒度的监督信息，要求模型在接收图像、参考文本以及目标对象的边界框作为监督信号的条件下，学习预测与参考文本匹配的目标掩码[19]。因此，模型需补足边界框信号中缺失的细节信息，得到更精细的像素级的掩码预测。在 REG 中，边界框信号属于完全监督学习信号。

关键点信号：如图 1(d) 所示，在 RIS 任务中，关键点信号提供目标对象的数个关键点，如对象中心点及其角点[20]。与边界框信号相比，关键点信号能更具体地反映对象的位置，但仍缺少完整的轮廓信息。因此，模型需充分利用这些关键点生成目标的像素级掩码预测。关键点信号能够在标注成本较低的情况下，作为像素级掩码信号的替代方案。

部分标注信号：在 RIS 任务中，部分样本具有完全监督信号，而其余样本仅具备不完全的监督信号或无监督信号。模型需结合这两种样本，通过有限的完整标注信息与标注不完全的数据相结合完成预测，以降低标注成本。其中，不完全的信号包括：1）图像−文本对[21]；2）仅图像−边界框−参考文本[22-23]；3）混合形式[24]，结合多种弱监督信号，例如，一部分对象仅具备边界框信号，而另一部分对象仅有关键点信号；4）仅图像[25-26]，无任何目标位置信息或文本描述，模型需间接利用其他具备完全监督信号的样本，进行区域预测。在 REG 任务中，训练数据不具备完整的监督信号，主要体现为完全监督信号（图像−边界框−参考文本）中参考文本的缺失，或图像中部分对象的边界框与参考文本信息的缺失。模型需基于已标注的信息完成学习，这些已标注信息可以是：1）图像−边界框[2]，缺少文本描述，模型仅基于目标的位置信息进行学习；2）图像−少量边界框−参考文本[27-28]，缺少图中一些对象的边界框与文本描述，模型需要利用已知监督信号，对未知目标的边界框进行预测。

无标注：如图 1(e) 所示，在这类学习方式中，模型通过未标注的图像数据或配对关系未知的图像−文本，来实现任务能力的建模。这种方式不依赖人工标注信号，能极大降低数据标注的成本。对于 RIS 任务，模型需在没有目标掩码及明确位置信息的情况下，从未标注图像中挖掘参考文本所对应目标的潜在语义区域；而在 REG 任务中，模型则需通过对应关系未知的图像−文本建模，推测目标对象的位置或边界。近年来，这类学习方式因其可扩展性和巨大的潜在应用价值，在 RIS 和 REG 任务中受到越来越多的关注。参考 Zhang 等[29] 的研究，部分监督信号类型的人工注释所需时间见表 1。

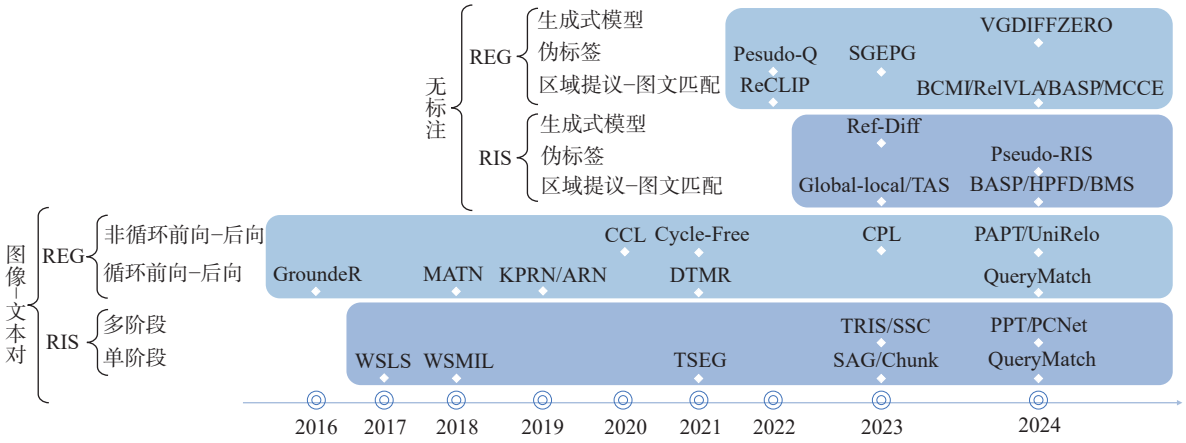
表 1　不同信号标注的平均时间成本

Table 1　Average time cost of different annotation signals

标注类型	标注内容	平均时间/s
类别标注	图像整体类别	1
边界框	实例对象的边界框	10
像素级轮廓标注	实例对象的像素轮廓	78

表 1 反映了不同标注所需的时间，类别标注是标注成本最低的形式，边界框次之，像素级轮廓在单个实例对象的标定平均需花费超过 1 min，而在 RIS 任务中，像素级标注需要与特定文本描述相对应，所需时间成本会更高。

在 RIS 和 REG 任务的弱监督学习研究中，图像−文本对及无标注数据的学习方式因其低标注成本和广泛适用性，成为研究的重点。相比之下，其他弱监督方式（如基于边界框或关键点标注）研究较少，通常针对特定任务或应用场景设计。基于这一趋势，本文重点围绕图像−文本对弱监督和基于无标注的弱监督这两类方法，系统梳理其在 RIS 和 REG 任务中的研究进展，并探讨未来的发展方向。图 2 从时间与类别角度展示了本文涉及的弱监督 RIS 与 REG 方法。依据模型的训练方式与特性，对不同弱监督信号下的 RIS 与 REG 方法进行了划分。



发表来源与方法对应关系, ECCV: GroundeR、Pseudo-RIS; CVPR: WSLS、MATN、Pseudo-Q、Global-Local、PPT、RelVLA; Doctoral. dissertation: WSMIL; ACM MM: KPRN、QueryMatch; ICCV: ARN、SAG、Chunk、TRIS、CPL; NeurIPS: CCL、PCNet; Arxiv: TSEG、SSC、Ref-Diff、BASP; TPAMI: DTMR; TMM: Cycle-Free; ACL: ReCLIP; Findings of ACL: EMNLP: TAS、SGEPG; MMM: PAPT; TOMM: UniRelo; PRICAI: HPFD; TCSVT: BMS、MCCE; Neurocomputing: BICM; ICASSP: VGDIFFZERO。

图 2　本文涉及的弱监督 RIS 与 REG 方法

Fig. 2　Paper focuses on the weakly supervised RIS and REG methods

2　图像−文本对的弱监督

图像−文本对仅提供了输入图像与文本的关联性, 它的弱监督性质在于: 模型需通过学习文本与图像全局特征的对齐关系, 隐式推断图像中与文本相关的对象位置。以下将分别讨论这一监督信号下的 RIS 和 REG 方法。

2.1　RIS

在 RIS 任务中, 由于图像−文本对仅提供图像级监督信号, 缺乏像素级的精确标注, 传统的像素级监督方式难以直接应用。针对这一问题, 现有方法主要沿两个方向展开: 一类方法聚焦于图文模态特征的深度交互, 通过跨模态特征融合策略增强图像与文本的对齐, 以端到端的训练方式尽可能从文本信号中挖掘隐含的位置信息; 另一类方法则借助多模态大语言模型(multimodal large language model, MLLM)及掩码提取器(mask extractor), 通过构造像素级伪标签, 结合阶段性训练策略, 引导模型逐步学习精确的目标定位能力。基于模型的训练范式, 本文将基于图像−文本对的 RIS 方法划分为单阶段方法和多阶段方法, 分别对应端到端特征交互范式和伪标签引导的逐步优化范式。

2.1.1　单阶段的 RIS 方法

单阶段的 RIS 方法旨在利用端到端的训练范式, 直接从图像−文本对中学习目标物体的定位信息。如图 3 所示, 这类方法通常围绕跨模态特征交互展开, 通过构建图文融合策略使模型在联合编码过程中捕捉文本描述与图像区域的对应关系, 以增强跨模态对齐能力。例如, 一些方法通过引入注意力机制或对比学习策略, 以挖掘文本信号中隐含的空间指引信息, 从而提升模型对目标区域的感知能力。

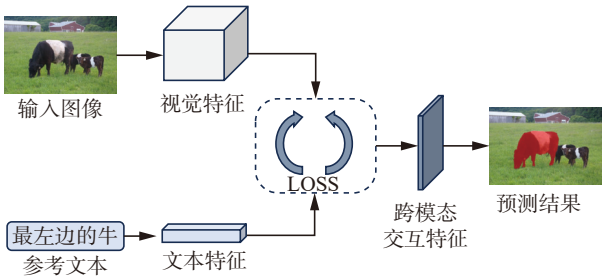


图 3　基于图像−文本对的单阶段 RIS 示意

Fig. 3　Illustration of single-stage RIS based on image-text pairs

初期的研究[30-31]主要通过多示例学习[32-33]或简单的视觉、语言编码器及交互模块, 关注如何利用较粗糙的注意力图或热力图及对应的相似度分数来尽可能还原掩码形式的目标对象。利用图像−文本对作为监督信号进行 RIS 任务的雏形体现在 Xiao 等[30]的研究上。其设计的方法 WSLS(weakly-supervised visual grounding of phrases with linguistic structures)由相对简单的视觉编码器 CNN(convolutional neural network)[34]、语言编码器 LSTM(long short-term memory)[35]、语义嵌入模块组成, 对应接收图像和图像相关的描述短语或短句, 其中语义嵌入模块由简单的多层感知机构成, 通过将视觉特征投射至与语言特征的共同特征空间中构建其与输入短语的关联。此外, 作者设计了一种分层的解析树结构, 通过词汇间的关系约束构建损失函数, 以消除歧义, 增强图像区域与短语配对的准确性。最终让模型输出对应文

本描述的像素级注意力掩码，得到预测区域。但该方法受语言解析器性能的影响较大。且可视化的注意力图也受图文编码器的特征编码能力影响，对于目标物体的边界轮廓无法较为准确地呈现。因此，面对更复杂的句子，模型难以从语义上更准确地理解。为实现端到端的句子定位，Dong[31] 提出 WSMIL（weakly supervised multi-instance learning）方法，该方法由两种相互独立的多示例学习[33] 损失函数，对通过图文特征（分别经 CNN 和 LSTM 编码）运算交互获得的单通道热力图进行处理。其以热力图中的像素点为对象，选取若干个固定数量的最大与最小值的像素点，分别视为前景像素集合与背景像素集合，以缓解传统多示例学习中只优化单个极值点可能带来的不稳定性，并对这两个集合进行约束，使得归一化的前景像素值被优化为接近 1，背景像素值被优化为接近 0，以此逐步提高模型对前景和背景区域的区分能力。这类单阶段方法的掩码预测由模型中间预测推测而来，精度较低，且难以处理稍微复杂的句子。

进一步地，Strudel 等[36] 提出了方法 TSEG（text grounded semantic segmentation），其使用 Transformer[37] 编码并计算图像块−文本的相似度分数，并在训练阶段使用多标签图像块分配机制来引导目标的类别划分。该方法先用独立的图像和文本编码器将图像块和文本各自编码，通过余弦相似度计算图像块与多条文本的相似度分数矩阵，对其通过全局权重池化的方式为每个图像块生成关于不同参考文本的分配概率，再利用归一化和插值的方式获得像素级的掩码，最终将图像块−文本的相似度分数还原为图像−文本的相似度分数。在这个过程中，模型可以利用图像块与不同参考文本的匹配分数与真实图像−文本对的分数计算损失，通过端到端的训练来实现基于图像−文本对的弱监督 RIS 任务，该方法进一步提升了模型在任务上的表现。然而，仅利用独立的图像块与不同文本的匹配关系，无法充分考虑到参考文本内部及各图像块之间的关联，难以处理包含对象关系描述的句子。

Lee 等[38] 和 Kim 等[39] 考虑到参考文本及图像间的关联性问题，对应提出了方法 Chunk（intra-chunk and inter-chunk consistency）和 SAG（shatter and gather），它们将关注点聚焦于如何通过挖掘图像与参考文本之间的关联、参考文本之间的相互作用以及文本内部语义关系，从而获取更准确的目标对象掩码。作为当中的关联技术，槽注意力[40]（slot attention）可以通过无监督方式学习图像中独立对象的表示。其核心在于利用一组可学习的嵌入向量，通过迭代的注意力操作，将输入数据中的信息逐步分配到各个槽，使每个槽最终表示一个对象。为更有效地发现细粒度视觉对象，SAG[39] 改进了传统的槽随机初始化方式，通过从多个对应特定语义的高斯分布中采样槽的初始值，确保初始化具有多样性。其设计的模型包括两个模块：1）自底向上的注意力模块，包含聚合模块和交互模块，分别负责将不同视觉实体的特征聚合到槽中，并在单一语义类别内区分不同实体。2）自顶向下的模态融合模块，通过交叉注意力机制，将文本特征与槽相互作用，生成跨模态嵌入向量，以推断文本与视觉实体之间的关联。此外，作者通过对跨模态嵌入与文本特征进行对比学习，以隐式捕获参考文本与图像实体间的关系。尽管该方法通过多种精心设计的注意力计算模块增强了图像与文本模态的对齐，但其注意力计算的迭代操作可能增加模型的复杂性。Chunk[38] 利用现成的视觉语言模型 ALBEF（align before fuse）[41] 进行图像−文本匹配，通过 Grad-CAM（gradient-weighted class activation mapping）[42] 生成视觉显著图，以定位图像中与每个单词相关的区域。Grad-CAM 是一种可视化技术，可以通过计算网络梯度与特征图的加权关系，生成反映特定输出与输入区域相关性的热力图。然而，单纯的 Grad-CAM 仅聚焦目标对象的局部区域，且无法充分考虑单词间的语义关系，因而定位结果的准确性受限。对此，Chunk 引入“词块内和词块间一致性”方法，通过构建基于名词块的语义关系提升 Grad-CAM 的鲁棒性和连贯性。此外，为了克服 Grad-CAM 局限于小范围区域，且难以准确呈现物体间边界的缺点，Chunk 提出两种优化方法：一是利用自注意力机制，根据像素间的关联度，将每个像素的定位分数传播到其语义相关的邻近像素点，从而扩展目标区域；二是通过无监督的对象形状先验方法（multiscale combinatorial grouping，MCG[43]），通过无监督的方式为图像中的对象生成多个掩码提议（mask proposals），再结合现有预测区域进行比较，筛选出最接近的掩码提议作为预测，从而进一步优化分割结果的边界。

