RefCLIP：一个负责弱监督参考表达理解的通用教师

雷金[12] *，罗将军[1]*周依依[12]孙小帅[12] †、姜灌南[3]安南舒[3]、纪荣荣[12]

[1]中国教育部多媒体可信感知与高效计算重点实验室，厦门大学，第361005页。中国

[2]厦门大学人工智能研究所，第361005页。中国

[3]当代安培瑞斯科技有限公司智能制造部。有限公司（CATL）。

{kingts, luogen}@stu.xmu.edu.cn,{zhouyiyi, xssun, rrji}@xmu.edu.cn,{jianggn, shuan01}@catl.com

2023年IEEE/CVF计算机视觉和模式识别会议（CVPR）| 979-8-3503-0129-8/23/$31.00©2023年IEEE | DOI：10.1

摘要

引用表达式理解（REC）是一项基于表达式建立引用基础的任务，它的开发受到昂贵的实例级注释的极大限制。现有的弱监督方法大多是基于两阶段检测网络，其计算成本很高。在本文中，我们采用高效的单级检测器，提出了一种新的弱监督模型RefCLIP。具体来说，RefCLIP将弱监督REC重新定义为一个锚定-文本匹配问题，可以避免现有方法中复杂的后处理。为了实现弱监督学习，我们引入了基于锚定的对比损失，通过大量的锚定-文本对来优化RefCLIP。在RefCLIP的基础上，我们进一步为现有的REC模型提出了第一个模型不可知的弱监督训练方案，其中RefCLIP作为一个成熟的教师，生成用于教学普通REC模型的伪标签。经过精心的设计该方案甚至可以帮助现有的REC模型获得比RefCLIP更好的弱监督性能。，TransVG和SimREC。为了验证我们的方法，我们对四个REC基准测试进行了广泛的实验。e.，RefCOCO，RefCOCO+，RefCOCOg和参考游戏。实验结果不仅报告了我们比现有的弱监督模型的显著性能提高。，+在RefCOCO上的24.87%，但也显示了5倍的推理速度。项目： https://refclip.github。io.

1.介绍

参考表达式理解（REC），也称为视觉接地[5,16]，目的是基于参考表达式在图像中定位目标实例

*同样的贡献。

沙痂通讯作者。

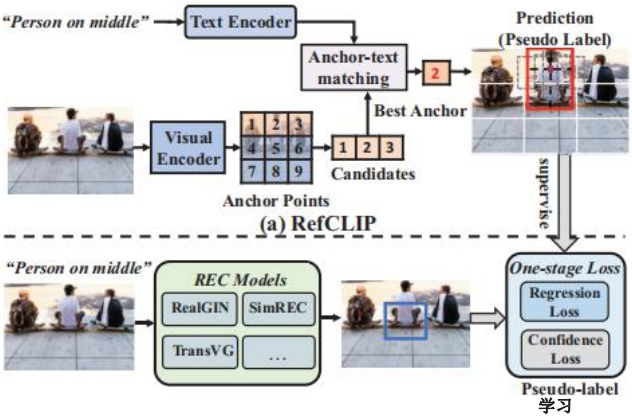


图1。所提出的RefCLIP和弱监督训练方案的说明。RefCLIP通过锚定-文本匹配从YOLOv3中选择目标边界框，通过基于锚定的对比学习进行优化。我们的培训方案使用RefCLIP作为一个成熟的教师来监督常见的REC模型，这不需要网络修改。

Sion[25-27、42、48]。作为一项跨模态识别任务，REC并不局限于一组固定的对象类别，而且在理论上能够进行任何开放式的检测[45]。这些吸引人的特性使REC越来越受到计算机视觉社区的关注[25,28,45-48]。然而，昂贵的实例级注释长期以来一直困扰着它的开发。

为此，弱监督REC模型的研究进展，该模型旨在学习仅基于语言信息[7,38,43]的检测。具体来说，现有的方法将Faster-RCNN [37]等两阶段对象检测器扩展到弱监督的REC模型。在方法方面，他们将REC视为一个区域-文本排序问题，首先通过Faster-RCNN提取图像的显著区域，然后通过跨模态匹配进行排序。为了实现弱监督训练，他们只使用表达式作为supervi-

979-8-3503-0129-8/23/$31.00©2023IEEE2681
DOI 10.1109/CVPR52729.2023.00263

通过语义重构[19,20,38]或跨模态对比学习[7,43]对排序模块进行优化。然而，由于使用了快速的-RCNN，这些方法在推理速度方面往往较差。

为了克服这些限制，我们对弱监督的REC采用了一期检测器。与FasterRCNN相比，YOLOv3 [36]等单级检测器在效率上具有明显的优势，但难以将其直接应用于现有的弱监督方案。最重要的是，现有的单级检测器[17,36]根据最后几个卷积层的特征来预测边界盒，也称为锚点[36]。在多尺度检测方面，将为一幅图像预测出数千个边界盒，因此将它们转换为区域特征变得更加耗时[1].然而，我们注意到卷积特征的接受域将比它们所代表的[29]的实际区域大得多，这表明单级检测器中的一个锚点可能包含足够的信息进行识别。

基于上述观察结果，我们将弱监督REC定义为一个锚定-文本匹配问题，并提出了一种新的弱监督模型RefCLIP。*具体来说，我们更改任务定义，其中检测到的区域是锚点有目标边界框的参考。*在这种情况下，我们可以直接对锚点进行排序，而不需要进行复杂的后处理，如ROI池化和NMS [37]。为了实现弱监督学习，RefCLIP执行基于锚的图像间和图像内部的对比学习，从而通过大量的锚-文本对学习视觉-语言对齐。值得注意的是，这种对比学习方案在负样本增强方面也表现出了优越的灵活性，这不受批处理大小的限制。

在本文中，我们还关注了弱监督REC的模型不可知的训练方案。包括RefCLIP在内，所有现有的解决方案都是特定于模型的，不能直接推广到现有的监督REC模型[5,25,42,45]。为此，我们进一步提出了第一个模型不可知的REC弱监督训练方案。具体来说，我们使用RefCLIP作为教师来制作伪标签，即。，边界框，以监督常见的REC模型。同时，我们还通过EMA [39]和数据增强[13]，缓解了伪标签噪声引起的确认偏差[1]。在该方案中，现有的REC模型可以不进行任何修改而进行弱训练，这使得我们的工作与现有的模型有很大的不同[7,18-20,38]。

