

指称表达理解与分割——MCN 的改进

摘要

Multi-task Collaborative Network (MCN) 模型于 2020 年在 CVPR 上提出，首次将指称表达理解和指称表达分割两个任务结合起来，在一个统一的模型中进行学习，使两个任务之间起到互相补充的效果。本文结合相关领域前沿论文与知识，在图像特征提取、多模态特征融合、图像数据增强、模型权重优化和检测框计算方式五个方面对 MCN 提出改进。实验证明，改进后的模型在性能表现上相比原 MCN 有显著提升。

关键词：指称表达理解；指称表达分割

1 引言

指称表达理解是一项根据文本指称表达检测图像中指定物体的任务，而指称表达分割和指称表达理解类似，是一项根据文本指称表达分割图像中指定目标的任务。同计算机视觉领域的目标检测任务和语义分割任务一样，它们的目的都是检测或分割出图像中的物体，但指称表达理解和指称表达分割更加注重于文本标签的自然表达属性，而不局限于固定的类别标签。正因为有这样更加符合人机交互的特性，近来指称表达理解和指称表达分割在计算机视觉领域和自然语言处理领域受到广泛关注。

由于文本标签的自然表达属性，指称表达理解和指称表达分割都要求模型能提取文本特征，能对图像和文本两个模态的特征进行准确配准。MCN 模型 [15] 首次提出将指称表达理解和指称表达分割两个任务结合起来共同学习，利用专注于细粒度分割的指称表达分割帮助指称表达理解更好地进行文本图像配准，利用专注于粗粒度检测的指称表达理解帮助指称表达分割更好地定位图像中的目标物体，从而使两个任务“相互补充”，达到更好的效果。

本文结合课程所学知识和目标检测、指称表达理解等相关领域论文，对 MCN 模型进行实验分析，并在图像特征提取、多模态特征融合、图像数据增强、模型权重优化和检测框计算方式五个方面对 MCN 模型作出改进，分别为将 MCN 模型中的图像编码器 DarkNet-53 替换成 YOLOv8 中的特征提取骨干网络，改进 MCN 模型中多尺度图像特征和文本特征的融合方式，对输入模型的训练图像进行随机缩放、利用 EMA [22] 优化模型权重以及将检测框计算方式改为 Anchor-free。在 RefCOCO [28] 的 val、testA 和 testB 上的实验结果表明，改进后的模型相较本地训练的基准模型有显著提升，且超过论文给出的模型结果，具体为在 REC 任务上分别超出 5.14%、5.50% 和 5.61%，在 RES 任务上分别超出 8.54%、8.75% 和 9.13%。

2 相关工作

2.1 指称表达理解

目前，指称表达理解（REC）的研究工作主要集中在监督学习方法上 [2, 7, 8, 10, 13–15, 27, 29, 30]，其主要可以分为两类，分别为二阶段方法和一阶段方法 [17]。二阶段方法 [8, 10, 27] 首先利用目标检测网络生成一系列的候选区域，然后根据指称表达对生成的候选框进行匹配分数排序，最终选出最符合的候选框，即为指称表达所指的物体，例如 MAttNet [27] 在第一阶段应用 Faster-RCNN [19] 生成候选目标框，然后在第二阶段使用三个模块计算外观、位置和关系三种不同类型的文本嵌入与生成候选目标框的匹配分数，用于确定最终的目标物体。近来，由于能同时兼顾推理速度和模型性能，一阶段方法 [2, 7, 13–15, 29, 30] 越来越受到关注。Luo 等人 [14] 在一阶段方法上进行大量的消融实验，提出了 SimREC 模型，该模型以更少的参数量和训练开销达到甚至超越了许多大规模预训练模型和二阶段方法的性能。与一般的一阶段方法不同，Su 等人 [20] 提出的 VG-LAW 不依赖于目标检测框架来提取图像特征，而是利用 transformer 架构将文本特征直接融入图像特征提取过程中，以提取与文本表达相关的图像特征。VG-LAW 模型的性能表现超过 SimREC，但随之也带来模型参数量的增加和训练开销的增大。由于数据标注工作需要耗费大量人力，近来弱监督方法 [4, 11, 12, 21, 24] 也越来越受到关注。

2.2 指称表达分割

早期指称表达分割相关研究工作 [6, 15, 16, 28] 主要应用卷积神经网络和循环神经网络分别提取图像和文本特征，然后对提取的特征进行混合来预测分割掩模。近来，越来越多的研究工作 [3, 5, 26] 开始通过应用图像和文本领域的 transformer 架构来提升指称表达分割模型的性能。此外，随着大规模预训练模型的发展，Wang 等人 [25] 提出借助 CLIP 模型强大的文本图像配准能力来提升指称表达分割模型的跨模态配准能力。