不依赖于 Grad-CAM 生成相对粗糙的热力图，Chen 等[44] 受方法 DETR（detection Transformer）[45] 的启发，提出了基于查询配对的方法 QueryMatch，利用一组可学习的查询（queries）表示候选目标，并通过 Transformer 交互以学习查询与视觉对象的对应关系。最终，正确的查询经解码层转换为

掩码或边界框预测。在该过程中, 查询的选择直接影响模型性能。为此, QueryMatch 采用基于查询的对比学习, 通过最大化正类查询与参考文本的相似度, 同时最小化负类查询的相似度来训练模型。然而, 简单增加负类查询数量并不能保证模型效果, 为了确保负类查询的质量, QueryMatch 设计了优化负类查询选择策略: 1)独特性, 计算待选负类查询与已选负类查询的相似度, 值越低, 表示该负类查询在特征空间中更具独特性; 2)区分度, 计算待选负类查询与参考文本的相似度, 值越低, 说明该查询越难区分, 能够更有效地挑战模型判别能力。QueryMatch 通过迭代评估查询的独特性和区分度, 优先选择难以区分且语义独特的负类查询, 以增强对比学习效果。这一方法显著提高了图文匹配的对齐能力, 同时优化了基于查询的视觉分割和定位任务。

2.1.2　多阶段的 RIS 方法

仅基于图像–文本对标注时, 上述单阶段方法多数难以实现精细的像素级分割结果。此外, 部分单阶段模型依赖更复杂的模块设计, 以增强跨模态特征对齐能力, 但可能带来较高的计算成本和优化难度。相比之下, 多阶段方法通过引入额外的中间训练阶段, 以弥补图像–文本对监督信号在像素级指导上的不足。如图 4 所示, 多阶段方法通常利用预设网络生成初步伪标签(Liu 等[18]), 再借助优化策略或掩码提议对伪标签进行进一步优化, 或直接将掩码提议作为新的伪标签。之后, 结果预测网络在此基础上进行训练, 以获得直接或间接的预测结果。结果预测网络可以是生成点提示(Dai 等[46])或进行图文匹配(Eiras 等[47])的网络, 也可以是直接输出掩码的分割网络。值得注意的是, 在此分类体系下, “多阶段”相对于“单阶段”而言, 因此包含了两阶段模型。

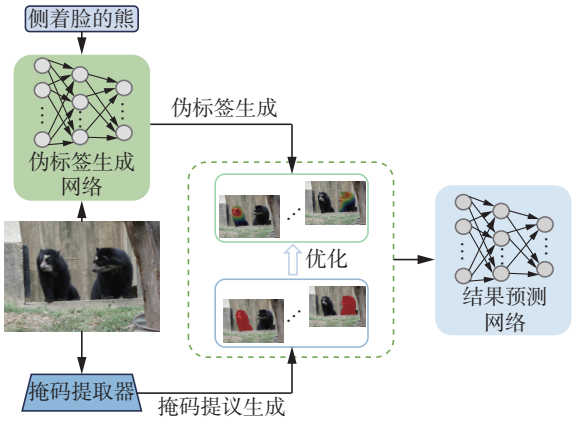


图 4　基于图像–文本对的多阶段 RIS 示意

Fig. 4　Illustration of multi-stage RIS based on image-text pairs

Liu 等[18] 提出方法 TRIS(referring image segmentation using text supervision), 采用响应图构建伪标签, 并利用这些伪标签作为监督信号, 对现有分割模型进行训练, 以实现目标对象的精准预测。该方法认为输入的文本描述已包含目标对象的关键属性, 并能提供足够的语义信息用于目标区域的定位。为充分挖掘图像与文本的对应关系, TRIS 设计了一种双向提示注意力机制, 计算输入图像的每个像素与多条参考文本的响应图, 以衡量像素与不同文本描述之间的关联度。此外, 为了增强伪标签的质量, TRIS 通过分类损失与校正损失, 将模型学习目标对象定位的任务转化为区分正负类文本的任务, 使其能够更有效地提取目标区域。随后, 通过对响应图进行筛选与增强, 以获取携带像素级监督信号的伪标签, 从而为后续的分割模型提供训练数据。然而, 类似于注意力图, 所生成的响应图可能会因目标区域边界模糊而导致伪标签的不准确性, 进而影响最终分割模型的性能。对此, Dai 等[46] 和 Eiras 等[47] 分别提出方法 PPT(point prompting)和 SSC(segment, select, correct), 它们利用预训练的掩码提取器生成更细粒度的掩码提议, 通过不同的选择方式或提示方法, 促进掩码提议和参考文本的正确配对。

作为一种掩码提取器, SAM(segment anything)[48] 在生成图像分割掩码方面表现优异, 其可在推理阶段接收边界框或像素点作为提示生成掩码, 或实现无监督的掩码提议(不含语义信息)生成。然而, 当简单以图像–文本对中提取的像素级语义响应作为输入提示时, 可能包含噪声, 或激活了参考文本中提及的非目标对象与属性, SAM 在处理这些提示时由于鲁棒性不足, 可能生成冗余或不完整的掩码。针对上述问题, PPT[46] 设计了一种结合多源课程学习策略的点提示框架。通过引入轻量级的点生成器, 为 SAM 生成可靠的点提示来获得更准确的预测掩码。首先, 利用多模态预训练模型 CLIP(constrastive language-image pre-training)[49] 提取图像和文本特征, 并通过交叉注意力层和多层感知机构成的轻量级点生成器, 接收上述特征以及一组可学习点查询, 用于生成目标内像素点的预测。其中, 点查询代表目标对象的分割掩码, 它能够有效抑制噪声干扰, 并通过将噪声响应划分为负类, 解决输入提示的鲁棒性问题。此外, PPT 通过构建基于 ImageNet[50](作为对象中心化数据)和 RIS 数据集[2-3](作为复杂场景数据)的多层次数据集, 利用课程学习策略,

依据任务目标难易程度逐步训练点提示生成网络：1）语义级别训练，基于 ImageNet 数据，学习对象的语义特征；2）参考级别训练，结合部分 RIS 数据集，学习对象的关系与参考文本处理能力；3）领域微调，仅在 RIS 数据集上微调，减少跨域分布差异，进一步提高模型在复杂场景中的表现。PPT 框架通过结合多源课程学习策略和点生成器，显著提高了点提示的鲁棒性和分割掩码的精确性。相比直接使用 SAM 处理弱监督输入提示，PPT 能够更好地捕获参考文本与目标对象之间的语义与空间关联，为复杂场景的 RIS 提供了更稳定的解决方案。然而，与 TRIS[18] 类似，PPT 训练时的性能仍会受制于构建伪标签的质量。

为保证伪标签质量，SSC[47] 将任务的核心分为伪标签构建、小规模数据筛选及配对模型优化 3 部分。1）伪标签构建：利用开放式目标检测器 Grounding DINO[51]（DETR with improved denoising anchor boxes）和掩码提取器（如 SAM 或 Freesolo[52]，free segmenting objects by locations），针对每条参考文本生成掩码提议。结合每条参考文本的关键名词与掩码提议的对应关系，构建一个较大规模的伪标签数据集。2）小规模数据筛选：使用 CLIP 提取参考文本的全局特征，并通过背景模糊等操作提取伪标签中掩码提议的特征，并计算二者的余弦相似度，筛选最相关的掩码，以构建小规模高质量预测数据集，该阶段的结果也可作为无标注场景中的预测。3）配对模型优化：将生成的小规模预测数据集作为标签，通过预训练来初始化配对模型的权重。随后对配对模型进一步训练，并利用图像语义与文本的对应性设计一种掩码配对损失函数，通过最大化模型的预测掩码与来自阶段 1）掩码提议之间的特征相似性，驱动模型优化图文配对性能。

与上述逐步优化的方法类似，Yang 等[53] 提出了方法 PCNet（progressive comprehension network）通过逐步理解目标相关的属性和关系，引导模型精准确定目标。不同于直接采纳掩码提议作为预测的做法，PCNet 借助掩码提议辅助优化中间预测，逐步提高精度。具体而言，该方法首先利用大语言模型（Mistral 7B[54]）将输入文本分解成多条短句，充当目标相关的语义线索，并采用类似 TRIS 的方法生成响应图，初步构建图文关系。为进一步细化响应图，PCNet 采用一种条件参考模块，通过多阶段更新文本和图像嵌入，结合目标区域的形状先验（来自掩码提议）优化响应图。在这一过程中，模型引入区域感知收缩损失，结合掩码提议的形状信息，逐步引导响应图收敛至更精确的目标形状。不同于直接采用掩码提议，这种做法可以避免因错误选择掩码提议而导致的严重偏差。此外，为增强实例感知能力，PCNet 设计了实例感知歧义消除损失，通过计算不同短句与响应图所对应掩码提议的对齐分数，约束不同描述的激活区域，使其互不重叠，显著提升了模型对实例间的辨别能力。

2.2　REG

由于缺乏直接的目标区域标注信息，传统的 REG 方法通常采用循环前向−后向流程（cyclic forward-backward pipeline）进行优化，这种方法通过视觉特征重构文本的方式，迭代增强模型定位目标的能力。然而，这类方法难以直接提升模型在推理阶段的图文对齐能力，导致定位精度受限。为克服这一问题，后续研究引入了对比学习、提示词微调和伪标签构建等策略，以提升模型的目标感知能力和对齐效果。这些方法不再依赖循环前向−后向流程，而是通过更直接的训练方式引导模型学习目标区域的特征分布。基于训练策略的差异，本文将图像−文本对监督信号下的 REG 方法分为循环前向−后向流程方法和非循环前向−后向流程方法。

2.2.1　循环前向−后向流程方法

循环前向−后向流程包含两个关键模块，如图 5 所示：1）前向模块，利用注意力机制或跨模态匹配，将参考文本映射到候选的目标区域，预测其在图像中的位置；2）重构模块，根据前向模块选择的目标区域，生成描述文本，并计算其与原始参考文本的差异（例如使用序列损失）。通过这一循环流程，模型能够在无目标区域标注的情况下，逐步优化前向模块的目标选择能力，使其更精准地定位图像中的参考对象。该类方法可进一步表示为

$$\boldsymbol{\alpha}_i = \mathrm{f}_{\mathrm{ATT}}(p, r_i) = \mathrm{MLP}(\boldsymbol{t}, \boldsymbol{v}_i), \quad i \in S \tag{1}$$

$$\boldsymbol{v}_{\mathrm{att}} = \mathrm{MLP}\left(\sum_{j=1}^{N} \boldsymbol{\alpha}_j \boldsymbol{v}_j\right), \quad j \in A \tag{2}$$

$$L_{\mathrm{rec}} = -\frac{1}{B}\sum_{b=1}^{B} \log(\mathrm{P}(\hat{p}|\boldsymbol{v}_{\mathrm{att}})) \tag{3}$$

其中，式 (1) 表示前向模块，式 (2) 和 (3) 表示重构模块。式 (1) 中，$\boldsymbol{\alpha}_i$代表注意力权重，$\mathrm{f}_{\mathrm{ATT}}$代表注意力函数，MLP代表多层感知机，$p$为参考文本，$r_i$为边界框提议（bounding box proposals），$\boldsymbol{t}$为文本特征，$\boldsymbol{v}_i$为视觉特征，S表明当前图像中所有边界框提议的集合。式 (1) 描述了前向模块计算参考文本与不同边界框提议之间关联性的过程。式

(2) 中，$\boldsymbol{v}_{\text{att}}$ 代表进一步编码后的加权视觉特征，A 表示模型关注的边界框提议，可根据前向模块（式 (1)）的计算结果，通过阈值等方式筛选得到。式 (3) 中，L_{rec} 为重构损失函数，B 代表批量大小，$\hat{p}$ 为输入的参考文本，$\mathrm{P}(\hat{p}|\boldsymbol{v}_{\text{att}})$ 表示输入 $\boldsymbol{v}_{\text{att}}$ 条件下，得到真实单词序列的概率分布。