为了验证所提出的RefCLIP和弱监督训练方案，我们在四个REC基准上进行了广泛的实验。*e.*，RefCOCO [32]，RefCOCO+ [32]，RefCOCOg [30]和参考游戏[10]，以及与一些最新的弱监督REC模型进行比较[18、22、38、41]。我们将我们的训练方案应用于几个具有代表性的REC模型，包括RealGIN [45]、TransVG [5]和SimREC [25]。实验结果表明，与现有的弱监督REC模型相比，我们的RefCLIP有明显的性能提高。*g.*，RefCOCO的+为21.25%。同时，通过我们的精心设计，这个建议训练方案甚至可以帮助这些REC模型获得弱监督REC的新的SOTA性能。总之，我们的主要贡献有三方面：

- 我们提出了一种新的单阶段对比模型RefCLIP，该模型通过基于锚点的跨模态对比学习实现了弱监督的REC，并显著提高了推理速度5倍。
- 我们为普通REC模型提出了第一个通用的弱监督训练方案，该方案可以使用RefCLIP生成的伪标签有效地提高任何REC模型。
- 所提出的RefCLIP在四个基准上优于现有的分配，我们的训练方案也有助于以前的REC模型获得新的弱监督的SOTA性能。

2.相关工作

.1.2.参考表达式理解法。

参考表达理解（REC）[26,42,45]，也称为视觉接地[5,16]或短语接地[6]，目的是基于给定的参考表达式在图像中的目标对象。REC的方法可以分为两类，i。*e.*，两阶段和阶段。两阶段方法[16,21,42]首先使用Faster-RCNN [37]等检测网络生成一组候选区域，然后进行区域-文本排序，选择目标区域。近年来，单阶段推理方法[14、24、26、45、48]因其较高的推理速度和优越的性能而受到越来越多的关注。早期的单阶段方法[26,45]主要由浅层多模态融合层组成。受到变压器[40]的巨大成功的启发，最近的研究人员[5,48]求助于深度反式-REC以前的架构。

2.2.弱监督的参考表达理解。

与完全监督的REC相比，弱超-由于缺乏box注释，被建议的REC更具挑战性。大多数现有的方法[7,19,20,22,38,41,43]都是由两阶段监督REC模型驱动的，并将弱监督REC表述为一个区域-文本排序问题。在这些方法中，主要的困难在于

[1]通过置信度滤波，这种处理仍然需要对COCO图像进行约26.6%的额外计算。

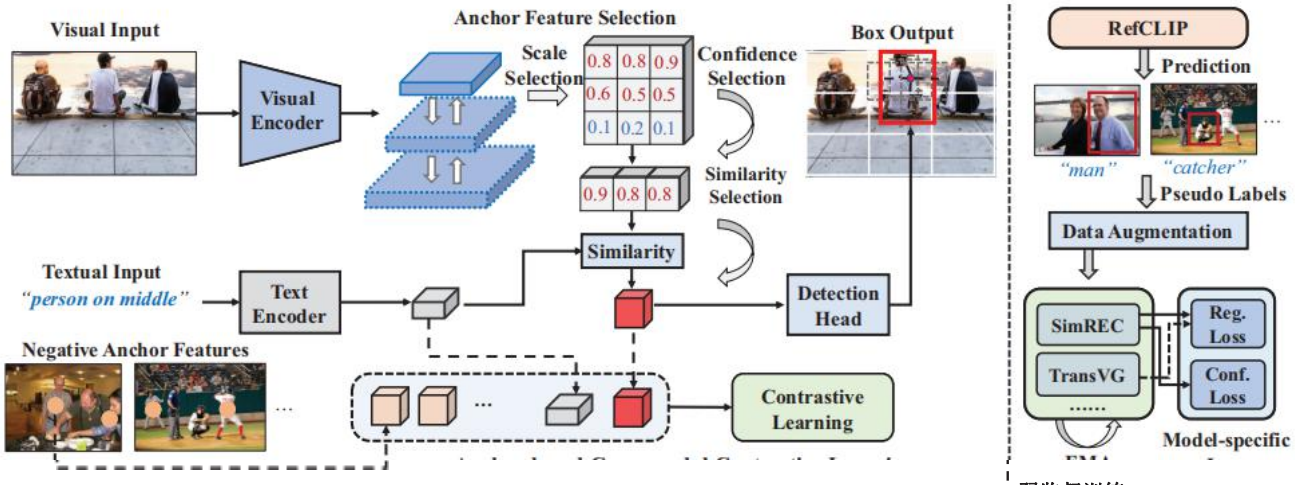


图2。提出的RefCLIP框架（左）和弱监督训练方案（右）。在RefCLIP中，图像和表达式首先由视觉和文本编码器进行处理。然后，RefCLIP过滤低值的锚，并返回最佳匹配的锚，用于边界框预测。RefCLIP通过基于锚定的对比学习进行弱监督。在我们的弱监督训练方案中，RefCLIP作为一个成熟的教师，为普通REC模型提供伪标签的训练。

如何从图像-文本对中提供有效的监督信号。为了解决这一问题，研究人员采用了诸如句子重建[19,20,38]和对比学习[7,43]等方法。特别是，句子重构选择排名得分最高的区域来重构输入表达式。与句子重构相比，基于对比学习的方法[7,43]从选定的区域和表达式中构建正、负样本对，并计算InfoNCE损失[34]。我们还注意到，[44]的早期工作很少探索弱监督REC的单阶段模型，但它们的性能仍然不如两阶段模型。与这些方法不同，RefCLIP是一个单阶段模型，具有创新的弱监督公式，i。e.，锚文本匹配。在RefCLIP的基础上，我们提出了一种新的弱监督训练方案，即伪标签学习，该方案适用于大多数REC模型，不需要任何网络修改。

3. 重新创建LIP

3.1. 问题定义

给定一个图像I和一个文本表达式T，引用表达式理解（REC）的目的是通过一个边界框b来定位目标实例。在现有的弱监督设置下[19,20,38]，该模型仅基于文本表达和图像学习检测，这是难以实现的。