3 方法

3.1 原文方法概述

如图 1 所示，MCN 主要包含图像编码器、文本编码器、多模态特征混合层、REC 和 RES 分支中的注意力模块和用于计算检测框或分割掩模的头部五个模块，图中红色框为原论文作者提出的两个创新方法，通过这两个方法和下采样拼接融合操作，作者将 REC 分支和 RES 分支联系起来。模型输入为文本和图像，输出为检测框和分割掩模。

其中，图像编码器为 Darknet-53 深度卷积神经网络，主要包含卷积层、激活层、批归一化层和残差连接，用于提取多尺度的图像特征，该网络源于 YOLOv3 [18] 中的特征提取骨干；文本编码器为 Glove 嵌入层加上 GRU 层和自注意力层，用于将文本单词转为词嵌入向量，并从向量中提取文本整体的语义特征；多模态特征混合层将顶层小尺度的图像特征与文本特征映射到同一空间维度后进行点乘、变换和上采样，用于混合提取的多尺度图像特征和文本语义特征；多尺度特征融合为 PANet [9] 架构，用于对不同尺度的特征进行双向融合，使得每一尺度的特征都能得到充分利用，同时也起到增强 REC 和 RES 两个任务协同学习的效果。

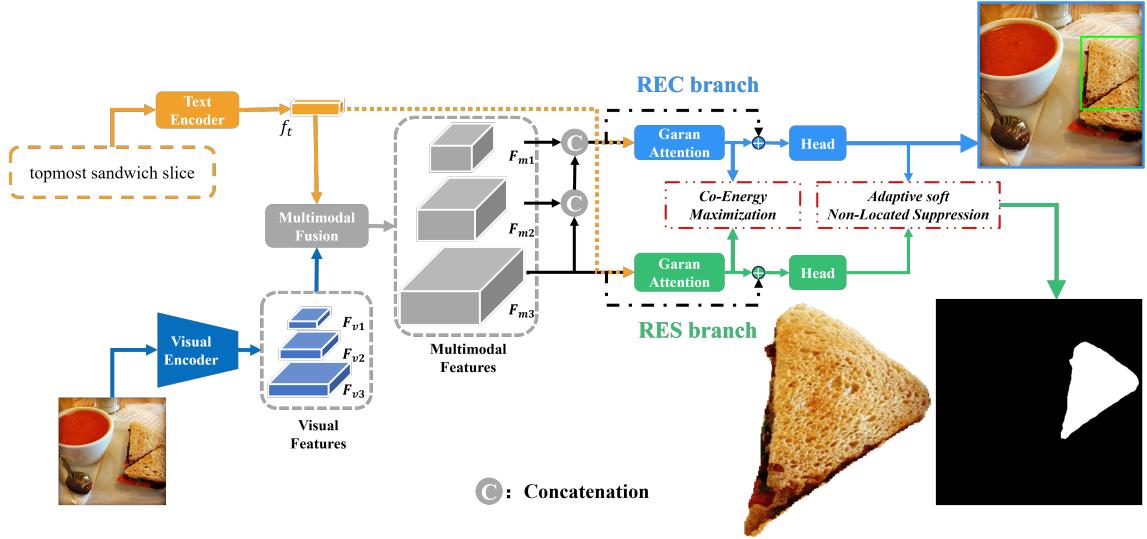


图 1. Multi-Task Collaborative Network [15] 框架示意图

3.2 改进图像编码器模块

图像编码器改进方式有两种，第一种是将原图像编码器替换为 YOLOv4 中的特征提取骨干，第二种是替换为 YOLOv8 的特征提取骨干，在本文实验中，YOLOv8 的特征提取骨干对图像的特征提取效果更好。