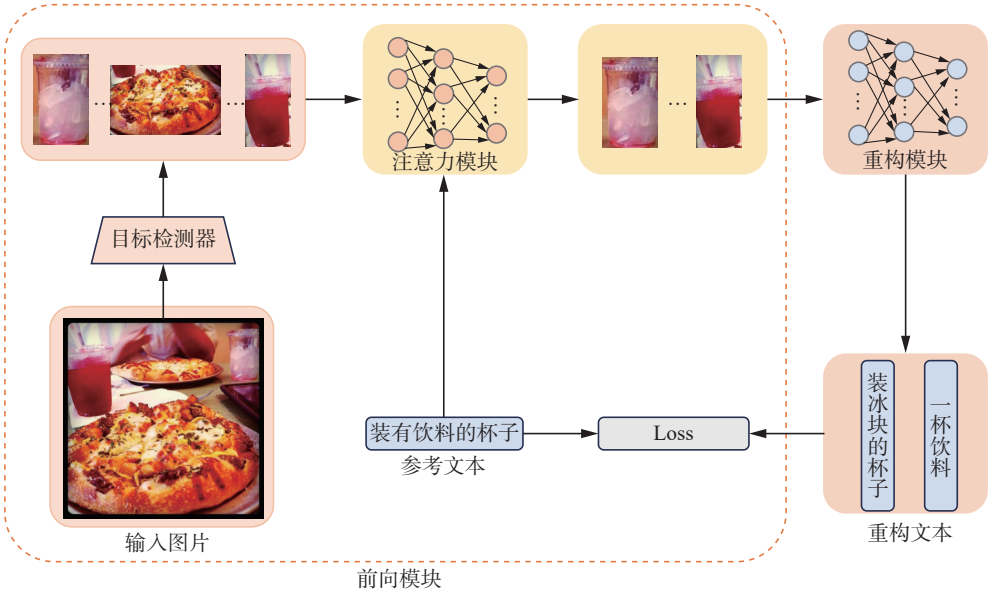


图 5　基于图像–文本对的循环前向–后向流程示意

Fig. 5　Illustration of cyclic forward-backward pipeline based on image-text pairs

因此，重构损失函数 L_{rec} 衡量了模型生成句子的概率与真实句子概率的匹配程度。在训练过程中，两个模块作为一个整体网络，模型通过反向传播，最大化重构模块中生成真实句子的概率，以优化整个网络。

作为循环前向–后向流程的典型做法，Rohrbach 等[55] 提出的 GroundeR（grounding of textual phrases in images by reconstruction）利用句子–区域匹配网络（LSTM 结构）作为前向模块，计算参考文本与边界框提议（由目标检测器生成）之间的匹配得分，训练时模型选择部分最相关的图像区域作为结果区域；后向模块直接将结果区域进行句子级重构。基于该循环过程，模型将重构句子与原始参考文本之间的差异作为损失函数，指导网络输出与参考文本尽可能接近的概率分布，从而优化整个网络。推理时，仅利用前向模块即可生成目标的边界框预测。在相同范式上，Liu 等[56] 提出 KPRN（knowledge-guided pairwise reconstruction network）方法，将参考文本解析为主体–对象对，把边界框提议按主体–对象关系分成多个提议对，通过两种信息的跨模态配对来增强模型对目标对象的理解。其利用编码器不同层的输出，将边界框提议编码为主体及对象特征，并计算它们与参考文本中的主体–对象对的跨模态相似度以进行匹配。同时设计两种可选择的边界框过滤策略以优化匹配质量：1）软过滤。基于主体特征与文本主体的相似度，为主体边界框提议分配不同权重，保留所有提议，并在后续步骤调整权重。2）硬过滤。基于相似度阈值，直接去除低置信度的主体边界框提议。在对象边界框提议选择时，直接匹配与文本对象特征相似度最高的对象边界框提议，以获得过滤后的主体–对象提议对。此外，KPRN 利用文本解析的方式引入属性标签，通过对主体边界框的属性分类，构建属性分类损失，以更好地区分同类别下的不同实例。不同于仅依赖单个边界框提议进行句子重构，KPRN 采用边界框提议对，结合目标对象的上下文信息，提供了更丰富的视觉特征，从而提升句子重构的完整性和匹配效果。另一些方法则通过解析后的参考文本和图像区域进行匹配，利用语言特征与视觉特征层面的关联性对网络进行优化。例如，Liu 等[57] 提出的另一种方法 ARN（adaptive reconstruction network），通过对文本和视觉特征进行解析和预定义的方式，将其各自划分为主体特征、位置特征和上下文特征，通过分层的注意力模块计算参考文本与边界框提议的匹配分数来定位目标边界框，以减轻多样的参考文本带来的偏差。训练时，除了基于句子重构的损失外，还通过最小化参考文本成分的特征与其相应视觉特征之间

的距离，来优化整体网络。Zhao 等[58] 通过多尺度的锚框 Transformer 网络 MATN（multi-scale anchored transformer），以边界框提议作为锚点，可以在整个特征图上搜索短语对应的图像区域，得到更准确的位置。结合对比重构损失，通过图像内部及不同图像之间的关系，优化短语之间的对比关系及相似度关系，以提高模型对不同短语的辨别能力，并增强模型面对相似短语时的鲁棒性。

然而，在上述多数做法中，重构模块均基于句子层面的逐句重构，当重构句子的含义与标签相同而表述不同时，通过逐字计算而来的损失值仍可能很大，容易导致网络难以收敛。针对这一问题，Sun 等[59] 提出了一种判别性三元组策略 DTMR（discriminative triad matching and reconstruction），通过将参考文本解析为一个或多个三元组来替代原始句子，从而简化表达并提高匹配精度。其中，每个三元组由 3 个单元构成：1）目标对象；2）参照对象，与目标对象相关联的其他对象；3）判别关系，描述目标对象与参照对象之间的区别关系。在匹配过程中，边界框提议对的第 1 个边界框对应目标对象，第 2 个边界框对应参照对象，两者的联合区域则对应判别关系部分。模型分别计算这 3 个单元之间的匹配（注意力分数），以评估边界框提议对与三元组的对应关系。为筛选配对结果，DTMR 设计了两种筛选策略：1）硬性方法，直接丢弃最高配对分数以外的提议；2）软性方法，根据匹配相似度为提议对分配权重。此外，DTMR 采用基于三元组层面的重构模块，不同于传统方法在句子层面逐字重构参考文本，该模块将边界框提议对进行三元组重构，并计算重构后的语言特征与原始三元组各单元之间的差异作为损失。这种设计更注重细粒度的语义对齐，从而提高整体的匹配效果。最终，通过整合每个三元组的匹配分数，模型可确定边界框预测结果。

2.2.2　非循环前向−后向流程方法

循环前向−后向流程方法的间接优化策略虽然提升了无区域标注时文本和区域的匹配精度，但可能存在偏差问题[60]：当前向模块通过参考文本选择的图像区域与目标区域完全偏离时，模型对这种配对的正确性是未知的，后向模块仍可能对偏离的预测区域重构出与参考文本相似的句子，而无法对这种偏离产生惩罚。基于这一问题，Sun 等[61] 后续摒弃了传统的循环前向−后向流程方法，提出一种无需循环的方法 Cycle-Free。它借鉴 DTMR[59] 的判别三元组与 KPRN[53] 中边界框提议对的构建方式，根据参考文本与每一候选图像区域对映射的文本描述之间的相似性来选择结果区域。具体而言，Cycle-Free 摒弃了传统对参考文本或其特征重构的重构模块，而是设计区域描述符来建立文本和视觉区域的关联。其由简单的全连接网络构成，可以接收来自边界框提议对的视觉特征，输出分别对应于边界框提议对中第 1、第 2 及整个边界框提议对的判别性描述。训练时，直接通过最小化这 3 种描述与参考文本三元组中各单元的差异（特征距离）来优化网络。为缓解区域描述符在训练初始阶段的性能不足，作者提出一种自主优化的策略，使网络能依据损失自主选择当前更简单的样本逐步增强自身性能。

在前述利用语言特征与视觉特征的相关性对网络进行优化的做法中，如何更好对齐视觉与文本的特征是个关键问题，Zhang 等[62] 提出 CCL（counterfactual contrastive learning）方法，通过构建反事实的正样本（保留关键信息）和负样本（破坏关键信息），直接优化视觉−文本对齐分数，缓解了现有方法依赖随机负样本或间接优化的缺陷。具体而言，其采用 3 种应用于网络推理中间过程的反事实转换策略，分别作用于特征、交互和关系 3 个层面：1）特征级扰动。改变候选区域的视觉特征，引入干扰或破坏局部信息，以提高模型的鲁棒性。2）交互级扰动。影响视觉−文本的交互方式，迫使模型在不同的匹配条件下学习稳定的跨模态对齐。3）关系级扰动。调整候选区域之间的上下文关联，使模型在缺乏显式关系线索时仍能正确理解目标。在训练过程中，CCL 通过对比学习最大化反事实正类与文本的匹配，同时最小化反事实负类的相似度，使模型能够在干扰条件下保持稳健的目标−文本对齐能力。此外，该方法不仅提升了模型的鲁棒性，还增强了其在缺乏显式关系线索时的目标定位能力。在模态融合阶段，CCL 采用注意力机制来增强视觉与语言模态间的交互，这在建立跨模态关联的同时也能优化单一模态的特征表示，从而进一步提高模型对文本和视觉语义的理解能力。但在实际推理过程中，单纯的注意力机制往往只关注最显著的部分，而难以涵盖整个物体，这容易导致无法预测完整的目标对象。

为缓解这一问题，Zhao 等[63] 提出方法 PAPT（part-aware prompt tuning），在借助 MLLM-X-VLM（vsion-language model）[64] 的基础上，使用 3 种不

同的可学习部分感知提示来对应引导模型关注目标对象的上、中、下 3 个部位, 以增强模型对局部信息的感知能力。具体而言, 对于提取的文本特征, 进行 3 种提示(可学习的矩阵)的拼接, 分别获得 3 种增强的文本表示; 同时, 视觉特征沿着由上而下空间维度被分为 3 个部分, 对应物体的不同部位, 以获得局部感知的视觉特征。在此基础上, 通过对比学习, 将不同物体的对应部分及不同物体的不同部分之间进行特征对比, 以提高模型对对象级区分与局部细粒度理解的能力。最终, 通过微调可学习的文本提示, 模型可在推理阶段获得更完整的 Grad-CAM 热力图, 从而在现有的边界框提议中更准确地筛选出最终预测目标。与文本提示类似, 视觉提示也是视觉−语言任务中一种常见的有效辅助方法[65], 它可以通过对图像进行细微加工, 如模糊、圆圈标记、点提示等操作, 帮助模型快速关注到目标对象, 从而促进定位学习。Zhang 等[66] 提出的 UniRelo(universal relocalizer)方法通过引入类别模块、颜色模块和空间关系模块来提升区域提议筛选精度。在提取视觉特征前, UniRelo 根据区域提议对图像进行模糊和裁剪等操作, 以增强模型对目标区域的关注, 降低背景噪声的影响。其中, 类别和颜色模块通过图文特征相似度计算的方式, 为每个区域提议分配预定的类别/颜色标签, 并计算该标签与参考文本的类别/颜色相似度, 获得对应的分数。空间模块基于区域提议的中心坐标和相对面积, 可计算区域的位置及相对大小, 以获得空间分数。最终, 通过加权融合上述分数获得综合评分, 用于筛选区域提议。该方法的最终评分能够集成到多数弱监督 REG 框架中, 而无需额外的训练。UniRelo 充分利用了输入图像的视觉模态信息, 通过显式建模类别、颜色、空间等视觉属性, 从细粒度上建立了与参考文本的联系。此外, 通过模糊裁剪等操作能降低背景的干扰。通过加强视觉层面的信息, UniRelo 利用视觉与参考文本的细粒度关联, 显著提升了区域提议的判别度与可靠性。

上面提到的多数方法用到了目标检测器获得区域提议, 并利用跨模态的相似度等方法对区域提议进行筛选。这类方式可能由于图像和文本模态的特征分布、表示方式等差异(跨模态异质性), 模型在学习过程中将文本和非对应的视觉区域匹配而产生错误的高置信度关联, 且由于弱监督场景中缺乏精确的目标标注, 这种错误无法被直接纠正, 模型可能会对这种错误关联赋予高置信度, 并在训练过程中不断积累, 对预测带来负面影响。对此, Liu 等[67] 提出一种置信度感知的伪标签学习方法 CPL, 以缓解跨模态异质性导致的伪标签噪声问题。其基于提示词工程为区域提议生成多视角的文本描述, 包括目标类别、属性、空间关系等, 并利用 BLIP(bootstrapped language-image pre-training)[68] 计算这些描述与参考文本的匹配得分, 以筛选高质量的区域−文本对作为伪标签。为进一步抑制错误匹配, CPL 引入置信度感知的伪标签验证模块, 通过 BLIP 计算图文匹配分数作为可信度指标, 并将其作为损失函数权重的一部分结合到模型训练中, 以减少错误伪标签对模型学习的干扰。

3　基于无标注的弱监督

为应对标注成本高的问题, 近年来研究者们逐渐采纳更宽泛的弱监督定义, 例如祁磊团队[15] 和 Shen 等[17] 将无标注数据的无监督学习也纳入弱监督学习的研究中。为更全面地呈现 RIS 与 REG 中弱监督技术的研究图景, 本文也将这些基于无标注数据的策略视为弱监督方法的一部分。

无标注数据要求模型在无外部监督的情况下, 仅依赖未标注图像或关联性未知的图像−文本完成任务。这意味着模型需要借助启发性知识或预训练模型的先验知识, 以实现目标对象的定位和分割。在 RIS 和 REG 中, 现有的方法通常依赖掩码提取器或目标检测器, 如 SAM[48]、Freesolo[52]、Mattnet(modular attention network)[9]、Grounding DINO[51] 等, 为图像生成候选区域(掩码或边界框提议)。随后, 利用 MLLM 处理跨模态信息, 以完成区域−文本匹配, 从而直接进行目标预测, 或生成伪标签以供模型进一步优化。此外, 部分方法探索了生成模型(如 stable diffusion[69])的应用, 利用其图文对齐能力实现目标定位与分割。在现有 RIS 和 REG 研究中, 基于 CLIP 进行候选区域的图文匹配是一种常见策略。本文将这一基于候选区域生成与筛选的范式归纳为“区域提议生成−图文匹配筛选”框架, 其基本流程见图 6。在该框架下, 方法首先对获取的区域提议(掩码或边界框)与参考文本进行编码, 通常采用 MLLM(如 CLIP)提取特征。随后, 通过计算跨模态相似度并进行对比分析, 从初始的区域提议中筛选出最符合参考文本描述的目标区域, 作为最终预测结果。