在这种情况下，现有的弱监督解通常采用预先训练的两阶段检测网络，e。*g.*，更快的-RCNN[37]，提供了一组候选绑定-接线盒B[2]，类似于现有的两阶段REC方法[16,21,42]。然后，将REC表述为一个区域-文本匹配问题，定义为由

$$b^{*}=\underset{b\in B}{\arg\max}\,\Phi(T、I、b)、(1)$$

其中b^{*}是最佳匹配的框，而Φ（•）是一个跨模态排序网络，它返回候选区域（框）和表达式之间的相似性。然后，该模型进行基于语义重建[19,20,38]或跨模态对比损失[7,43]的弱监督训练。尽管有可行性，但这个解决方案需要复杂的后处理，e。*g.*，ROI池用于区域特征提取，这极大地限制了其推理速度。

为此，我们求助于像YOLOv3 [36]这样的高效的单级探测器来构建我们的RefCLIP。RefCLIP还利用了YOLOv3的检测能力。但在实践中，我们将REC任务简化为一个锚定-文本匹配问题，i。*e.*，哪个锚最有可能有目标框：

$$a^{*}=\underset{a\in A}{\arg\max}\,\phi(T,I,a),(2)$$

其中a^{*}是最佳锚点，A表示YOLOv3中的锚点的集合，ϕ（•）是一个简单的线性排序模块。为了解释，像YOLOv3这样的单级检测器的预测是基于输出特征图的网格特征，也被称为锚点。通过知道哪个锚点是正确的，我们可以大大缩小候选框的范围，最终得到最自信的盒子作为预测。

2一些方法使用MSCOCO的官方注释作为候选注释。

更重要的是，通过等式2、我们可以直接使用卷积主干来提取锚定特征，而不需要进行复杂的后处理。为了实现弱监督优化，我们进一步在图像内外进行基于锚定的对比学习。

3.2. 锚定选择

RefCLIP的框架如图所示。2.类似于流行的跨模态对比学习方案，i。*e.*，CLIP [35]，RefCLIP还将视觉和文本特征投射到一个联合语义空间上，并通过许多多模态对学习视觉-语言对齐。

在RefCLIP中，使用所有锚点作为候选锚点会阻碍对比学习的效率和质量。这是因为单级探测器[17,36]通常是多尺度的，所以它们有数千个候选锚点，其中大部分是背景的或低质量的。

因此，RefCLIP需要过滤掉大多数低值锚点，如图所示。2.首先，我们只保留最后一个卷积特征映射的锚点。为了解释，在最近的REC数据集[10,30,32]中，大多数对象都相对较大，可以被小分辨率特征图中的锚点检测到。其次，我们根据剩余的锚点的置信度得分e对其进行过滤。*g.*，选择前10%的锚点。

然后，RefCLIP计算这些候选锚点与联合语义空间中的表达式之间的相似性，然后返回最佳匹配的锚点作为正锚点进行对比优化。

.3.3.基于锚点的对比性学习

为了实现弱监督学习，我们引入了一种基于锚定的跨模态对比学习方案。具体来说，给定一个图像I和一个表达式T，我们首先使用检测网络和语言编码器来提取它们的特征，记为$F_v \in R^{h \times w \times d}$和$f_t \in R^d$各自地然后，用F中相应的特征来表示一个锚点v，表示为$f_a \in R^d$.

在锚选择后，我们将所选择的锚f_a和文本特征f_t在同一语义空间上，它们的相似性由

$$\mathrm{sim}(f_a, f_t) = (f_a W_a)^T (f_t W_t), \tag{3}$$

其中W_a和W_t是投影矩阵，而sim（·）可以看作是等式中的轻量级排名模块吗2.

在REC中，图像中的目标实例和表达式通常是一对一匹配的。理论上，只有一个锚是积极的例子，其余的都是消极的，特别是那些被过滤掉的。因此，我们定义了图像之间和图像内部的对比损失：

$$\mathcal{L}_c = -\log \frac{\exp\left(\mathrm{sim}(f_{a_0}^i, f_t^i)/\tau\right)}{\sum_{n=0}^{N}\sum_{j=0}^{M} \mathbb{I}_{\neg(i=j \wedge n\neq 0)} \exp\left(\mathrm{sim}(f_{a_n}^j, f_t^i)/\tau\right)}, \tag{4}$$

其中$f_{a_n}^j$锚点是从一批和$f_{a_0}^i$是图像i的积极之一。$\mathbb{I}_{\neg(i=j \vee n\neq 0)}$是指示器函数，当i = j和n0时等于0。≠N和M分别表示每幅图像和批线大小的负锚点的数量。τ为温度[9]。在N方面，我们根据它们的置信度分数来选择负锚点。

从等式4，我们可以看到RefCLIP在增加负样本方面的灵活性。原则上，更多的负样本可以更好地促进优化。然而，在现有的图像级对比学习方案中，负性例子的数量仅限于批处理大小[4]或依赖于外部堆栈[8]。在我们的基于锚定的方案中，负样本的数量可以是批量大小的多倍，大大提高了训练效率。

.4.3网络设置

如图所示。2、RefCLIP由一个预先训练过的单级检测器，i组成。e.，YOLOv3 [36]，一个语言编码器和一个多尺度融合模块[25,26]。语言编码器是一个双向的GRU [2]，然后是一个自我注意层[40]。在交叉模态匹配之前，我们采用多尺度融合模块[26]来融合三个尺度的语义信息。

在推理过程中，RefCLIP首先选择最佳匹配的锚点，在此基础上使用检测头来预测边界盒。由于一个锚点可能产生几个盒子[36]，我们使用置信度最高的一个作为预测。

4. 基于伪标签的弱监督训练方案

在本节中，我们介绍了一种新的基于伪标签的任意REC模型的训练方案，这也是在REC中的首次尝试。在该方案中，RefCLIP通过教师的伪标签进行普通REC模型的教学，可以帮助他们在不做任何修改的情况下推广到弱监督REC。

给定一个图像-文本对（I，T），我们首先使用RefCLIP来生成伪标签b。在此之后，我们构造了一个三联体（I，T，b）来监督共同的REC模型，其目标可以被定义为

$$\text{最小} L_s(I, T, b; \theta_s), \tag{5}$$

其中θ_s表示模型参数，和L_s是损失函数，可以是两阶段模型[42]的排名损失，也可以是单阶段模型[5,45]的回归损失。 *RefCLIP生成的伪标签仍然可能有噪声和质量低，导致一个称为确认偏差[1]的关键问题。*这一问题意味着训练信号可能被噪声样本所主导，累积的误差最终会限制性能