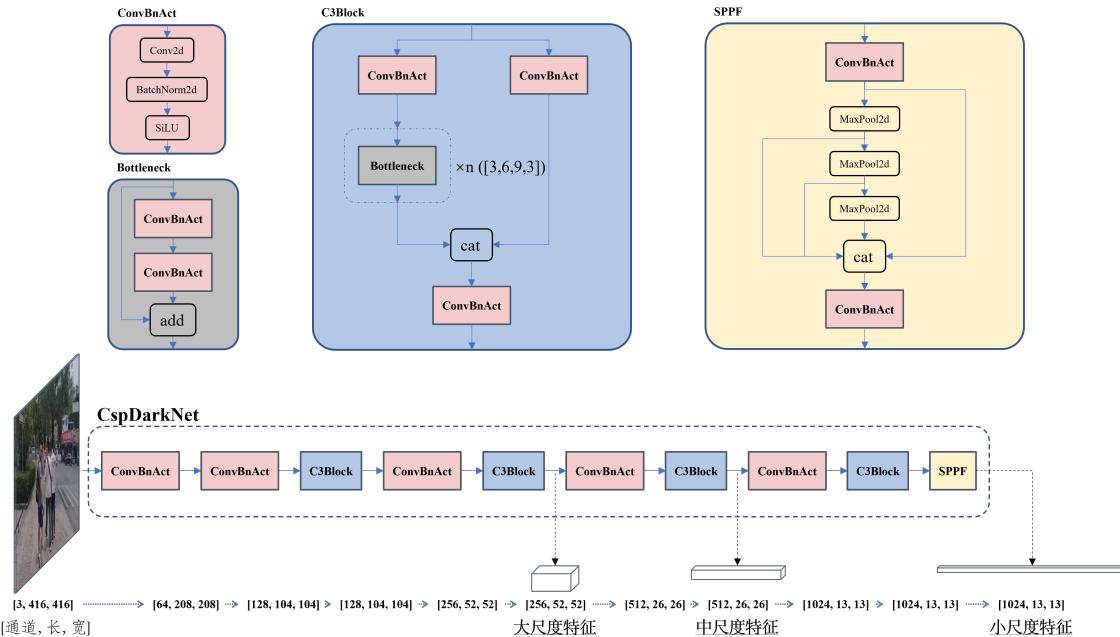
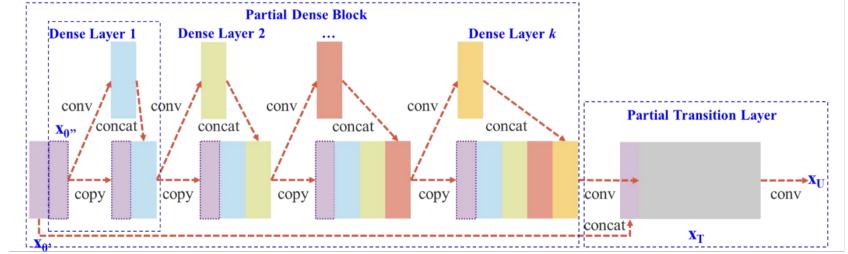


图 2. CSPDarkNet [15] 特征提取示意图

第一种改进方式的图像编码器模块如图 2 所示。本文将原图像编码器 Darknet-53，如图 3a 所示，替换为 CSPDarknet-53。CSPDarknet-53 源于 YOLOv4 [1] 检测网络，在 YOLOv4 论文中已经被证明更适合于检测任务。如图 3 所示，相比 Darknet-53，CSPDarknet-53 引入了 CSPNet (Cross Stage Partial Network) [23] 中的 Cross Stage Partial Connection 方式，如图 3b 所示，对应图 2 中的 C3Block，通过分割和合并特征的策略使梯度能够更好地在网络中传播，这种策略已被证明能够在降低计算量和内存消耗的同时，提升卷积神经网络的学习能力。

Type	Filters	Size	Output
Convolutional	32	3×3	256×256
Convolutional	64	$3 \times 3 / 2$	128×128
Residual			128×128
Convolutional	32	1×1	
Convolutional	64	3×3	
Residual			64×64
Convolutional	128	$3 \times 3 / 2$	64×64
Convolutional	64	1×1	
Convolutional	128	3×3	
Residual			64×64
Convolutional	256	$3 \times 3 / 2$	32×32
Convolutional	128	1×1	
Convolutional	256	3×3	
Residual			32×32
Convolutional	512	$3 \times 3 / 2$	16×16
Convolutional	256	1×1	
Convolutional	512	3×3	
Residual			16×16
Convolutional	1024	$3 \times 3 / 2$	8×8
Convolutional	512	1×1	
Convolutional	1024	3×3	
Residual			8×8
Avgpool			
Connected			Global
Softmax			1000

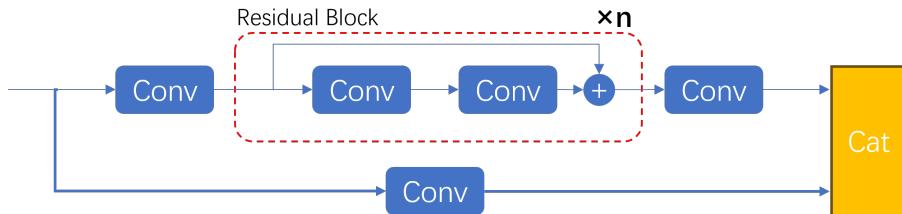
(a) DarkNet-53 [18]