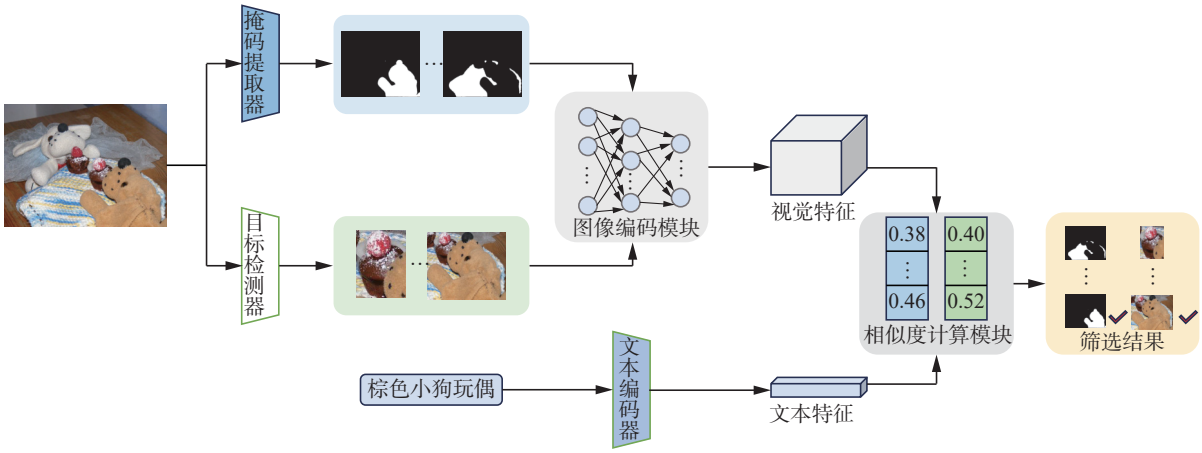


图 6　区域提议生成−图文匹配筛选框架示意

Fig. 6　Illustration of region proposal generation and image-text matching framework

3.1　RIS

在区域提议生成−图文匹配筛选框架下，Yu 等[70] 首先提出一种 RIS 策略 Global-Local。该策略利用全局−局部上下文特征来增强对目标对象的语义理解与关系建模，可以实现无需训练的零样本学习[71]（zero-shot learning）。Global-Local 基于 CLIP 文本编码器和改进的 CLIP 视觉编码器，分别提取文本与图像的全局−局部特征以增强跨模态匹配。其中，全局视觉特征既包含目标掩码区域，也保留了周围环境信息，可帮助模型捕捉对象间的上下文关系；局部特征直接来源于裁剪后的目标掩码区域，可令模型更加聚焦于目标对象本身；全局文本特征由完整的参考文本构成，以提供完整语义信息，局部文本特征则由参考文本解析得到的目标名词短语获得，以强化目标实体的语义表示。最终，通过加权融合全局−局部特征，计算跨模态相似度分数，用于目标掩码提议的筛选。通过全局−局部上下文的联合建模，Global-Local 提升了零样本 RIS 任务中的目标定位和语义对齐能力。然而，由于 MLLM 原本任务的训练特性，如 CLIP 的预训练任务仅用于图像分类，其更容易在全局层面对整个图像和文本进行配对，而不擅长参考文本和图像区域这种更细粒度层面的配对任务，同时也欠缺了处理空间匹配的能力。对此，Suo 等[72] 设计了一种弥补 CLIP 在无标注 RIS 任务中不足的框架 TAS（text augmented spatial-aware）。该框架由一个掩码提取器、CLIP 的图文编码器、标题生成器、相似度分数计算模块以及一个空间校正模块构成。其中，相似度分数计算模块计算 3 种相似度分数：1）裁剪掩码区域与文本特征的相似度，以初步评估二者的配对关系；2）利用标题生成器 BLIP-2[73] 为经背景模糊的掩码区域生成描述，并计算其与参考文本的相似度，以缓解跨模态域差距带来的特征对齐困难；3）利用 BLIP-2 为目标对象生成负类描述，并计算其与视觉特征的相似度，帮助排除背景对象的干扰。此外，空间校正模块基于掩码提议的中心坐标，结合参考文本中的方位词对掩码提议进行筛选，以提升模型的空间感知能力。通过结合多视角相似度计算与空间约束，TAS 有效增强了 CLIP 在区域级别的图像−文本匹配任务中的表现。

在前述掩码提议生成−筛选的框架中，掩码提议的质量至关重要，若掩码提议存在太多冗余区域，则会带来过多干扰项，同时影响预测精度。对此，Li 等[74] 关注高质量掩码的获取和特征信息挖掘上，提出了 HPFD（hierarchica prompts and frequency domain fusion）方法。为获得高质量掩码，其依次结合使用 3 种预训练模型：1）RAM++（recognize anything plus model）[75]，为图像内各实例生成类别标签；2）Grounding DINO[51]，根据类别标签为对象实例生成细致的边界框；3）HQ-SAM（segment anything in high quality）[76]，作为 SAM 的升级版本，根据边界框提示为相应对象生成细致的掩码。为加强模型的空间理解能力，HPFD 解析参考文本中的方位词，利用非线性激活函数在对应方向坐标轴上生成 1~0 逐渐衰减的方向偏置矩阵，该矩阵通过 Hadamard 乘积，将方位信息以权重的方式融合至原始图像中。在图文特征提取上，HPFD 沿用了 Global-Local 的特征提取方式，并借鉴 TAS 的空间校正模块和负类文本生成方法，优化目标掩码筛选。为充分挖掘物体的轮廓和边缘特征信息，不同于 Global-Local 直接加权结合全局与局部特征的做法，HPFD 利用非下采样轮廓波变换[77] 对图像特征进行融合，用哈尔小波转换[78] 对文本特征进行融合。其中，图像和文

本特征均包括综合特征(类似前述全局特征)及焦点特征(类似前述局部特征)。最终, 通过计算融合的图文特征之间的相似度, 筛选出最优目标掩码。在排除非目标对象区域方面, 与文本模态上的负类描述设计不同, Li 等[79] 将重点放在视觉层面, 提出了 BMS(bidirectional mask selection)方法, 区别于 TAS 通过负类文本描述对背景进行约束, BMS 侧重于视觉层面的正负掩码对比, 通过双向对比机制提高目标选择的准确性。其将负类掩码考虑进任务中, 以充分利用目标对象周围的上下文语义。在同一图像中, BMS 将每个掩码提议之外的所有掩码提议集合作为负类掩码, 当前掩码提议则视为正类掩码。之后, 分别计算正类掩码(结合模糊操作)和负类掩码与参考文本的相似度, 并以加权求和的方式融合, 获得每个掩码提议的最终得分。此外, BMS 根据正负掩码的互补信息, 提出一种自适应的掩码融合策略, 该策略可以动态调整两种相似度分数的融合权重, 以便适应不同场景, 提高零样本学习的鲁棒性。

不同于上述直接基于相似度对比筛选掩码提议方法, Yu 等[80] 提出一种基于伪标签生成的 RIS 框架 Pseudo-RIS, 以实现自动化、高质量伪标签构建。在掩码和描述的生成中, 利用高泛化性的掩码提取器(如 SAM[48])为输入图像生成掩码区域, 再针对每个掩码区域用文本生成模型(如 CoCa (contrastive captioners)[81])生成描述。为保证伪标签的质量, Pseudo-RIS 提出了保证描述独特性的采样和过滤策略。前者结合全图掩码的词汇生成概率分布, 抑制高频通用词, 促进低频但具有区分性的词汇生成, 以提高目标描述的独特性。描述过滤策略综合评判了描述的唯一性和正确性: 1)通过比较目标描述与目标掩码, 及其与其余掩码的相似性大小, 评价目标描述的唯一性; 2)计算目标掩码与描述中目标名词短语的相似度来评价目标描述的正确性。结合唯一性和正确性指标, 以量化目标描述质量, 筛选出高质量伪标签。通过自动化伪标签生成与优化, Pseudo-RIS 能提供高质量的完全监督信号。

在区域提议生成−图文匹配筛选和伪标签生成的方法外, Ni 等[82] 借助生成模型, 提出 Ref-Diff (referring diffusional segmentor)框架, 以克服传统判别式模型在区域−文本匹配和空间感知方面的不足, 提升视觉元素的精确定位能力。Ref-Diff 通过引入生成模型(如 Stable Diffusion[69]), 在无标注的零样本学习 RIS 任务中隐式建模视觉−文本关系, 直接生成掩码提议, 而无需依赖第三方掩码提取器。另外, Reff-Diff 可集成判别式模型, 并通过最终的相似度加权融合判别式和生成方法的掩码−文本匹配结果。在判别式通道中, Reff-Diff 设计了一种基于文本方位词的位置权重分配策略, 使参考文本的位置信息嵌入图像特征(详见 HPFD[74] 方法)。之后, 通过 CLIP 编码后的图像−文本特征相似度计算, 作为判别式通道的分数。在生成通道中, 文本与图像的交叉注意力矩阵由生成模型的图文编码器提取, 反映了各文本标记(token)与区域特征的关联性。考虑到不同标记对视觉区域的关注度不同, Reff-Diff 采用语法分析识别句子中的根标记, 聚合了来自其他标记的上下文信息, 能在隐空间中更有效地捕获全局语义。随后, 通过归一化根标记对应的交叉注意力矩阵, 并还原至原图尺度, 再基于预设阈值进行二值化, 生成一系列掩码提议。最终, 通过计算掩码提议与根标记的交叉注意力矩阵之间的相似度, 评估参考文本与掩码提议的匹配程度。Reff-Diff 利用了生成模型的隐式表示能力, 实现端到端的掩码生成与匹配, 避免了外部掩码提取器的依赖, 从而提升模型在不同场景中的适应性。

3.2　REG

REG 任务中, 在区域提议生成−图文匹配筛选的场景下, 由于模型的预测通常源自目标检测器生成的边界框提议, 因而严重依赖目标检测器的性能。为缓解对目标检测器的依赖性, Shi 等[83] 提出了一种双向跨模态匹配框架 BICM(bidirectional cross-modal matching), 通过设计多个模块促进图文融合与匹配: 1)查询感知的注意力图。类似 Grad-CAM, 从文本−图像匹配的注意力图中提取显著区域, 并通过阈值操作生成边界框提议, 实现从文本到图像的无监督区域提取; 2)跨模态目标匹配。为确保边界框提议的多样性, 综合模块 1 生成的边界框与目标检测器的边界框, 并结合目标检测器预测的类别标签, 计算各边界框与参考文本的跨模态相似度; 3)相似性融合。通过加权融合模块 1 生成的注意力分数与模块 2 计算的相似度分数, 整合两种边界框提议的匹配结果, 提升匹配精度; 4)知识自适应配对。先利用 CLIP 匹配图文对, 筛选可靠的匹配样本, 采用模块 3 的相似度预测结果作为伪标签, 训练一个全连接网络, 用于计算不同视觉特征(全局图像、区域特征)和文本特征(完整文本、类别名称)之间的相似性。最终, 加权融合模块 3 和模块 4 计算的相似度, 得到最终的边界框预测。相比传统方法, BICM 在自适应优化匹配结果的同时, 降低了

对目标检测器性能的依赖。除此之外，用于图文配对的模型 CLIP 仍然面临与 RIS 类似的问题，例如缺乏空间感知能力及区域−文本配对能力不足等。

为缓解 CLIP 在空间定位方面的局限性，Subramanian 等[84] 提出了基于关系解析的方法 ReCLIP（a strong zero-shot baseline for referring expression comprehension），其设计了一种启发式空间关系解析器，用于增强模型对图像中物体关系的理解。ReCLIP 在常规边界框提议−文本相似度匹配框架的基础上，引入空间关系建模，预定义了 7 种物体间的空间关系（包括方位、大小、内外关系等）。为解析文本中的空间信息，ReCLIP 采用现成的文本解析器（如 system for dependency parsing, spaCy[85]），将参考文本解析为语义树，其中每个名词短语作为树的节点，依赖路径作为实体之间的关系。在解析过程中，ReCLIP 通过递归更新节点状态，结合子节点信息，衡量其是否指代某个目标对象，并推导其与该对象的空间关系。最终，模型结合空间关系解析器与匹配分数，筛选出最符合参考表达的目标区域，从而提升 CLIP 在 REG 任务中的空间定位能力。然而，ReCLIP 主要关注空间关系，但未充分建模物体间的语义关系。为增强对物体关系的理解能力，Han 等[86] 提出了一种基于图像−文本结构相似性的零样本 REG 方法 RelVLA（relationship-enhanced vision language alignment）。该方法通过构建三元组，对齐图像和文本的关系结构，并计算三元组级别的相似度，以筛选目标边界框。方法主要包括 3 个核心步骤：1）文本三元组构建。利用 ChatGPT 解析参考文本，识别主体，并构造“主、谓、宾”形式的三元组，表示主体与其他对象的关系。2）视觉三元组构建。基于边界框提议，两两组合（包括自身）形成三元组，其中谓语由“主”“宾”边界框的并集区域表示。此外，结合参考文本中的方位信息，对候选三元组进行位置关系过滤。3）视觉−文本三元组匹配。计算视觉三元组与文本三元组的元素级相似度，并通过不同的加权方式，还原至实例级别的相似度分数，以此作为最终边界框筛选依据。RelVLA 通过显式建模对象间的语义关系，有效提升了 MLLM 在零样本 REG 任务中的匹配准确性。