后

天花板利用最新的研究进展[23,31]，我们实施了两种设计来缓解这个问题。

具体来说，我们对输入图像进行数据增强，e。*g.*，随机调整了[13]的大小，以防止模型过早地过度拟合伪标记数据。此外，我们采用指数移动平均线（EMA）[39]的REC模型，定义为

$$\theta_s^t \leftarrow \alpha\,\theta_s^{t-1} + (1-\alpha)\,\theta_s^t, \tag{6}$$

式中，α为EMA系数，t为训练步骤。如等式中定义的6、EMA将逐步集成在不同训练状态下的REC模型，从而防止决策边界从向有噪声的样本移动。最后，在我们的训练方案中的梯度更新是：

$$\theta_s^t = \hat{\theta}_s - \gamma \sum_{k=1}^{t-1} (1-\alpha^{-k+(t-1)}) \frac{\partial \mathcal{L}_s(I,T,b;\theta_s)}{\partial \theta_s^k}, \tag{7}$$

其中θˆs表示初始模型的权重。

虽然该方案与完全监督训练相似，但在训练过程中没有使用任何地面真实边界框，这与弱监督REC [19,20]的定义相一致。

5. 实验

.1. 5数据集和度量

RefCOCO [32]从19,994张MSCOCO [15]图像中有142,210个引用表达式和50,000个对象。RefCOCO的表达式主要是关于绝对空间信息的。RefCOCO+ [32]包含了来自19,992张MSCOCO图像的49,856个边界框的141,564个参考表达式。RefCOCO+的数据分割与RefCOCO相同。然而，RefCOCO+的描述是关于相对的空间信息和外观，如颜色和纹理。RefCOCOg [30,32]有26,711张图像中的54,822个边界框的104,560个参考表达式。与RefCOCO和RefCOCO+相比，RefCOCOg的表达时间更长、更复杂。在这里，我们在实验中使用了RefCOCOg的谷歌分裂[30]。**参考游戏[10]有来自SAIAPR12数据集的19,997张图像，99,220个边界框和120,072个参考表达式。**我们根据伯克利分割将数据集划分为训练，val，测试。我们使用IoU@0.5作为度量标准。如果预测值与地面真实框之间的伊奥单位大于0.5，则预测是正确的。

5.2. 实施细节

我们将输入图像的大小调整为416×416。RefCOCO、RefCOCOCO+和RefCOCOCOg的输入文本的最大长度设置为15，参考游戏的最大长度设置为20。为了RefCLIP，我们使用YOLOv3 [36]作为检测器提取锚定特征，在MS-COCO [15]上进行预训练，去除上述三个数据集的val和测试集图像。为了与参考游戏中的[21,41]进行公平的比较，我们使用在视觉基因组[12]上预训练的YOLOv3作为我们的RefCLIP的检测器。在训练过程中，YOLOv3的参数是固定的。语言编码器的维度被设置为512。通过多尺度融合，将锚定特征投影到512个。在基于锚点的对比学习中，线性投影的维数为512，每张图像默认使用2个负锚点。所有模型均由Adam [11]优化器进行训练，学习速率不变为1e-4。训练时代和批处理大小分别设置为25和64。对于弱监督训练方案，我们将随机调整大小作为输入图像的数据增强。EMA系数设置为0.9997。RealGIN、SimREC和TransVG的其他配置与它们的默认设置相同。

.3. 5定量分析

消融RefCLIP。帐单1显示了RefCLIP中两种主要设计的烧蚀结果。、*锚定选择和负锚定增强（NAA）*。NAA表示在不改变批大小的情况下添加负样本。我们首先可以观察到，锚点过滤对于RefCLIP是至关重要的。在没有任何滤波规则的情况下，RefCLIP的性能实际上远远不令人满意，这证实了我们对锚定噪声的动机。在这种情况下，一个简单的比例选择可以在很大程度上提高性能，e。*g.*，RefCOCO的+为17%。当与基于置信度的过滤相结合时，这两个数据集的性能都可以进一步提高。最后一行的结果，i。*e.*，在图像内添加负锚也有利于REC性能，可以在非常有限的额外成本的情况下提高对比学习。

帐单2显示了不同设置的效果。我们首先注意到，52×52或26×26的比例会导致性能的急剧下降，尤其是前者。如上所述，现有REC数据集中的参考值相对较大，所以在这些尺度上，目标边界框几乎没有分布在预测上，这也解释了为什么52×52的精度为零。在这种情况下，最小的尺度，i。*e.*，13×13，是最好的选择。即便如此，YOLOv3的锚点仍然是多余的。如选项卡所示。2、通过基于置信度过滤高达80%或90%的锚点，性能仍然可以略有提高。这些结果很好地证实了我们关于对比学习的锚定冗余的假设。

在选项卡中。4，我们检验了负样本量对对比学习的影响。具体来说，我们调整了每幅图像的负锚的数量和批大小，i。e、N和M，定义于等式中4.我们首先观察到，较大的批处理大小有利于对比

表1。RefCLIP的消融研究。“比例”指的是比例的大小经文”Conf.”是置信度过滤器。“NAA”表示负的锚定增强。

锚杆选择		对比学习	雷夫科科	雷富科+
比例尺	会议	纳亚	val	val
–	–	–	33.71	29.11
/	–	–	50.75	36.65
/	/	–	53.30	40.07
/	/	/	**60.36**	**40.39**

表2。对RefCLIP的锚点选择设置的影响。

锚杆选择	设置	雷夫科科 val	雷富科+ val
比例尺挑选	所有	48.75	38.14
	52 × 52	0.00	0.00
	26 × 26	11.23	7.19
	13 × 13	**60.36**	**40.39**
信赖过滤	100%	20.84	39.74
	20%	59.31	**41.06**
	10%	**60.36**	40.39
	5%	48.46	39.69