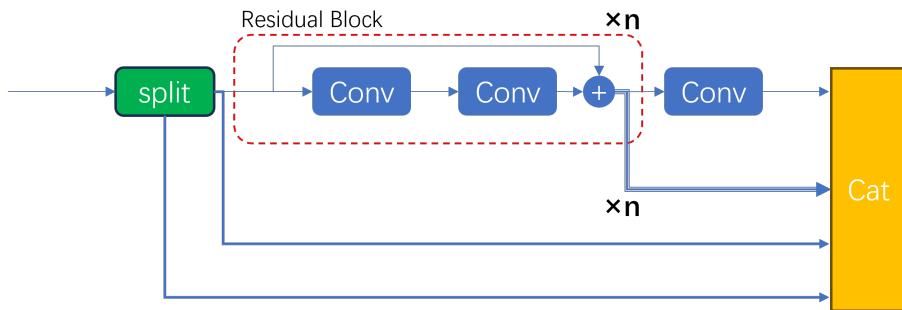
(b) Cross Stage Partial DenseNet [23]

图 3. DarkNet-53 特征提取网络示意图和 Cross Stage Partial DenseNet 架构示意图

第二种改进方式的图像编码器模块与第一种改进方式的主要区别在于 Cross Stage Partial Connection。如图 4 所示，图 4a 在图 4b 的基础上进一步调整分割方式以及增加了每个残差模块特征的输出，该方式源于今年推出的 YOLOv8，在目标检测任务中展现出了更优越的特征提取能力。



(a) YOLOv4 中的 Cross Stage Partial Connection



(b) YOLOv8 中的 Cross Stage Partial Connection

图 4. YOLOv4 和 YOLOv8 特征提取骨干网络中 Cross Stage Partial Connection 的差异

3.3 改进图像文本特征混合方式

原图像文本特征混合方式如图 5a 所示。原论文具体做法是首先对顶层小尺度图像特征和文本语义特征进行变换和点乘混合特征，后通过对小尺度的混合特征上采样叠加的方式来实

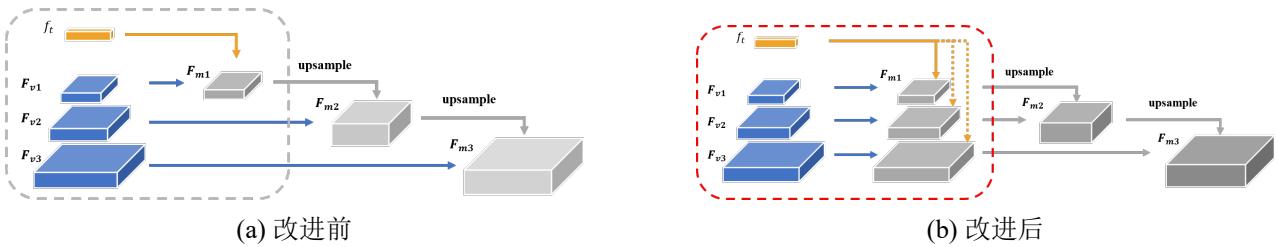


图 5. 改进前和改进后的特征融合方式示意图

现其他尺度图像特征与文本特征的混合，具体公式如下：

$$f_{m_1}^l = \sigma(f_{v1}^l \mathbf{W}_{v1}) \odot \sigma(f_t \mathbf{W}_t) \quad (1)$$

$$\mathbf{F}_{m_2} = [\sigma(Upsample(\mathbf{F}_{m_1}) \mathbf{W}_{m_1}), \sigma(\mathbf{F}_{v2} \mathbf{W}_{v2})] \quad (2)$$

$$\mathbf{F}_{m_3} = [\sigma(Upsample(\mathbf{F}_{m_2}) \mathbf{W}_{m_2}), \sigma(\mathbf{F}_{v3} \mathbf{W}_{v3})] \quad (3)$$

其中 \mathbf{W}_v 、 \mathbf{W}_t 和 \mathbf{W}_m 为可学习的投影权重矩阵， σ 为 *LeakyReLU* 激活函数， f_{v1}^l 为小尺度图像特征 \mathbf{F}_{v1} 中的向量， $f_{m_1}^l$ 为小尺度多模态特征 \mathbf{F}_{m_1} 中的向量，*Upsample* 为二维双线性插值上采样操作，将特征图像边长扩大为原来的两倍， $[\cdot]$ 为拼接操作， \mathbf{F}_{v2} 和 \mathbf{F}_{v3} 分别为中尺度和大尺度的图像特征， \mathbf{F}_{m_2} 和 \mathbf{F}_{m_3} 分别为中尺度和大尺度的多模态混合特征。