除了空间感知能力的不足，Wang 等[87] 指出 CLIP 在文本到图像检索任务中存在检索幻觉。即当 CLIP 需要根据输入文本从多个候选图像中选取匹配图像时，会选择相似度分数最高的图像，即便在图像到文本的检索任务中，该模型能够正确预测图像内容。作者认为，这种幻觉部分源于 CLIP 在处理不同图像时，对同一组文本的相似度计算结果存在分布不均衡的问题。为缓解这一问题，作者提出 BASP 方法，其核心思想是引入辅助提示文本，利用这些提示文本与图像的相似度分数，对原始图文相似度分数进行归一化平衡。首先，计算一组辅助提示文本（根据常见类别，由大语言模型 ChatGPT3[88] 生成）与图像的相似度；之后采用 Softmax 归一化调整原始图文匹配分数，使不同图像的分数分布更加均衡，以优化最终的匹配决策。此外，BASP 同样适用于区域提议生成−图文匹配筛选框架下的 RIS 任务中，以缓解 CLIP 在零样本检索中的匹配偏差，提高图文对齐的鲁棒性。针对 CLIP 在细粒度多模态对齐任务中的偏差问题，Qiu 等[89] 指出由于 CLIP 的训练机制，模型更关注显著的视觉特征，但这些特征可能与参考文本的描述属性无关，导致其在细粒度跨模态匹配上的表现受限。为解决这一问题，作者提出了一种基于 MLLM 的跨模态对比熵模型 MCCE（MLLM-driven cross-modal contrastive entropy model），旨在增强图文交互，消除显著性偏差。方法主要分为 3 个步骤：1）多文本线索生成。利用 MLLM 生成细粒度的目标描述，同时构造正类和负类线索，可以从多个角度丰富文本信息。2）多线索跨模态交互。计算多条正类线索与全局视觉特征的注意力权重，引导视觉特征的提取，以便获得更聚焦于与文本语义一致的区域；3）对比相似度熵计算。通过相似度计算与对比学习，优化跨模态对齐机制，利用正类线索强化不同属性的文本线索与目标区域间的匹配，提高模型识别能力；利用负类线索抑制无关视觉区域的干扰，减少模型对显著性区域的偏好，从而提高对参考文本的精准理解。MCCE 通过 MLLM 自动构造多样文本线索并结合对比学习机制，有效缓解了 CLIP 在细粒度跨模态匹配中的偏差。

在伪标签生成方面，Jiang 等[90] 设计了 Pseudo-Q（pseudo language queries for visual grounding）方法，通过为无标注的图像生成文本查询作为伪标签，以提供完全监督信号，从而减少对人工标注的依赖。在伪标签构造中，生成的文本查询由名词、属性、空间关系组成：名词和属性采用具备属性分类功能的目标检测器[91]，识别物体类别与属性，并通过置信度阈值筛选结果，以提升文本描述的可靠性；空间关系关注对象间的水平、垂直、深度关系。此外，Pseudo-Q 结合生成的文本查询，设计固定的提示词模板，以增强定位模型对

文本描述的适应性。Pseudo-Q 用于训练的定位模型不同于传统方法仅在最终编码阶段进行特征交互，其跨模态融合模块在每个编码阶段均执行视觉−文本注意力交互，以充分捕获不同模态的上下文信息，提升匹配精度。Pseudo-Q 通过自动构造文本查询，为定位任务提供了完整的监督信号，其在跨模态融合中的逐层交互机制提升了视觉与文本特征对齐的精细度，使模型更具泛化能力。然而，Pseudo-Q 在图中对象关系上仅利用空间位置关系构造伪标签，未能充分捕捉对象间的复杂语义关系，且缺乏对伪标签质量的有效验证，可能导致文本描述的清晰度和准确性受限。

为解决上述问题，Wu 等[92] 提出了一种基于场景图增强的伪标签生成方法 SGEPG（scene graph enhanced pseudo-query generation），以提升伪标签的质量和信息丰富度。方法主要有 4 个基本组成部分：1）核心场景图生成。场景图（scene graph）包含了关于图像的结构化语义信息（如类别、属性、成对关系等），可作为重要的先验知识。借助 IETrans[93] 方法为图像生成初步场景图，图中的顶点代表对象实例，边代表不同对象间的关系。为补充更全面的类别和属性及空间关系，采用类似 Pseudo-Q[90] 的方式获取目标信息，并将其作为顶点对场景图进行更新，以获得核心场景图。2）非歧义查询生成。利用设计的固定模板可将核心场景图转化成文本查询，然而直接基于完整场景图生成文本查询可能导致表述混乱、歧义增加。对此，利用场景图子图的不相交性，即确保目标对象对应的子图（描述其上下文关系的子结构）不与其他对象的子图存在包含关系，来消除视觉定位中的歧义；3）多样性驱动的查询重写。直接转换场景图文本可能导致表述固定、缺乏自然性，影响模型泛化能力。采用 GPT-3.5[94] 生成多样化的表述方式，使文本更加贴近人类表达习惯，并增强数据泛化性；4）视觉−语言模块。采用 VLTVG[10] 结构，利用伪标签训练 REG 模型。通过场景图建模和歧义查询过滤等措施，SGEPG 显著提升了伪标签的质量。

与区域提议生成−图文匹配筛选和伪标签生成等判别式方法不同，Liu 等[95] 提出了 VGDIFFZERO（diffusion models for zero-shot visual grounding）方法，借助预训练的扩散模型 Stable Diffusion[69] 来完成无标注的 REG 任务。该方法指出，扩散模型在 REG 任务中具有两大优势：一是具有强大的视觉文本对齐能力，二是具备对空间关系及细粒度解耦概念的充分知识。VGDIFFZERO 将 REG 任务视为独立区域提议的筛选问题，主要包括两个阶段：1）噪声注入。利用目标检测器生成的区域提议，并分别进行：掩蔽−生成掩码提议集，保留全局上下文信息；裁剪−生成裁剪提议集，关注局部细节。接着，将两种区域提议编码到潜在空间，并注入高斯噪声，模拟扩散模型的前向扩散过程。2）噪声预测。去噪 U-Net[96] 接收文本嵌入和加噪后的潜在向量，执行去噪预测；计算预测噪声与真实采样噪声之间的偏差，衡量模型对区域−文本对齐的预测精度，误差越小，表明该区域提议与文本的语义匹配度越高。最终，VGDIFFZERO 选择预测误差最小的区域提议作为最终预测结果，确保在全局与局部上下文信息的综合考量下，完成区域提议的选择。

4　相关数据集与评价指标

由于任务特性的相似性，RIS 和 REG 通常会使用相同的数据集，主要包括 RefCOCO[3]、RefCOCO+[3]、RefCOCOg（G-Ref）[2,97] 和 ReferItGame[98] 等。然而，由于分割与定位任务的侧重点不同，二者在评价指标上存在差异。本节将介绍 RIS 与 REG 现有常见的几个通用数据集及相应评价指标。

4.1　通用数据集

ReferItGame、RefCOCO、RefCOCO+和 RefCOCOg 数据集具体构成见表 2。

表 2　RIS 与 REG 常见数据集的具体构成信息

Table 2　Detailed composition of common datasets for RIS and REG

数据集	图像总数	指代对象总数	文本描述总数	对象类别数
ReferItGame[98]	19 894	96 654	130 525	238
RefCOCO[3]	19 994	50 000	142 209	80
RefCOCO+[3]	19 992	49 856	141 564	80
RefCOCOg[2,97]	26 711	54 822	104 560	80

ReferItGame[98] 数据集包含多张图像及关于图中对象的文本描述，主要基于 Image CLEF IAPR 图像检索数据集[99] 构建。其文本描述通过双人互动游戏收集，游戏中包括两个参与者，其中一个玩家被分配到带有分割对象的图像，并对该对象进行描述；另一个玩家只获得该描述和对应图像，并被要求点击描述的对象。然而，该数据集中的图像通常仅包含给定类别中的单个对象，虽强调了上下文信息，但忽略了对象的细节。

RefCOCO 和 RefCOCO+数据集涵盖了多张图

像及多条关于图中实例的文本描述，其中图像与真实标签（类别、边界框、像素级掩码）来自 Microsoft COCO[100] 数据集，文本描述通过 ReferItGame 收集。二者的区别在于，RefCOCO+强调的是目标对象的外观特征描述，而舍弃了方位词的使用，而 RefCOCO 并无该类限制。2 个数据集均包含一个训练集、一个验证集（val）和两个测试集（testA 和 testB）。其中，验证集包含多种场景和类别，testA 主要包含人物类目标，testB 主要包含非人物类目标，如动物、物品等。

RefCOCOg 与 RefCOCO 和 RefCOCO+相同，同样基于 Microsoft COCO[100] 数据集构建，文本描述通过 Amazon Mechanical Turk 收集。RefCOCOg 可被划分为训练集、验证集和测试集。验证集和测试集的划分方式有 google（val（G））和 umd（val（U），test（U））两种；在 umd 划分方式下，训练集、验证集和测试集的图像不存在重叠。该数据集的特点是具有更长、更复杂的表达式，因而语义更加复杂，更具挑战性。

4.2　评价指标

在 RIS 任务中，评价指标主要集中于预测掩码与真实掩码之间的重叠情况，主要包括全局交并比（overall intersection over union，Overall IoU）、平均交并比（mean intersection over union，Mean IoU）、精度百分比（Prec@X）等；REG 任务的评价指标则主要针对预测框的定位准确程度，通常使用 0.5 精度百分比（Prec@0.5）。此外，内存占用和推理速度也是评价一个方法性能的重要指标。

全局交并比定义为所有测试样本中预测掩码与真实掩码的交集总面积和并集总面积的比值。它反映了模型在整体数据集层面的分割精度，可表示为

$$f_{\text{oIoU}}=\frac{\sum_{i=1}^{N}P\cap G}{\sum_{i=1}^{N}P\cup G} \tag{4}$$

平均交并比分别计算每个样本预测掩码与真实掩码的交并比（IoU），再对所有样本取平均值。作为 RIS 的核心评价指标，它衡量了模型在样本级别的分割效果，能反映模型在不同样本上的表现是否稳定，计算公式为

$$f_{\text{mIoU}}=\frac{1}{N}\sum_{i=1}^{N}\frac{P\cap G}{P\cup G} \tag{5}$$

式中：N指测试样本总数，P表示预测结果，G表示真实标签。

精度百分比（Prec@X）表示预测掩码与真实掩码的 IoU 大于阈值 X 的样本比例。常用的阈值可由易到难设置为 0.5、0.7、0.9。

0.5 精度百分比（Prec@0.5）表示预测区域与真实区域的 IoU 大于 0.5 的样本比例。该指标强调模型是否准确找到目标对象的位置，而不注重预测框的精细边界。

模型的内存占用指模型在运行时峰值内存消耗，若占用过大，则会增加硬件资源需求，从而影响模型在不同场景下的适用性。推理速度（单位时间内处理的样本数）则决定了模型的响应延迟和吞吐量，直接影响实际应用的实时性。目前，大多数研究并未报告模型的内存占用与推理速度，这在一定程度上限制了对模型部署成本和实用性的评估。

5　主要方法结果评测与讨论

针对现有图像−文本对及无标注数据场景下典型的 RIS 与 REG 方法，在 3 个常用数据集上的结果进行总结，并对结果进行讨论。其中，RIS 采用 Mean IoU 指标，对应结果见表 3，REG 采用 Prec@0.5 指标，对应结果见表 4。

表 3 和表 4 中，$\text{KPRN}_{\text{soft+attr}}$ 表示采用软过滤策略与属性分类损失的 KPRN（详见 2.2.1 节），UniRelo+ARN 表示将方法 UniRelo 集成至 ARN 中，SSC（SS）表示方法 SSC 第 2 个阶段的预测结果。骨干网络指用于提取视觉特征的网络，ViT-G 指 Scaling vision Transformers[101] 中的结构，ViT-S/16 出自 Vision Transformer[102]，ResNet-101 为典型的 ResNet[103] 网络。ALBEF-B/16 指的是采用 ALBEF[41] 预训练权重的 ViT-B/16，CLIP-RN50、CLIP-B/32、CLIP-L/14 分别指的是采用 CLIP 预训练权重的 ResNet-50、ViT-B/32 和 ViT-L/14，M2f-Swin-B、X-VLM-Swin-B 分别指采用 Mask2-former[104] 及 X-VLM[64] 预训练权重的 Swin-B，BLIP-B/16 指采用 BLIP[68] 预训练权重的 ViT-B/16，SD-UNet 指采用 Stable Diffusion[69] 预训练权重的 U-Net 结构。符号“+”表示同时采用两种网络，例如在 UniRelo 中，模型用两种网络分别提取相同对象的不同特征，并在后续与其他特征进行交互；在 ReCLIP 中则利用两种网络作为两个分支，最终融合来自两个分支的输出分数作为结果。符号“&”的左边表示该方法生成伪标签时用到的骨干网络，右边表示利用伪标签训练的模型的骨干网络。“—”表示原论文未提及。