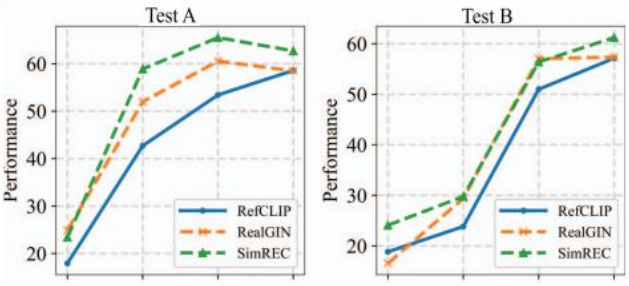


图3。RefCLIP的性能对普通REC模型的影响。*e.*，RealGIN和SimREC，关于RefCOCO测试分割。

学习，但随着批规模的增加，优点将变得边缘。因此，我们只测试了最大批处理大小为64。第二个块显示了图像中负锚点的影响。我们可以观察到，N = 2不会带来太多的额外成本，但其性能的提高是显著的，这表明我们在负锚定增强方面的优势。我们还注意到，使用更多的负锚会适得其反。*g.*，N = 3，这与现有的对比学习研究[8]不一致。一个潜在的原因是，RefCLIP只需要优化语言编码器和联合语义空间，这使得在现有的数据规模上很容易进行过拟合。

弱监督训练的消融。我们在Tab的方案中进一步研究了EMA和数据增强的影响。3.我们首先可以观察到，该训练方案对弱监督的REC是有效的。在所有三个分裂上，弱监督RealGIN和RefCLIP之间的性能差距并不明显。同时，在数据增强和EMA的帮助下，RealGIN的性能得到了全面的提高，表明了其对模型训练的有效性。

表3。对所提出的弱监督列车的消融研究ing方案。RealGIN是基础模型，并使用RefCLIP作为参考。

模型	方法 8月EMA		我的测试		
重新创建LIP	– –		**60.36** **57.13**	**58.58**	
RealGIN	–	–	57.36	57.34	56.33
	/	–	58.99	58.51	55.66
	/		59.43	58.49	**57.36**

表4。RefCLIP中阴性样本量的消融。N和M表示每幅图像的负锚点的数量和批数的大小。

约束性学习	设置	Neg。数量	RefCOCO val	RefCOCO+ val
M	16	15	48.98	40.08
	32	31	52.74	40.98
	64	63	53.30	40.07
N	1	63	53.30	40.07
	2	**126**	**60.36**	**40.39**
	3	189	44.41	38.66
	5	315	42.98	38.46

图3说明了RefCLIP的性能对测试的REC模型的影响。第一个观察结果是RefCLIP的质量极大地影响了这些常见的REC模型的弱监督性能。然而，我们也可以看到，RefCLIP的性能并不总是我们的训练方案的性能上限。当被测模型具有更好的多模态推理能力或更先进的REC设计时，其性能在不同设置下很容易超过RefCLIP。*g.*、SimREC和RealGIN。这些结果极大地验证了我们的方案对现有的REC模型的推广。

与最先进的技术进行比较。我们通过比较Tab中的一组弱监督的REC模型，来检查我们的弱监督训练方案和RefCLIP。5.在选项卡中。5，我们比较了提出的RefCLIP和常见的REC模型，包括单阶段REC模型[5,25,45]和两阶段REC模型[42]由我们的方案与更弱监督方法。之前的最佳性能是通过方法[18,20,38]的设置。即便如此，RefCLIP在大多数分割上都可以优于这些方法，最高可达21个。1%的RefCOCO val。

帐单5还显示了由我们的弱监督训练方案训练的现有REC模型的结果表示为RefCLIP_模型名称。我们可以看出，我们的训练方案可以帮助常用的REC模型在多次分割上轻松超越现有的SOTA性能。，71.27关于RefCOCO测试B。我们还观察到，MAttNet的性能提高比现有的更明显。g.，+14.14%的RefCOCO测试。根据这些结果，我们的假设是，两阶段的REC模型不需要学习边界盒回归，从而重新学习

表5。在四个REC基准数据集上与最先进的方法进行了比较。地面真相的建议意味着使用MSCOCO的官方注释作为候选人。为了进行公平的比较，这些方法的推理速度没有进行比较。重新创建LIP_ModelName表示在我们的弱监督训练方案中由RefCLIP训练的常见REC模型。

方法	我的测试	RefCOCO+ val 测试A测试b	RefCOCOg val-g	请参阅游戏 试验	推理速 度
地面真相建议书：					
VC [33]CVPR18 - 33.29 30.13 - 34.60 31.58 30.26 -					
ARN [19]ICCV19 38.05 36.43 36.47 34.53 36.40 36.12 39.62 -					
KPRN [20]MM19 36.34 35.28 37.72 37.16 36.06 39.29 38.37 33.87					
DTWREG [38]TPAMI21 39.21 41.14 37.72 39.18 40.01 38.08 43.24 -					
获得[18]TPAMI22 38.08 38.25 38.59 37.54 37.58 37.92 45.33 36.86					
重新创建LIP_MAtt网[42]我们的	**69.31 67.23 71.27**	**43.01 44.80 41.09**	**51.31**	-	-
检测到的建议：					
VC [33]CVPR18 - 32.68 27.22 - 34.68 28.10 29.65 14.50					
KAC Net[3]CVPR18---------15.83					
马[44]CVPR18——13.61					
ARN [19]ICCV19 32.17 35.25 30.28 32.78 34.35 32.13 33.09 26.19 5.7fps					
IGN [43]NeurIPS20 34.78 37.64 32.59 34.29 36.91 33.56 34.92 -					
DTWREG [38]TPAMI21 38.35 39.51 37.01 38.91 39.91 37.09 42.54 5.9fps					
RelR [22]CVPR21 - - - - - - - 37.68					
NCE+Distillation [41]CVPR21 - - - - - - 38.39					
RefCLIP（我们的）60.36 58.58 57.13 40.39 40.45 38.86 47.87 39.58 31.3fps				重新创建LIP_RealGIN[45]（我们的）59.43 58.49 57.36 37.08 38.70 35.82 46.10 37.56 51.7fps	重新创建LIP_SimREC[25]（我们的）62.57 62.70 61.22 39.13 40.81 36.59 45.68 42.33 54.8fps
重新创建LIP_TransVG [5]（ours）64.08 63.67 63.93 39.32 39.54 36.29 45。	**70 42.64 19.3fps**				

在很大程度上回避了弱监督REC的困难。更重要的是，RefCLIP或我们的单级基础模型的推理速度都比现有的弱监督模型要快得多。，与DTWREG [38]相比，RefCLIP将推理速度提高了一个数量级。这些结果很好地证实了RefCLIP和我们的培训方案的有效性。