通过上采样叠加混合特征的方式，会使得文本信息在上采样的过程越来越粗粒度化，为了能够更加细粒度地混合图像文本特征，同时增加混入的文本信息量，本文对图像文本特征混合方式进行改进。

改进后的图像文本特征混合方式如图 5b 所示，具体做法为提前对所有尺度的图像特征和文本语义特征利用变换和点乘进行混合，具体特征混合公式如下：

$$f_{m_1}^l = \sigma(f_{v1}^l \mathbf{W}_{v1}) \odot \sigma(f_t \mathbf{W}_t) \quad (4)$$

$$f_{m_2}^l = \sigma(f_{v2}^l \mathbf{W}_{v2}) \odot \sigma(f_t \mathbf{W}_t) \quad (5)$$

$$f_{m_3}^l = \sigma(f_{v3}^l \mathbf{W}_{v3}) \odot \sigma(f_t \mathbf{W}_t) \quad (6)$$

其中 W_v 和 W_t 为可学习的投影权重矩阵， σ 为 *LeakyReLU* 激活函数， f_{v1}^l 、 f_{v2}^l 和 f_{v3}^l 分别为小尺度图像特征 \mathbf{F}_{v1} 中的向量、中尺度图像特征 \mathbf{F}_{v2} 中的向量和大尺度图像特征 \mathbf{F}_{v3} 中的向量， $f_{m_1}^l$ 、 $f_{m_2}^l$ 和 $f_{m_3}^l$ 分别为小尺度多模态特征 \mathbf{F}_{m_1} 、中尺度多模态特征 \mathbf{F}_{m_2} 和大尺度多模态特征 \mathbf{F}_{m_3} 中的向量。混合得到多尺度多模态特征后，按照原论文多尺度特征融合方式对不同尺度的特征按照 PANet [9] 方式进行双向融合。

3.4 改进指称表达理解头部

原论文检测框计算方式和损失计算方式如 6 左侧所示，检测框基于 anchor 进行计算，anchor 尺寸和数量需要人为预先设定，损失计算方式为对预测框和真实框的中心点坐标计算二值交叉熵损失，对长宽计算均方差损失。

本文参考 SimREC [14] 的检测头进行改进，改进后的检测框计算方式和损失计算方式如图 6 右侧所示，检测框直接从模型输出特征中计算得出，不依赖预先设定的 anchor，且损失函数改为计算预测框和真实框面积的交并比 IoU。

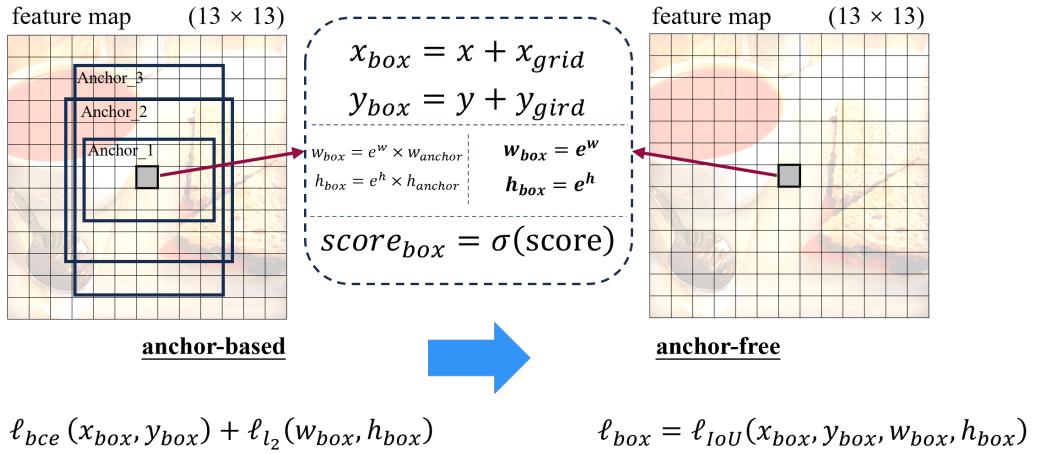


图 6. 改进检测框计算方式

3.5 数据增强

原文中 MCN 模型输入地图像尺寸固定为 416×416 ，为了使模型能够适应多尺度图像输入，同时提高模型对不同尺度目标的检测和分割能力，本文将固定输入尺寸改进为随机缩放输入尺寸，将输出图像尺寸扩充为 224×224 到 640×640 ，其中边长均为 32 的整数倍。