表 3　基于图像−文本对与无标注数据的 RIS 在 3 个数据集上的 Mean IoU 表现

Table 3　Mean IoU performance of RIS based on image-text pairs and unlabeled data across three datasets　%

标签	方法	骨干网络	RefCOCO			RefCOCO+			RefCOCOg		
			val	testA	testB	val	testA	testB	val(U)	test(U)	val(G)
图像−文本对	TSEG[36]	ViT-S/16	25.44	—	—	22.01	—	—	—	—	22.05
	SAG[39]	ViT-S/16	34.76	34.58	35.01	28.48	28.60	27.98	—	—	28.87
	TRIS[18]	CLIP-RN50&Swin-B	31.17	32.43	29.56	30.90	30.42	30.80	36.19	36.23	36.00
	Chunk[38]	ALBEF-B/16	31.06	32.30	30.11	31.28	32.11	30.13	—	—	32.88
	PPT[46]	CLIP-B/16	46.76	45.33	46.28	45.34	45.84	44.77	—	—	42.97
	QueryMatch[44]	M2f-Swin-B	59.10	59.08	58.82	39.87	41.44	37.22	—	—	43.06
	PCNet[53]	CLIP-RN50	52.20	58.40	42.10	47.90	56.50	36.20	46.80	46.90	47.30
	SSC[47]	CLIP-L/14&Swin-B	56.03	64.73	38.64	46.89	55.45	33.88	48.18	48.61	49.41
无标注	Global-Local[70]	CLIP-RN50	26.70	24.99	26.48	28.22	26.54	27.86	33.02	33.12	32.79
		CLIP-B/32	26.20	24.94	26.56	27.80	25.64	27.84	33.52	33.67	33.61
	Ref-Diff[82]	SD-U-Net	37.21	38.40	37.19	37.29	40.51	33.01	44.02	44.51	44.26
	SSC(SS) [47]	CLIP-L/14	36.95	43.77	27.97	37.68	46.24	29.31	41.41	47.18	47.57
	TAS[72]	CLIP-RN50	39.91	42.85	35.85	43.99	50.58	36.44	47.68	47.41	48.69
		CLIP-B/32	39.84	41.08	36.24	43.63	49.13	36.54	46.62	46.80	48.05
	BMS[79]	—	40.77	49.64	29.66	41.70	51.18	30.16	48.32	50.16	50.26
	Pseudo-RIS[80]	ViT-G&CLIP-RN50	41.05	48.19	33.48	44.33	51.42	35.08	45.99	46.67	46.80
	HPFD[74]	CLIP-RN50	41.66	44.64	37.80	44.80	51.49	37.52	49.24	49.15	50.02
		CLIP-B/32	41.71	43.04	38.13	44.32	49.92	37.36	48.40	47.93	49.68

表 4　基于图像−文本对与无标注数据的 REG 在 3 个数据集上的 Prec@0.5 表现

Table 4　Prec@0.5 performance of REG based on image-text pairs and unlabeled data across three datasets　%

标签	方法	骨干网络	RefCOCO			RefCOCO+			RefCOCOg		
			val	testA	testB	val	testA	testB	val(G)	val(U)	test(U)
图像−文本对	ARN[57]	ResNet-101	34.26	36.01	33.07	34.53	36.01	33.75	34.66	—	—
	CCL[62]	ResNet-101	34.78	37.64	32.59	34.29	36.91	33.56	34.92	—	—
	$KPRN_{soft+attr}$[56]	ResNet-101	36.34	35.28	37.72	37.16	36.06	39.29	36.65	—	—
	DTMR[59]	ResNet-101	38.35	39.51	37.01	38.91	39.91	37.09	42.54	—	—
	Cycle-Free[61]	ResNet-101	39.06	39.80	37.09	38.18	39.60	37.31	—	—	—
	QueryMatch[44]	M2f-Swin-B	66.02	66.00	65.48	44.76	46.72	41.50	48.47	—	—
	UniRelo+ARN[66]	CLIP-RN50+CLIP-B/32	64.69	62.01	66.46	46.41	46.94	44.71	50.50	—	—
	UniRelo+DTMR[66]	CLIP-RN50+CLIP-B/32	66.03	65.90	66.73	49.76	51.69	47.92	59.07	—	—
	CPL[67]	BLIP-B/16&ResNet-50	66.75	69.77	63.44	50.65	55.30	45.52	55.19	53.80	53.92
	PAPT[63]	X-VLM-Swin-B	68.16	77.53	58.22	68.62	77.32	57.24	—	—	—
无标注	VGDIFFZERO[95]	SD-UNet	27.95	30.34	29.11	28.39	30.79	29.79	—	33.53	33.24
	Pseudo-Q[90]	ResNet101&ResNet-50	56.02	58.25	54.13	38.88	45.06	32.13	49.82	46.25	47.44
	ReCLIP[84]	CLIP-RN50	41.53	40.78	45.55	44.53	45.88	42.87	—	57.66	56.37
		CLIP-B/32	45.77	46.99	45.24	45.34	48.45	42.71	—	56.96	56.15
		CLIP-RN50+CLIP-B/32	45.78	46.10	47.07	47.87	50.10	45.10	—	59.33	59.01
	RelVLA[86]	CLIP-B/32	48.24	48.40	49.15	45.64	47.59	42.79	—	57.60	56.64
	MCCE[89]	CLIP-RN50+CLIP-B/32	60.78	68.99	53.33	59.03	67.81	48.13	—	65.56	62.30
	SGEPG[92]	ResNet101&ResNet	69.61	72.34	65.67	55.21	60.67	46.76	—	61.33	60.61

结合表 3 和表 4 的观察结果，在图像−文本对的弱监督信号场景下，早期 RIS 和 REG 方法的骨干网络均未使用 MLLM 权重，随着 MLLM 的引入，方法性能得到显著提升，在 REG 任务中更为明显，而 RIS 的性能提升幅度较小，这可能与任务的细粒度需求有关。REG 主要依赖区域级别的跨模态对齐，而 RIS 需要像素级精准分割，涉及更复杂的视觉特征解析。因此，MLLM 虽能增强整体表现，但有效的跨模态对齐策略仍是关键。例如，对比 QueryMatch 与 Chunk 可以发现，尽管前者未使用 MLLM 预训练参数，但性能仍明显优于后者，且综合性能在单阶段的 RIS 方法中位列第 3，这归因于 QueryMatch 高效的对比学习策略，其通过 Transformer 交互学习目标对象的表征，并结合筛选策略选取高质量负类 query，从而优化图文对齐能力，实现了图像−文本特征的有效交互。相比之下，Chunk 依赖 Grad-CAM 生成热力图，并采用名词块一致性优化目标区域。然而，由于仅基于图像−文本对的弱监督信息，Chunk 缺乏精确的区域级别对齐能力，导致名词块与真实目标区域匹配不稳定，进而影响最终的定位效果。另一方面，当面临复杂的参考文本时，基于图像−文本对信号生成的响应图容易出现目标模糊或错误激活非目标区域的问题，难以保证伪标签质量。因此，多阶段方法通常结合掩码提议，并利用 MLLM（如 CLIP）进行掩码提议−文本匹配，以逐步构造或优化已构造的伪标签，供结果预测网络进一步训练。在此类方法中，PPT 采用多层次课程学习，使模型逐步从简单任务过渡到复杂任务。SSC 则通过掩码提议筛选与修正机制，先利用掩码提取器生成初步掩码提议，再结合小规模高质量预测数据集优化训练，确保了伪标签的可靠性，并实现了最优性能。另一方面，PCNet 采用不同的优化思路，利用掩码提议引导模型优化自身生成的响应图，以逐步提升预测的精度。此外，结合大语言模型解析文本，通过不同短语的激活区域间的约束，优化文本与目标区域的匹配。这一策略规避了单一掩码提议筛选带来的误差，使跨模态特征对齐更加稳定，从而实现了次优的性能。

在 REG 任务中，早期的循环前向−后向流程方法因优化目标的间接性，难以有效提升前向模块的目标定位能力。例如，当前向模块预测的区域偏离真实目标区域时，后向模块仍可能生成与参考文本相似的描述，无法对预测偏差提供有效约束。结合其性能表现，该方法逐渐被 MLLM 驱动的图文匹配方法取代，如利用 CLIP 实现更高的区域−文本匹配精度，或结合文本解析增强细粒度特征对齐能力。另一类优化策略是构造伪标签进行完全监督训练，如利用目标检测器生成候选区域，并借助 MLLM 进行跨模态特征匹配，同时结合不同的标签质量增强策略，以提升伪标签的可靠性。例如，CPL 借助 BLIP 配对候选区域与构造的描述性文本，将计算的匹配置信度分数结合到模型训练中，优化伪标签质量，并在众多方法中实现了次优的效果。另一方面，提示词微调的方法 PAPT 表现出色，其通过设计可学习的提示，引导 MLLM 关注目标对象的不同部分，以优化目标区域的理解和表征。同为提示词辅助方法，UniRelo 注重视觉模态的利用，通过显式建模候选区域的类别、颜色和空间属性，实现与文本的细粒度关系建模，从而提升区域筛选的准确性。同时，其评分机制可直接嵌入其他方法中，具备较强的灵活性。

在无标注数据场景下，由于缺乏外部监督信号，现有 RIS 和 REG 方法主要依赖预训练模型的知识，结合启发式策略，通过区域提议与文本配对进行预测或构建伪标签。因此，二者的整体框架较为相似。然而，由于任务对精细度要求的差异，二者在区域筛选策略上又有所不同。RIS 任务中的 HPFD 采用逐级提示策略，结合不同定位与分割模型，将上一级模型的输出作为下一级模型的提示，逐步优化掩码质量。这种分层级的提示机制相比于使用单一模型直接生成的掩码提议，能够更精细地捕捉目标轮廓，因而实现了最优性能。TAS 和 BMS 在区域提议与参考文本匹配策略上引入了负类构造策略，以提升目标筛选的准确性。其中，TAS 通过负类文本描述，利用 CLIP 辅助目标对象与非目标对象的文本区分能力；而 BMS 通过构造负类掩码，在视觉特征空间中强化目标的可辨性。目前，大多数区域提议筛选方法依赖于 CLIP 进行跨模态匹配，但 CLIP 在局部目标定位任务上可能存在偏差。为此，REG 中的 MCCE 通过 MLLM 生成更丰富的文本线索，增强 CLIP 处理细粒度图文匹配的能力，减少了 CLIP 在区域级图像匹配时的偏差，达到了次优的性能。

在 RIS 任务的伪标签构建方法中，Pseudo-RIS 在多项指标上仅次于 HPFD，这可能得益于其对生成文本的选择和过滤策略，确保了伪标签中文本描述的准确性和独特性。在 REG 中，SGEPG 通过场景图建模有效建立了目标对象之间的清晰

关系, 并通过歧义过滤, 减少语义混淆, 从而提升伪标签的可靠性, 达到了最优的表现。此外, 借助文本多样性增强策略, 确保伪标签的语言表达更具区分性, 有助于模型更精准地学习目标对象之间的关系。

在另一种方法中, Ref-Diff 和 VGDIFFZERO 借助生成模型 Stable Diffusion 完成任务。Ref-Diff 直接利用扩散模型的交叉注意力矩阵生成掩码提议, 并结合语法解析选取目标区域。然而, 该方法缺乏进一步的掩码优化机制, 可能导致预测掩码质量不稳定。VGDIFFZERO结合目标检测器生成区域提议, 并利用 Stable Diffusion 的去噪过程, 通过计算文本与区域的噪声预测偏差选择最终目标区域。尽管这种方式利用扩散模型的视觉−文本对齐能力, 但 Stable Diffusion 模型主要用于文本引导的图像生成, 偏向匹配全局信息, 其视觉−文本对齐能力可能难以满足 RIS 和 REG 定位任务中的对齐要求, 导致其性能相较于其他方法存在一定差距。

6　挑战与展望

出于对低标注成本的需求, 近年来, 大量基于弱监督信号的 RIS 和 REG 方法相继涌现, 在降低标注成本的同时提升了任务性能, 但与完全监督方法相比仍存在一定差距。这些方法通常借助掩码提取器、目标检测器或 MLLM, 以利用预训练模型在大规模数据中学到的先验知识来弥补监督信号的不足。本节将讨论基于图像−文本对和无标注数据的 RIS 与 REG 方法所面临的挑战, 并探讨未来发展方向。

6.1　存在的问题与挑战

在图像−文本对信号中, 模型需从文本与全局图像的关系中隐式推导目标, 而无法利用更精细的标签对预测损失直接进行校正, 这使得细粒度的特征对齐尤为困难。为缓解此问题, 一些方法基于 Transformer 架构设计模态融合模块, 或采用对比学习策略促进跨模态特征对齐。这些方法虽能捕捉复杂语义关系并支持端到端训练, 但在弱监督信号下, 图像与文本的语义层次可能不匹配, 单纯依赖 Transformer 可能导致模型过度偏向单一模态, 影响可靠的特征映射学习。同时, 数据规模增大会导致计算资源需求激增。另一方面, 若图像与文本的特征分布差异较大(如跨领域数据), 对比学习可能难以稳定匹配不同模态的特征。因此, 在弱监督场景下, 如何设计高效的跨模态交互策略, 使图文特征细粒度对齐, 仍是 RIS 和 REG 任务的关键挑战。