.4.5定性分析

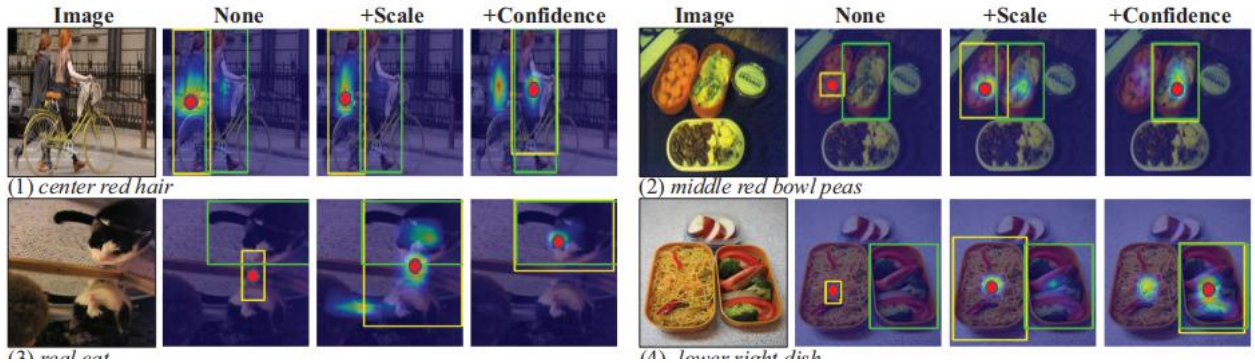
为了深入了解所提出的RefCLIP和训练方案，我们在图中进一步可视化了不同设置下的预测。4.从无花果。4-a，我们可以看到，如果没有任何过滤，RefCLIP的视觉-语言对齐能力是非常有限的。同时，该模型也很容易选择大小不合适的盒子，例如。，第二个和第四个例子。这种情况可以通过规模选择来很好地缓解。*e.*，″+scale″.有了置信滤波，我。*e.*，“+置信度”，RefCLIP的预测精度进一步提高，验证了我们对锚点冗余的关注。图4-b为不同负样本量下RefCLIP的预测。可以看出，适当增加负锚点可以极大地提高对比学习，使锚点-文本匹配更加准确。*g.*，第一个例子。最后，我们将RefCLIP与图中所训练的基础REC模型。4-c.可以看出，这些常见的REC模型的预测并不总是与他们的老师RefCLIP一致。当这些模型具有更强的推理能力时。它们甚至可以显示出比RefCLIP，e更好的跨模态对齐。*g.*，第7个和第8个例子。这些结果也很好地证实了我们的训练方案的推广和优越性。

6. 限制和未来的工作

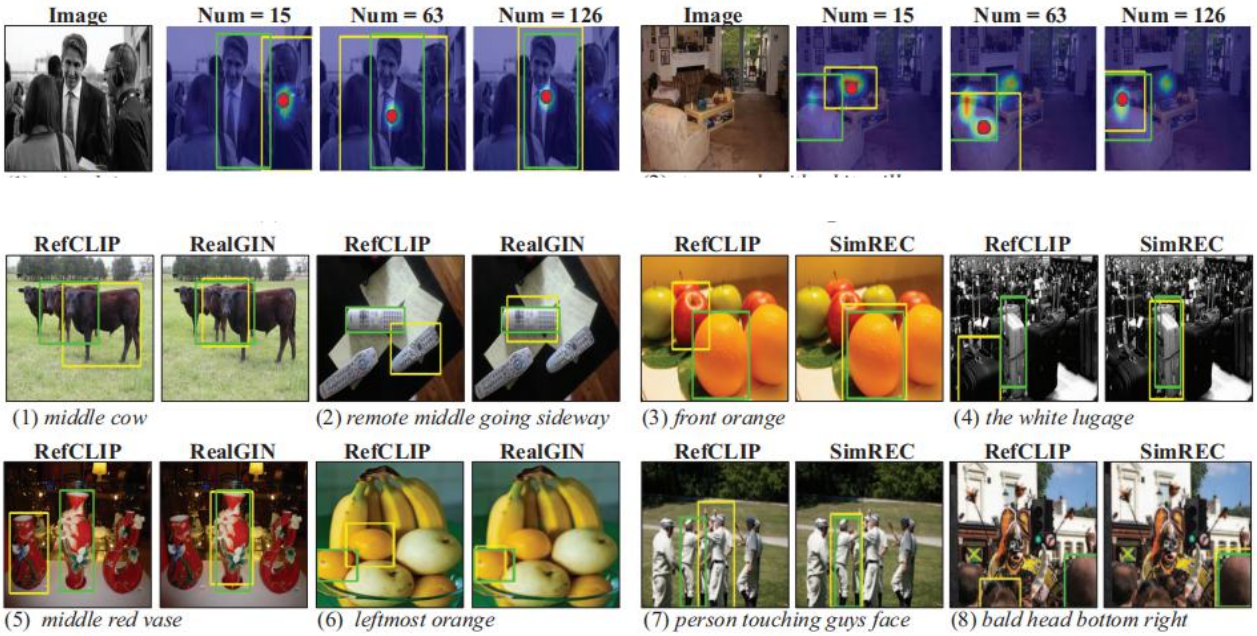
RefCLIP的检测规模是为REC任务设计的，这可能会限制其在小目标检测中的性能。此外，我们的弱训练方案可能会导致学生模型在更容易的样本上表现更好，从而导致在更具挑战性的数据集上的教学质量降低。未来的研究将集中于解决这些限制，并扩大我们的方法在其他多模式任务中的应用。

7. 结论

在本文中，我们主要关注有效的和一般的弱监督REC。具体来说，我们首先提出了一种新的弱监督模型，称为RefCLIP。为了避免复杂区域特征提取，RefCLIP将REC重新定义为锚文本匹配问题，实现了弱su-



具有不同锚点过滤规则的RefCLIP的(a)预测。



RefCLIP和弱监督常见REC模型的(c)预测。

图4。由我们的弱监督学习方案训练的RefCLIP和普通REC模型的可视化。黄色和绿色的盒子分别是预测的和地面真实的。子图(a)显示，比例选择和置信度过滤可以帮助RefCLIP更好地选择目标框。子图(b)中的示例反映了更大的负样本量对锚定-文本匹配的好处。在子图(c)中，我们可以看到由我们的方案弱训练的常见REC模型的预测并不总是与他们的老师RefCLIP一致，有时甚至更好。

通过基于锚的对比学习进行优化优化。在RefCLIP的基础上，我们进一步提出了第一个针对常见REC模型的建模无关的弱监督训练方案，其中RefCLIP作为伪标签学习的教师。该方案适用于大多数现有的REC模型，不进行任何网络修改。在四个基准测试上的实验结果不仅表明RefCLIP比现有的弱监督REC模型的性能提高，而且证实了我们的训练方案的有效性和泛化能力。

致谢这项工作得到了Na-的支持。
中国国家重点研发计划（NoZD0118201），.2022
国家杰出青年科学基金项目（第62025603号），国家自然科学基金项目。U21B2037, No.U22B2051, No.62176222, No.62176223, No.62176226, No.62072386, No.62072387, No.62072389, No.62002305和No.62272401)；福建省自然科学基金资助项目（编号2021J01002，No. 2022J06001）；中央高校基本科研业务费基金资助项目(编号No.20720220068).