3.6 模型权重优化

本文在原模型的基础上，引入 Exponential Moving Average (EMA) [22] 用于在测试阶段调整模型权重。其具体公式如下：

$$W'_i = \begin{cases} W_0, & i = 0 \\ \alpha W'_{i-1} + (1 - \alpha) W_i, & i = 1, 2, 3, \dots \end{cases}$$

$$\alpha = \min(\beta, \frac{i}{i+9}), \quad i = 1, 2, 3, \dots \quad (7)$$

其中， W_0 为随机初始化的模型权重， W 为训练阶段模型的权重，通过反向传播算法进行更新， W' 为测试阶段模型的权重， i 为模型权重更新次数， β 为超参数，设置为 0.9997。每次对测试阶段模型权重的调整都基于之前测试阶段调整后的模型权重和当前训练阶段的模型权重。

使用 EMA 在测试阶段对模型当前以及历史的权重进行加权平均，使模型权重融合更多的历史状态，可以削弱后期训练阶段反向传播对模型权重的影响，从而提高模型整体的稳定性。

4 复现细节

4.1 与已有开源代码对比

本文所选论文模型代码¹已经开源，在复现与改进指称表达理解分支头部时参考使用了 SimREC 代码库²，在改进图像编码器时参考使用了 YOLOv8 代码库³。

本文除了在指称表达理解头部和图像编码器部分进行改进，还通过随机缩放训练时输入模型的图像尺寸来增强数据，从而提升模型对不同尺度物体的检测和分割效果；通过调整多模态特征混合方式，来增强文本语义特征与多尺度图像特征的混合效果；通过引入 EMA 对模型当前权重和历史权重进行加权平均，从而提升模型的稳定性和训练时的收敛速度。

本文模型均在 PyTorch 框架上进行训练测试。本文所使用的图像编码器均在 MS-COCO 数据集上经过目标检测任务预训练。模型输入图像尺寸为 416×416 ，输入文本限制最大单词数为 15。模型训练的批量数设置为 32，训练轮数为 39，优化器学习率设置为 0.0001，学习率调节器中设置学习率前 3 轮从 $1e-7$ 线性增长到 $1e-4$ ，对模型进行预热，学习率衰减因子为 0.2，分别在 30 轮、35 轮和 37 轮进行学习率衰减，其他参数与原论文保持一致。

4.2 数据集与评价指标

本文使用到的数据集为 RefCOCO [28]，该数据集包含 142210 句指称表达，50000 个指称表达对应的目标框和 19994 张 MS-COCO 中的图像。RefCOCO 中的指称表达主要包含空间绝对位置信息、颜色以及类别名。实验中 RefCOCO 数据集按照 unc 的划分方式，分别划分为 train、val、testA 和 testB 四个部分，对应模型的训练、验证和测试阶段。

本文使用的评价指标与原论文保持一致。在 REC 任务上，本文使用 IoU@0.5 作为评价指标，即预测框和真实框面积的交并比大于或等于 50% 的比例。在 RES 任务上，本文使用 mIoU 作为评价指标，即所有验证或测试样例中预测掩模和真实掩模面积的平均交并比。

4.3 创新点

- 应用 YOLOv8 特征提取骨干网络对 MCN 图像编码器进行改进，提升模型性能。
- 改进 MCN 多模态特征混合层，使多模态特征混合入更多更细粒度的文本特征，提升模型性能。
- 改进指称表达理解分支头部，将 anchor-based 改为 anchor-free，进一步提升模型指称表达理解分支的性能。
- 引入随机缩放方法对训练图像进行增强，使模型能够检测和分割更多尺度的物体，提升模型性能。
- 引入 EMA 优化模型权重，进一步提升模型性能。

¹<https://github.com/luogen1996/MCN>

²<https://github.com/luogen1996/SimREC>

³<https://github.com/ultralytics/ultralytics>

5 实验结果分析

由于原论文使用的代码框架为 Tensorflow，本文使用的代码框架为 PyTorch，为了更准确地进行比较，本文按照论文作者给定的参数进行 MCN 模型训练，将训练好的模型作为基准模型。本文首先将改进后的模型与原模型进行比较，然后对每个改进的模块进行消融实验，最后对模型进行定性分析。