为缓解监督信号不足的问题, 一些方法(如 TRIS[18]、Pseudo-Q[90] 等)通过离线构建伪标签为模型提供完全监督信号, 其核心在于借助预训练模型, 结合不同标签生成与筛选策略, 以提升伪标签精度。然而, 伪标签的质量直接影响模型训练效果, 尽管现有方法采取优化措施, 但仍难以达到真实标签的质量。此外, 离线生成伪标签需要额外的计算资源和时间成本, 相比端到端训练, 可能带来更高的计算开销。因此, 如何在提升伪标签质量的同时降低计算成本, 也是一个关键挑战。

在无标注的场景下, 现有 RIS 和 REG 方法主要依赖 MLLM 及掩码提取器(目标检测器)。MLLM 具备良好的图文特征对齐能力, 还能对参考文本进行深度解析, 增强输入文本的语义信息。然而, 由于训练任务的差异, MLLM 更擅长全局图像−文本匹配, 而在 RIS 和 REG 任务所需的区域−文本匹配上能力有限, 同时可能产生“检索幻觉”等偏差。此外, MLLM 的空间位置感知能力不足, 而当前多数方法中, 掩码或边界框提议被独立地进行筛选, 区域间的空间关系容易被忽略, 进一步加大了对空间理解的难度。同时, 掩码提取器或目标检测器生成的区域提议质量也直接影响最终预测性能。因此, 如何缓解现有模型在这两类任务上的缺陷, 是 RIS 与 REG 任务不容忽视的问题。

方法的实用性是另一大挑战。现有多数方法仅在标准数据集上进行验证, 而在现实场景中, 模型可能处理未见类别的目标, 这要求模型须具备较强的自适应能力以分割或定位未知的目标。同时, 实时性是实际应用的关键需求, 然而当前研究主要侧重提升预测精度, 较少关注推理速度与计算效率。例如, 在 REG 任务中, MCCE[89] 简要对比了自身及 ReCLIP[84] 等方法在 RefCOCO、RefCOCO+和 RefCOCOg 数据集上的推理速度, 然而并未指明实验所用设备。结果显示, 在单个样本推理上, ReCLIP 的速度为 0.67 s, MCCE 为 1.58 s。这表明 MCCE 尽管在定位准确性上优于 ReCLIP, 但推理速度却慢了一倍以上。当面临大批量数据的任务时, ReCLIP 所需的总处理时间可能明显少于 MCCE。因此, 需要考虑如何平衡精度与效率, 提升方法在动态环境中的适应能力, 以满足更广泛的应用需求。

6.2　未来的发展方向

未来, 弱监督场景下的 RIS 和 REG 方法或将

围绕跨模态细粒度特征对齐、伪标签优化、预训练模型适配、精细化区域提取、低资源场景泛化等方向发展。

对于跨模态细粒度特征对齐，未来研究不仅在于提升跨模态引导机制的灵活性，同时有必要探索动态对齐策略，使模型可以根据不同语言表达，灵活调整视觉注意力，如现有研究中的 PAPT[63] 方法，通过文本信息动态调整视觉注意力，使图像特征的提取更符合语言描述。此外，还可引入结构化对齐方式，如细粒度建模文本中语义成分与图像中特定对象乃至特定区域的关系，从而缩小跨模态差距。

在伪标签优化上，提高质量并降低计算成本是重要挑战。未来可从两个方向推进：1）通过跨模态一致性约束及迭代优化，来剔除噪声，使伪标签在多轮更新中逐渐收敛至更高精度；2）探索与主模型联合训练的端到端伪标签生成机制，如采用教师−学生模型，通过模型自身的预测迭代改进伪标签，减少离线生成带来的计算开销。

在预训练模型适配方面，现有方法广泛依赖 MLLM 及掩码提取器（目标检测器），如 SSC[47]、HPFD[74] 等结合不同模型，通过逐级提示对掩码提议进行优化，从而提高预测精度，但引入多个模型容易增加计算开销。此外，MLLM 在弱监督任务中的应用仍具潜力，例如 RelVLA[86]、MCCE[89] 等方法已探索利用 MLLM 解析参考文本或生成额外描述，以增强输入语义信息。但 MLLM 在区域级特征对齐及空间感知能力上仍有不足，已有研究（如 BASP[87]、MCCE[89]、ReCLIP[84] 等）通过辅助文本增强策略或空间解析模块进行补充。未来，研究或将从轻量化适配与知识迁移优化两个方向改进。例如，前者通过设计专门针对局部区域理解的适配模块，使预训练模型能在 RIS 或 REG 任务中保持高效而不依赖冗余模块，从而避免引入过多参数；后者可探索如何在预训练知识迁移过程中增强区域级别的语义与空间感知，从而提高模型在目标定位和分割中的适配性。

在精细化区域提取方向上，未来可采用更成熟的掩码提取器（目标检测器），并结合自身方法预测（如基于 Transformer 的跨模态融合策略）对提议的目标区域进行空间约束（如 PCNet[53] 的做法），通过显式建模区域间的语义关系，使模型获得更精细的预测结果。还可结合语言提示动态调整区域的空间边界，而非依赖固定的检测结果，从而实现粗粒度候选到精细化掩码的渐进式优化，提升最终分割与定位的准确性。

最后，低资源场景下的泛化能力仍是挑战。未来不仅能借助领域自适应或数据高效训练等技术，还可通过合成数据增强。不局限于 PPT[46] 合成样本的做法，还可通过生成模型获得“语义变体描述”，提升语言多样性以增强模型鲁棒性。此外，可结合数据高效训练方法，发展参数高效微调或少样本适配策略，使模型能在有限数据下快速适应新场景。

7　结束语

弱监督学习可显著降低模型对标注信息的需求，缓解 RIS 与 REG 的高昂标注成本问题，因此其研究具有重要意义。本文首先阐述了弱监督的 RIS 与 REG 的背景与重要性，并介绍了常见的弱监督信号，其中图像−文本对与无标注数据的获取成本最低，与之相关的研究更受关注。随后，围绕基于图像−文本对及无标注的弱监督 RIS 与 REG 方法展开分析，并在相同监督场景下对其进行分类与比较：在图像−文本对标注下，根据训练特点将 RIS 分为单阶段与多阶段方法，根据方法特性将 REG 分为“循环前向−后向流程”及“非循环前向−后向流程”方法；在无标注场景，依据模型方法特性统一将 RIS 和 REG 方法归为 3 类，即“区域提议生成−图文匹配筛选”、伪标签构造、借助生成模型的方法。接着，概述了相关数据集与评价指标，并在公开数据集上对典型方法的性能进行评估，同时探讨其挑战和发展方向。结果表明，无标注数据的方法性能普遍不如图像−文本对标注的方法，此外，MLLM 的引入可有效辅助各类方法性能的提升，但面临着自身能力限制及任务迁移时带来的问题。另外，面对现实应用场景的复杂需求，方法的实时性、精确度和泛化性等仍是重大挑战。未来，关于弱监督的 RIS 和 REG 或将基于更成熟的 MLLM 和掩码提取器（目标检测器）进一步发展，并就兼顾模型的精度、推理效率及泛化性提出一系列策略。

参考文献：

[1]　HU R, ROHRBACH M, DARRELL T. Segmentation from natural language expressions[C]//Computer Vision–ECCV 2016: 14th European Conference. Amsterdam: Springer, 2016: 108−124.

[2]　MAO Junhua, HUANG J, TOSHEV A, et al. Generation and comprehension of unambiguous object descriptions[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 11−20.

[3] YU Licheng, POIRSON P, YANG Shan, et al. Modeling context in referring expressions[C]//Computer Vision– ECCV. Cham: Springer International Publishing, 2016: 69−85.

[4] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2980−2988.

[5] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence. Montreal: IEEE, 2017: 1137−1149.

[6] CHEN Haonan, TAN Hao, KUNTZ A, et al. Enabling robots to understand incomplete natural language instructions using commonsense reasoning[C]//2020 IEEE International Conference on Robotics and Automation. Paris: IEEE, 2020: 1963−1969.

[7] GU Jing, STEFANI E, WU Qi, et al. Vision-and-language navigation: a survey of tasks, methods, and future directions[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: USAACL, 2022: 7606−7623.

[8] HU Yutao, WANG Qixiong, SHAO Wenqi, et al. Beyond one-to-one: rethinking the referring image segmentation[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 4044−4054.

[9] YU Licheng, LIN Zhe, SHEN Xiaohui, et al. MAttNet: modular attention network for referring expression comprehension[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1307−1315.

[10] YANG Li, XU Yan, YUAN Chunfeng, et al. Improving visual grounding with visual-linguistic verification and iterative reasoning[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 9489−9498.

[11] 邱爽, 赵耀, 韦世奎. 图像指代分割研究综述[J]. 信号处理, 2022, 38(6): 1144−1154.
QIU Shuang, ZHAO Yao, WEI Shikui. A survey of referring image segmentation[J]. Journal of signal processing, 2022, 38(6): 1144−1154.

[12] 项伟康, 周全, 崔景程, 等. 基于深度学习的弱监督语义分割方法综述[J]. 中国图象图形学报, 2024, 29(5): 1146−1168.
XIANG Weikang, ZHOU Quan, CUI Jingcheng, et al. Weakly supervised semantic segmentation based on deep learning[J]. Journal of image and graphics, 2024, 29(5): 1146−1168.

[13] 陈震元, 王振东, 宫辰. 图像级标记弱监督目标检测综述[J]. 中国图象图形学报, 2023, 28(9): 2644−2660.
CHEN Zhenyuan, WANG Zhendong, GONG Chen. Image-level labeled weakly supervised object detection: a survey[J]. Journal of image and graphics, 2023, 28(9): 2644−2660.

[14] 李文生, 张菁, 卓力, 等. 基于 Transformer 的视觉分割技术进展[J]. 计算机学报, 2024, 47(12): 2760−2782.
LI Wensheng, ZHANG Jing, ZHUO Li, et al. Overview of Transformer-based visual segmentation techniques[J]. Chinese journal of computers, 2024, 47(12): 2760−2782.

[15] 祁磊, 于沛泽, 高阳. 弱监督场景下的行人重识别研究综述[J]. 软件学报, 2020, 31(9): 2883−2902.
QI Lei, YU Peize, GAO Yang. Research on weak-supervised person re-identification[J]. Journal of software, 2020, 31(9): 2883−2902.

[16] 蒋弘毅, 王永娟, 康锦煜. 目标检测模型及其优化方法综述[J]. 自动化学报, 2021, 47(6): 1232−1255.
JIANG Hongyi, WANG Yongjuan, KANG Jinyu. A survey of object detection models and its optimization methods [J]. Acta automatica sinica, 2021, 47(6): 1232−1255.

[17] SHEN Wei, PENG Zelin, WANG Xuehui, et al. A survey on label-efficient deep image segmentation: bridging the gap between weak supervision and dense prediction[J]. IEEE transactions on pattern analysis and machine intelligence, 2023, 45(8): 9284−9305.

[18] LIU Fang, LIU Yuhao, KONG Yuqiu, et al. Referring image segmentation using text supervision[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 22067−22077.

[19] FENG Guang, ZHANG Lihe, HU Zhiwei, et al. Learning from box annotations for referring image segmentation[J]. IEEE transactions on neural networks and learning systems, 2022, 35(3): 3927−3937.

[20] LI Hui, SUN Mingjie, XIAO Jimin, et al. Fully and weakly supervised referring expression segmentation with end-to-end learning[J]. IEEE transactions on circuits and systems for video technology, 2023, 33(10): 5999−6012.

[21] ZANG Ying, CAO Runlong, FU Chenglong, et al. RES-Match: Referring expression segmentation in a semi-supervised manner[J]. Information sciences, 2025, 694: 121709.

[22] QU Mengxue, WU Yu, WEI Yunchao, et al. Learning to segment every referring object point by point[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 3021−3030.

[23] NAG S, GOSWAMI K, KARANAM S. SafaRi: adaptive sequence transformer for weakly supervised referring expression segmentation[C]//Computer Vision–EC-

CV 2024. Cham: Springer Nature Switzerland, 2024: 485−503.

[24] HUANG Minglang, ZHOU Yiyi, LUO Gen, et al. Towards omni-supervised referring expression segmentation[C]//2024 IEEE International Conference on Multimedia and Expo. Niagara Falls: IEEE, 2024: 1−6.

[25] SUN Jiamu, LUO Gen, ZHOU Yiyi, et al. RefTeacher: a strong baseline for semi-supervised referring expression comprehension[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 19144−19154.

[26] YANG D, JI J, MA Y, et al. SAM as the guide: mastering pseudo-label refinement in semi-supervised referring expression segmentation[C]//International Conference on Machine Learning. Toronto: PMLR, 2024: 56139−56155.

[27] HU Ronghang, ROHRBACH M, ANDREAS J, et al. Modeling relationships in referential expressions with compositional modular networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 4418−4427.

[28] ZHU Haidong, SADHU A, ZHENG Zhaoheng, et al. Utilizing every image object for semi-supervised phrase grounding[C]//2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2021: 2210−2219.

[29] ZHANG Dingwen, HAN Junwei, CHENG Gong, et al. Weakly supervised object localization and detection: a survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 44(9): 5866−5885.

[30] XIAO Fanyi, SIGAL L, LEE Y J. Weakly-supervised visual grounding of phrases with linguistic structures[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 5253−5262.

[31] DONG T. Weakly supervised learning from referring expression: Challenge and directions[D]. Urbana-Champaign: University of Illinois at Urbana-Champaign, 2018.

[32] DIETTERICH T G, LATHROP R H, LOZANO-PÉREZ T. Solving the multiple instance problem with axis-parallel rectangles[J]. Artificial intelligence, 1997, 89(1/2): 31−71.