参考文献

[1]埃里克·阿拉佐，迭戈·奥尔特戈，保罗·艾伯特，诺埃尔·奥康纳和凯文·麦吉尼斯。深度半监督学习中的伪标记和确认偏差。*2020年国际神经网络联席会议（IJCNN），第1-8页*。2020年IEEE。2，4

[2]，赵庆云，和本虎。神经机器翻译通过联合学习来对齐和翻译。*arXiv预印本，arXiv：1409.0473,2014年*。4

陈[3]，高济阳、南华山。知识有助于弱监督短语接地的一致性。*发表在IEEE计算机视觉和模式识别会议论文集上，第4042-4050页，2018页*。7

[4] Ting Chen，西蒙·科恩布利斯，穆罕默德·诺鲁齐和杰弗里·辛顿。一个简单的视觉表征对比学习的简单框架。*在关于机器学习的国际会议上，第1597-1607页*。PMLR，2020.4

[5]邓家俊、杨正元、陈天郎、周文刚、李后强。带变压器的端到端视觉接地。*请参见《IEEE/CVF计算机视觉国际会议论文集》，第1769-1779页，2021页*。1，2，4，6，7

[6] Pelin Dogan，列昂尼德·西格尔和马库斯·格罗斯。神经顺序短语接地（序列接地）。*发表在IEEE/CVF计算机视觉和模式识别会议论文集上，第4175-4184页，2019页*。2

[7] Tanmay Gupta，阿拉什·瓦达特，盖尔、切希克、杨晓东、考茨和德里克·霍姆。弱监督短语接地的对比学习。*在欧洲计算机视觉会议上，第752-768页*。施普林格，2020年。1，2，3

何[8]开明、范浩奇、吴宇新、谢素明、吉希克。无监督视觉表征学习的动量对比。*发表在IEEE/CVF计算机视觉和模式识别会议论文集上，第9729-9738页，2020页*。4，6

[9]，杰弗里·辛顿，奥里奥尔葡萄酒公司，杰夫·迪恩，等人。在神经网络中提炼知识。*arXiv预印本arXiv：1503.02531,2(7)，2015年*。4

[10] Sahar卡兹姆扎德，维森特奥多内斯，马克马滕，和塔玛拉伯格。参考游戏：指自然场景照片中的物体。*在2014年自然语言处理中的经验方法（EMNLP）会议论文集中，第787-798页，2014年*。2，4，5

[11]饮食协会，金马和吉米·巴。一种随机优化的方法。*arXiv预印本arXiv：1412。6980,* 2014.5

[12]兰杰·克里希纳，朱玉克，奥利弗·格罗斯，贾斯汀·约翰逊，哈塔健二，约书亚·克拉维茨，斯蒂芬妮·陈，扬妮斯·卡兰蒂迪斯，李佳李，大卫·阿沙玛，等。视觉基因组：使用众包密集的图像注释来连接语言和视觉。*《计算机视觉国际杂志》，123(1)：32-73,2017*。5

[13]亚历克斯·克里热夫斯基，伊利亚·苏茨克弗和杰弗里·辛顿。基于深度卷积神经网络的图像集分类。*神经信息处理系统的研究进展，2012年25日*。2，5

廖[14]、刘思、李冠斌、王飞、陈彦杰、陈谦、李波。一种用于参考表达式理解的实时跨模态相关滤波方法。在*IEEE/CVF计算机视觉和模式识别会议论文集，第10880-10889页，2020页*。2

[15]宗-林毅、迈克尔·梅尔、塞尔日·贝隆吉、詹姆斯·海斯、彼得罗纳、拉曼南、彼得多尔和劳伦斯·齐特尼克。微软coco：上下文中的常见对象。*在欧洲计算机视觉会议上，第740-755页*。施普林格，2014年。5

[16]大庆Liu，张汉王，冯武，查郑军。学习通过组装神经模块树网络来进行视觉接地。*发表在IEEE/CVF计算机视觉国际会议论文集上，第4673-4682页，2019页*。1，2，3

[17] Wei Liu、安圭洛夫、埃尔汉、基斯蒂、里德、傅成阳、亚历山大·伯格。单镜头多盒探测器。*在欧洲计算机视觉会议上，第21-37页*。施普林格，2016年。2，4

刘[18]学景、李梁、王淑慧、查郑军、李子超、齐天、黄清明。实体增强自适应重构网络。*《IEEE《模式分析与机器智能学报》，2022年*。2，6，7

刘[19]学景、李梁、王淑慧、查郑军、德浩、黄清明。弱监督参考表达式接地的自适应重构网络。*发表在IEEE/CVF计算机视觉国际会议论文集，第2611-2620页，2019页*。2，3，5，7

刘[20]学景、李梁、王淑慧、查郑军、李苏、黄清明。弱监督参考表达接地的知识引导成对重构网络。*发表在第27届ACM多媒体国际会议论文集上，第539-547页，2019年*。2，3，5，6，7

刘[21]、王子豪、邵景、王小刚、李宏生。通过跨模态注意引导擦去来改进参考表达式基础。*《IEEE/CVF计算机视觉和模式识别会议论文集》，1950-1959页，2019页*。2，3，5

刘永飞[22]、波湾、林马、何旭明。弱监督视觉接地的关系感知实例细化。*发表在IEEE/CVF计算机视觉和模式识别会议论文集上，第5612-5621页，2021页*。2，7