5.1 定量分析

5.1.1 改进前后对比

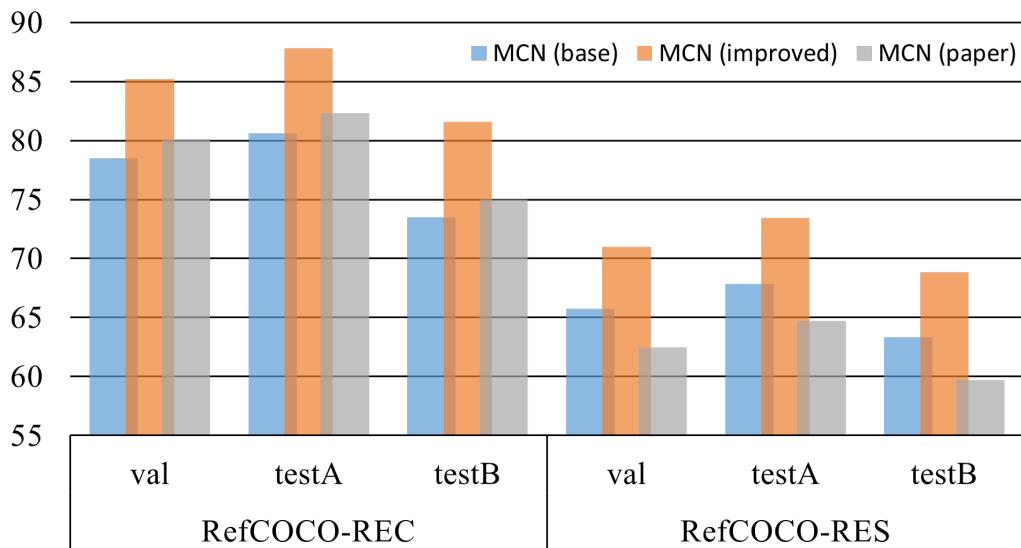


图 7. MCN 改进前后的对比以及与原论文的对比

图 7 中橙色表示改进后的模型 MCN (improved)，蓝色表示本地训练的基准模型 MCN (base)，灰色表示论文中的模型 MCN (paper)。从图 7 中可以明显看出，虽然框架不同导致模型在性能上产生差异，但改进后模型在 REC 和 RES 分支上都超过了本地训练的基准模型以及论文中的模型，初步证明了模型改进的有效性。

5.1.2 消融实验

本文在基准模型上进行增量消融实验，即逐步在基准模型上增加上文提到的改进，验证每种改进方法的有效性，实验结果如表 1 所示，其中每一行表示增加改进方法后的效果，粗体表示最优性能，下划线表示次优性能，最后一行斜体为论文中给出的模型性能。

从实验结果可以看出，随着改进方法的增加，改进模型相较于基准模型在性能上不断提升，最终在 val、testA 和 testB 上均达到最优，在 REC 任务中，改进后的模型相较基准模型分别提升了 6.74%、7.18% 和 8.10%，在 RES 任务中，分别提升了 5.25%、5.61% 和 5.49%，且均超过论文中给出的模型性能。

其中，对指称表达理解头部进行改进，只提升了模型的指称表达理解性能，对模型的指称表达分割性能没有明显影响。

Models	RefCOCO-REC (IoU@0.5)			RefCOCO-RES (mIoU)		
	val	testA	testB	val	testA	testB
MCN (base)	78.48 _{+0.00}	80.61 _{+0.00}	73.49 _{+0.00}	65.73 _{+0.00}	67.84 _{+0.00}	63.35 _{+0.00}
+ CSPDarknet53	79.12 _{+0.64}	82.34 _{+1.73}	73.94 _{+0.45}	67.17 _{+1.44}	69.82 _{+1.98}	63.94 _{+0.59}
+ Three Fusion	80.29 _{+1.81}	84.73 _{+4.12}	74.41 _{+0.92}	68.39 _{+2.66}	71.20 _{+3.36}	64.74 _{+1.39}
+ Random Resize	81.46 _{+2.98}	85.51 _{+4.90}	77.21 _{+3.72}	68.49 _{+2.76}	71.43 _{+3.59}	65.61 _{+2.26}
+ EMA	82.62 _{+4.14}	86.07 _{+5.46}	77.74 _{+4.25}	69.45 _{+3.72}	72.10 _{+4.26}	66.34 _{+2.99}
+ Anchor-free	<u>83.72</u> _{+5.24}	<u>86.69</u> _{+6.08}	<u>78.90</u> _{+5.41}	69.22 _{+3.49}	72.08 _{+4.24}	<u>66.40</u> _{+3.05}
+ YOLOv8-backbone	85.22 _{+6.74}	87.79 _{+7.18}	81.59 _{+8.10}	70.98 _{+5.25}	73.45 _{+5.61}	68.84 _{+5.49}
MCN (paper)	80.08	82.29	74.98	62.44	64.70	59.71

表 1. 对改进后的模型在 RefCOCO 上进行增量消融实验

5.2 定性分析

如图 8 所示，其中绿色框为检测框，蓝色部分为分割掩模，从左到右第一列为原始图像，第二列为基准模型的预测结果，第三列为改进后模型的预测结果，第四列为真实值，从给出的 3 个样例中可以观察到，改进后的模型对输入文本语义的理解更为准确，分割的边界也更为精确。