[33] CINBIS R G, VERBEEK J, SCHMID C. Multi-fold MIL training for weakly supervised object localization [C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 2409−2416.

[34] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015−04−10) [2024−12−12]. https://arxiv.org/pdf/1409.1556.

[35] GRAVES A. Long short-term memory[M/OL]. Supervised sequence labelling with recurrent neural networks. Berlin: Springer Berlin Heidelberg, 2012: 37−45. [2024−12−12]. https://doi.org/10.1007/978-3-642-24797-2_4.

[36] STRUDEL R, LAPTEV I, SCHMID C. Weakly-supervised segmentation of referring expressions[EB/OL]. (2022−05−02) [2024−12−12]. https://arxiv.org/abs/2205.04725.

[37] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc, 2017: 6000−6010.

[38] LEE J, LEE S, NAM J, et al. Weakly supervised referring image segmentation with intra-chunk and inter-chunk consistency[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 21813−21824.

[39] KIM D, KIM N, LAN Cuiling, et al. Shatter and gather: learning referring image segmentation with text supervision[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 15501−15511.

[40] LOCATELLO F, WEISSENBORN D, UNTERTHINER T, et al. Object-centric learning with slot attention[J]. Advances in neural information processing systems, 2020, 33: 11525−11538.

[41] LI J, SELVARAJU R, GOTMARE A, et al. Align before fuse: Vision and language representation learning with momentum distillation[J]. Advances in neural information processing systems, 2021, 34: 9694−9705.

[42] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[J]. International journal of computer vision, 2020, 128(2): 336−359.

[43] ARBELÁEZ P, PONT-TUSET J, BARRON J, et al. Multiscale combinatorial grouping[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 328−335.

[44] CHEN Shengxin, LUO Gen, ZHOU Yiyi, et al. QueryMatch: a query-based contrastive learning framework for weakly supervised visual grounding[C]//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne: ACM, 2024: 4177−4186.

[45] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//Computer Vision–ECCV 2020. Cham: Springer International Publishing, 2020: 213−229.

[46] DAI Qiyuan, YANG Sibei. Curriculum point prompting for weakly-supervised referring image segmentation [C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 13711−13722.

[47] EIRAS F, OKSUZ K, BIBI A, et al. Segment, select, correct: a framework for weakly-supervised referring segmentation[EB/OL]. (2023−10−23) [2024−12−12]. https://arxiv.org/abs/2310.13479.

[48] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 3992−4003.

[49] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. Cambridge: PMLR, 2021: 8748−8763.

[50] DENG Jia, DONG Wei, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 248−255.

[51] LIU Shilong, ZENG Zhaoyang, REN Tianhe, et al. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection[C]//Computer Vision–ECCV 2024. Cham: Springer Nature Switzerland, 2024: 38−55.

[52] WANG Xinlong, YU Zhiding, DE MELLO S, et al. FreeSOLO: learning to segment objects without annotations[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 14156−14166.

[53] YANG Zaiquan, LIU Yuhao, LIN Jiaying, et al. Boosting weakly supervised referring image segmentation via progressive comprehension[C]//The Thirty-eighth Annual Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc, 2024: 93213−93239.

[54] JIANG A Q, SABLAYROLLES A, MENSCH A, et al. Mistral 7B[EB/OL]. (2023−10−10) [2024−12−12]. https://arxiv.org/abs/2310.06825.

[55] ROHRBACH A, ROHRBACH M, HU Ronghang, et al. Grounding of textual phrases in images by reconstruction[C]//Computer Vision–ECCV 2016. Cham: Springer International Publishing, 2016: 817−834.

[56] LIU Xuejing, LI Liang, WANG Shuhui, et al. Knowledge-guided pairwise reconstruction network for weakly supervised referring expression grounding[C]//Proceedings of the 27th ACM International Conference on Multimedia. Nice: ACM, 2019: 539−547.

[57] LIU Xuejing, LI Liang, WANG Shuhui, et al. Adaptive reconstruction network for weakly supervised referring expression grounding[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 2611−2620.

[58] ZHAO Fang, LI Jianshu, ZHAO Jian, et al. Weakly supervised phrase localization with multi-scale anchored transformer network[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 5696−5705.

[59] SUN Mingjie, XIAO Jimin, LIM E G, et al. Discriminative triad matching and reconstruction for weakly referring expression grounding[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(11): 4189−4195.

[60] WANG Ning, SONG Yibing, MA Chao, et al. Unsupervised deep tracking[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 1308−1317.

[61] SUN Mingjie, XIAO Jimin, LIM E G, et al. Cycle-free weakly referring expression grounding with self-paced learning[J]. IEEE transactions on multimedia, 2021, 25: 1611−1621.

[62] ZHANG Zhu, ZHAO Zhou, LIN Zhijie, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding[J]. Advances in neural information processing systems, 2020, 33: 18123−18134.

[63] ZHAO Chenlin, YE Jiabo, SONG Yaguang, et al. Part-aware prompt tuning for weakly supervised referring expression grounding[C]//MultiMedia Modeling. Cham: Springer Nature Switzerland, 2024: 489−502.

[64] ZENG Yan, ZHANG Xinsong, LI Hang. Multi-grained vision language pre-training: aligning texts with visual concepts[C]//International Conference on Machine Learning. Baltimore: PMLR, 2022: 25994−26009.

[65] JIA Menglin, TANG Luming, CHEN B C, et al. Visual prompt tuning[C]//Computer Vision-ECCV 2022. Cham: Springer Nature Switzerland, 2022: 709−727.

[66] ZHANG Panpan, LIU Meng, SONG Xuemeng, et al. Universal relocalizer for weakly supervised referring expression grounding[J]. ACM transactions on multimedia computing, communications, and applications, 2024, 20(7): 1−23.

[67] LIU Yang, ZHANG Jiahua, CHEN Qingchao, et al. Confidence-aware pseudo-label learning for weakly supervised visual grounding[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 2816−2826.

[68] LI Junnan, LI Dongxu, XIONG Caiming, et al. BLIP: bootstrapping language-image pre-training for unified

vision-language understanding and generation[C]//International Conference on Machine Learning. Baltimore: PMLR, 2022: 12888−12900.

[69] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 10674−10685.

[70] YU S, SEO P H, SON J. Zero-shot referring image segmentation with global-local context features[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 19456−19465.

[71] LAROCHELLE H, ERHAN D, BENGIO Y. Zero-data learning of new tasks[C]//Proceedings of the 23rd National Conference on Artificial Intelligence-Volume 2. Chicago: AAAI Press, 2008: 646−651.

[72] SUO Yucheng, ZHU Linchao, YANG Yi. Text augmented spatial aware zero-shot referring image segmentation[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: USAACL, 2023: 1032−1043.

[73] LI J, LI D, SAVARESE S, et al. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models[C]//International Conference on Machine Learning. Honolulu: PMLR, 2023: 19730−19742.

[74] LI Changlong, ZHUANG Jiedong, HU Jiaqi, et al. Zero-shot referring image segmentation with hierarchical prompts and frequency domain fusion[M]//PRICAI 2024: Trends in Artificial Intelligence. Singapore: Springer Nature Singapore, 2024: 215−228.

[75] HUANG Xinyu, HUANG Yijie, ZHANG Youcai, et al. Open-set image tagging with multi-grained text supervision[EB/OL]. (2023−11−16) [2024−12−12]. https://arxiv.org/abs/2310.15200.

[76] KE L, YE M, DANELLJAN M, et al. Segment anything in high quality[J]. Advances in neural information processing systems, 2023, 36: 29914−29934.

[77] DA CUNHA A L, ZHOU Jianping, DO M N. The nonsubsampled contourlet transform: theory, design, and applications[J]. IEEE transactions on image processing, 2006, 15(10): 3089−3101.

[78] CHEN C F, HSIAO C H. Haar wavelet method for solving lumped and distributed-parameter systems[J]. IEE proceedings - control theory and applications, 1997, 144(1): 87−94.

[79] LI Wenhui, PANG Chao, NIE Weizhi, et al. Bidirectional mask selection for zero-shot referring image segmentation[J]. IEEE transactions on circuits and systems for video technology, 2024, 35(1): 911−921.

[80] YU S, SEO P H, SON J. Pseudo-RIS: distinctive pseudo-supervision generation for referring image segmentation[C]//Computer Vision–ECCV 2024. Cham: Springer Nature Switzerland, 2024: 18−36.

[81] YU Jiahui, WANG Zirui, VASUDEVAN V, et al. CoCa: contrastive captioners are image-text foundation models[EB/OL]. (2022−06−14)[2024−12−12]. https://arxiv.org/abs/2205.01917.

[82] NI Minheng, ZHANG Yabo, FENG Kailai, et al. Ref-diff: zero-shot referring image segmentation with generative models[EB/OL]. (2023−09−01)[2024−12−12]. https://arxiv. org/abs/2308.16777.

[83] SHI Hengcan, HAYAT M, CAI Jianfei. Unpaired referring expression grounding via bidirectional cross-modal matching[J]. Neurocomputing, 2023, 518: 39−49.

[84] SUBRAMANIAN S, MERRILL W, DARRELL T, et al. ReCLIP: a strong zero-shot baseline for referring expression comprehension[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: USAACL, 2022: 5198−5215.

[85] HONNIBAL M, JOHNSON M. An improved non-monotonic transition system for dependency parsing [C]//Conference on Empirical Methods in Natural Language Processing, EMNLP 2015. Lisboa: Association for Computational Linguistics (ACL), 2015: 1373−1378.

[86] HAN Zeyu, ZHU Fangrui, LAO Qianru, et al. Zero-shot referring expression comprehension via structural similarity between images and captions[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 14364−14375.

[87] WANG Hanyao, ZHAN Yibing, LIU Liu, et al. Towards alleviating text-to-image retrieval hallucination for CLIP in zero-shot learning[EB/OL]. (2024−06−27)[2024−12−12]. https://arxiv.org/abs/2402.18400.

[88] FLORIDI L, CHIRIATTI M. GPT-3: its nature, scope, limits, and consequences[J]. Minds and machines, 2020, 30(4): 681−694.

[89] QIU Heqian, WANG Lanxiao, ZHAO Taijin, et al. MCCE-REC: MLLM-driven cross-modal contrastive entropy model for zero-shot referring expression comprehension[J]. IEEE transactions on circuits and systems for video technology, 2025, 35(1): 754−768.

[90] JIANG Haojun, LIN Yuanze, HAN Dongchen, et al. Pseudo-Q: generating pseudo language queries for visual grounding[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 15492−15502.

[91] ANDERSON P, HE Xiaodong, BUEHLER C, et al. Bot-

tom-up and top-down attention for image captioning and visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 6077−6086.

[92] WU Cantao, CAI Yi, LI Liuwu, et al. Scene graph enhanced pseudo-labeling for referring expression comprehension[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: USAACL, 2023: 11978−11990.

[93] ZHANG Ao, YAO Yuan, CHEN Qianyu, et al. Fine-grained scene graph generation with data transfer[C]// Computer Vision−ECCV 2022. Cham: Springer Nature Switzerland, 2022: 409−424.

[94] LIN S, HILTON J, EVANS O. Teaching models to express their uncertainty in words[J/OL]. Transactions on Machine Learning Research. [2024−12−12]. https://openreview.net/forum?id=8s8K2UZGTZ.

[95] LIU Xuyang, HUANG Siteng, KANG Yachen, et al. VGDIFFZERO: text-to-image diffusion models can be zero-shot visual grounders[C]//ICASSP 2024−2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul: IEEE, 2024: 2765−2769.

[96] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. Cham: Springer International Publishing, 2015: 234−241.

[97] NAGARAJA V K, MORARIU V I, DAVIS L S. Modeling context between objects for referring expression understanding[C]//Computer Vision–ECCV 2016. Cham: Springer International Publishing, 2016: 792−807.

[98] KAZEMZADEH S, ORDONEZ V, MATTEN M, et al. ReferItGame: referring to objects in photographs of natural scenes[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: USAACL, 2014: 787−798.

[99] GRUBINGER M, CLOUGH P D, MÜLLER H, et al. The IAPR TC-12 benchmark: a new evaluation resource for visual information systems[C]//International workshop onto Image. Genoa: LREC, 2006: 13−23.

[100] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Computer Vision–ECCV 2014. Cham: Springer International Publishing, 2014: 740−755.

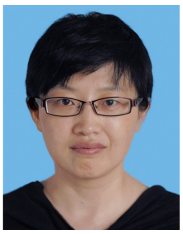
[101] ZHAI Xiaohua, KOLESNIKOV A, HOULSBY N, et al. Scaling vision transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 1204−1213.

[102] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[C/OL]. International Conference on Learning Representations. (2021−06−03) [2024−12−12]. https://openreview.net/forum?id=YicbFdNTTy.

[103] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[104] CHENG Bowen, MISRA I, SCHWING A G, et al. Masked-attention mask transformer for universal image segmentation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 1280−1289.

作者简介：



张磊，教授，中国人工智能学会认知系统与信息处理专委会委员，主要研究方向为机器学习、计算机视觉。主持国家自然科学基金项目 4 项，近 5 年发表学术论文 30 余篇（包括研究领域的顶刊和顶会）。E-mail：zhanglei@gdupt.edu.cn。

黄咏秋，硕士研究生，主要研究方向为计算机视觉。E-mail：smile_n_n@163.com。

李欣，副教授，主要研究方向为信号及图像处理、计算机视觉与机器人。E-mail：lixin@gdupt.edu.cn。