刘延成[23]、马志耀、何子健、郭嘉文、陈人、张培昭、武比臣、基拉、金刚彼得。半监督目标检测的无偏教师。*arXiv预印本arXiv：2102.09480,2021年*。5

[24]罗创、周依依、吉荣荣、孙小帅、苏金、林嘉文、齐天。采用级联分组注意网络进行参考表达分割。*发表在第28届ACM国际多媒体会议论文集上，第1274-1282页，2020年*。2

[25]罗创、周依依、孙嘉木、黄树斌、孙小帅、叶启祥、吴永健、季荣荣。什么

单阶段参考表达理解超越多模态融合：实证研究。*arXiv预印本arXiv：2204.07913,2022年*。1，2，4，6，7

[26]罗创、周依依、孙小帅、曹刘娟、吴成林、邓成诚、吉荣荣。多任务协作网络的联合引用表达理解和分割。*发表在IEEE/CVF计算机视觉和模式识别会议论文集上，第10034-10043页，2020年*。1，2，4

[27]罗创、周依依、孙小帅、丁兴浩、吴永健、黄飞跃、高岳、纪荣荣。通过动态卷积来实现语言引导的视觉识别。*arXiv预印本arXiv：2110.08797,2021年*。1

[28]罗创、周依依、孙小帅、王燕、曹刘娟、吴永健、黄飞跃、季荣荣。通过对视觉和语言任务的组级转换来实现轻量级转换器。IEEE图像处理交易报，31：3386-3398,2022。1

[29]罗文杰、李玉佳、乌尔塔森、泽梅尔。理解深度卷积神经网络中的有效接受域。*神经信息处理系统的进展，2016年29日*。2

[30]·毛俊华、黄乔纳森、亚历山大·托舍夫、奥纳·坎布鲁、艾伦·尤耶和凯文·墨菲。生成和理解明确的对象描述。在*IEEE计算机视觉和模式识别会议论文集，第11-20页*。2，4，5

[31]彭美、林江营、周依依、沈云亨、罗将军、孙小帅、曹刘娟、傅荣荣、徐强、纪荣荣。负责半监督目标检测的主动教师。*发表在IEEE/CVF计算机视觉和模式识别会议论文集上，第14482-14491页，2022页*。5

[32] Varun K纳加拉贾，弗拉德I莫里乌和拉里戴维斯。建模对象之间建模上下文，以引用表达式理解。*在欧洲计算机视觉会议上，第792-807页*。施普林格，2016年。2，4，5

[33]介绍，包括牛、张汉王、卢志武和张施福。变分上下文：利用视觉和文本上下文来建立引用表达式。IEEE平台上的模式分析和机器智能交易，43(1)：347-359，2019.7

[34]亚伦，李和葡萄酒。使用对比预测编码的表示学习。*arXiv预印本，arXiv：1807.03748,2018年*。3

[35]亚历克·雷德福，金正男，克里斯·哈勒西，阿迪提亚·拉梅什，加布里埃尔·高，桑德希尼·阿加瓦尔，吉里什·萨斯特里，阿曼达·阿斯凯尔，帕梅拉·米什金，杰克·克拉克等。从自然语言监督中学习可转移的视觉模型。*在机器学习国际会议上，*第8748-8763页。PMLR，2021.4

[36]约瑟夫·雷德蒙和阿里·法尔哈迪。Yolov3：一个渐进式的改进。*arXiv预印本，arXiv：1804.02767,2018年*。2，3，4，5

[37]少卿，任、何开明、罗氏、孙吉安。更快的r-cnn：利用区域建议网络实现实时目标检测。*神经信息处理系统的研究进展，2015年28日*。1，2，3

孙[38]明杰，小继民、林英吉、刘思、刘志。弱引用表达接地的判别三联征匹配与重建。*IEEE关于模式分析和机器智能的交易，43（11）：4189-4195,2021*。1，2，3，6，7

[39] Antti·塔瓦宁和哈里·瓦尔波拉。平均教师是更好的榜样：权重平均一致性目标提高了半监督深度学习的结果。*神经信息处理系统的进展，2017年30日*。2，5

[40] Ashish瓦斯瓦尼，诺姆沙泽尔，尼基帕尔马，雅各布乌斯科-

狮子王，琼斯，艾丹和戈麦斯，Łukasz Kaiser，和伊利亚·波罗苏欣。你所需要的就是注意力。*神经信息处理系统的进展，2017年30日*。2，4

王利伟、黄静、尹力、许坤、阳正元、董宇。通过对比知识精馏改进弱监督视觉基础。*发表在IEEE/CVF计算机视觉和模式识别会议论文集上，第14090-14100页，2021页*。2，5，7

[42]、林哲、沈晓辉、杨姬、新路、莫希特·班萨尔和塔玛拉·伯格。用于参考表达理解的模块化注意网络。*发表在IEEE计算机视觉和模式识别会议论文集上，第1307-1315页，2018年*。1，2，3，4，6，7

[43]朱张，周赵，林志杰，何秀强，等。弱监督视觉语言基础下的反事实对比学习。*神经信息处理系统的进展，33：18123-18134,2020*。1，2，3，7

[44]方赵，李建书，吉安赵，和嘉实峰。基于多尺度锚定变压器网络的弱监督短语定位。*发表在IEEE计算机视觉和模式识别会议论文集上，第5696-5705页，2018年*。3，7

[45]周依依、吉荣荣、罗将军、孙小帅、苏金、丁兴浩、林嘉文、齐天。一种用于一期推理表达理解的实时全局推理网络。*IEEE神经网络和学习系统学报，2021年*。1，2，4，6，7

[46]周依依、吉荣荣、孙小帅、苏金、德予孟、高岳、沈春华。很多都是瘟疫：视觉问答的细粒度学习。*IEEE关于模式分析和机器智能的交易，44(2)：697-709,2019年*。1

周[47]依依、天河人、朱朝阳、孙小帅、刘建庄、丁兴浩、徐明亮、吉荣荣。Trar：通过变压器中的注意力跨度进行视觉问题回答。*请参见《IEEE/CVF计算机视觉国际会议论文集》，第2074-2084页，2021页*。1

[48]朝阳Zhu、周依依、沈永亨、罗将军、潘兴家、林明宝、陈超、曹刘娟、孙小帅、吉荣荣。Seqtr：一个简单而通用的视觉接地网络。*arXiv预印本arXiv：2203.16265,2022年*。1，2