如图 8 中的第 1 行，指称表达为“咖啡旁别的玻璃水杯”，从基准模型的预测结果可以看出基准模型的理解出现错误，而改进后的模型给出了正确的结果；图 8 中的第 3 行，通过比较改进前后模型的分割掩模输出结果，可以明显看出改进后的模型给出了较为准确的分割边界。

从以上实验结果可以进一步得出，改进后的模型性能相较基准模型有显著提升。

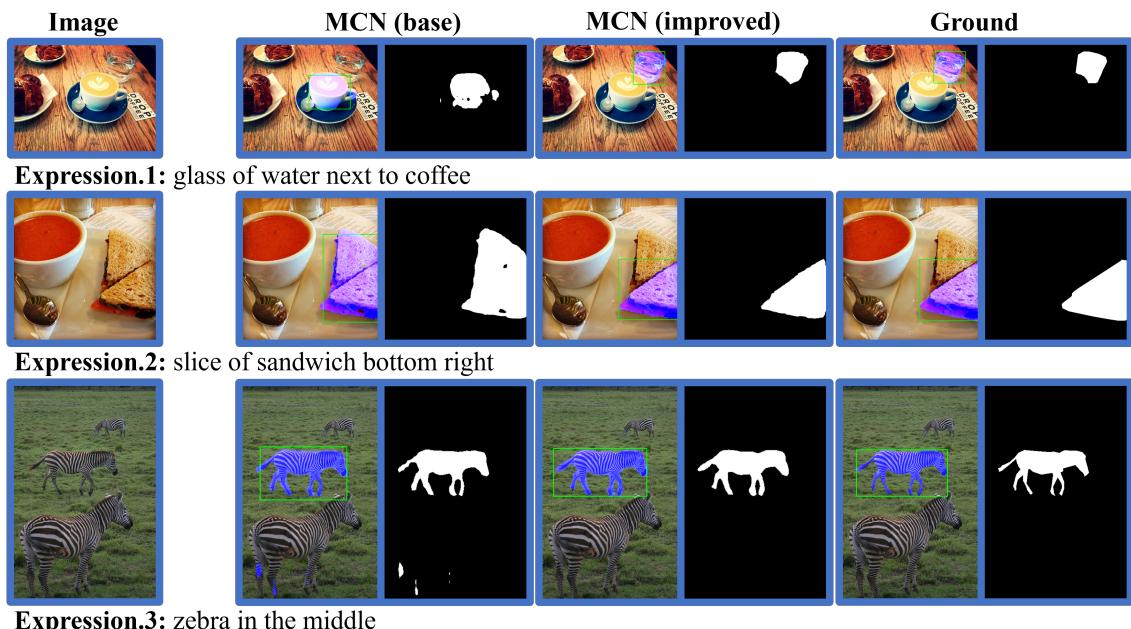


图 8. 改进前后模型测试结果样例

6 总结与展望

本文结合课程所学图像特征提取、文本特征提取等内容，对 MCN 模型进行实验。同时在原 MCN 模型的基础上，本文结合课堂所学知识与相关领域前沿论文，在图像特征提取、多尺度图像特征与文本语义特征融合、图像数据增强、模型权重优化和检测框计算方式五个方面作出改进，经过实验证明改进后的模型在 REC 和 RES 任务上的表现均有显著提升。

在指称表达分割分支中，还可以参考相关指称表达分割算法进行优化改进，在提升指称表达分割性能的同时进一步提升指称表达理解的性能。

参考文献

- [1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection, April 2020.
- [2] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-End Visual Grounding With Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021.
- [3] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-Language Transformer and Query Generation for Referring Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021.
- [4] Lei Jin, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. RefCLIP: A Universal Teacher for Weakly Supervised Referring Expression Comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2681–2690, 2023.
- [5] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. ReSTR: Convolution-Free Referring Image Segmentation Using Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18145–18154, 2022.
- [6] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring Image Segmentation via Recurrent Refinement Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2018.
- [7] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A Real-Time Cross-Modality Correlation Filtering Method for Referring Expression Comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10880–10889, 2020.
- [8] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to Assemble Neural Module Tree Networks for Visual Grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4673–4682, 2019.

- [9] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path Aggregation Network for Instance Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018.
- [10] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving Referring Expression Grounding With Cross-Modal Attention-Guided Erasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1950–1959, 2019.
- [11] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Zechao Li, Qi Tian, and Qingming Huang. Entity-Enhanced Adaptive Reconstruction Network for Weakly Supervised Referring Expression Grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3003–3018, March 2023.
- [12] Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. Relation-aware Instance Refinement for Weakly Supervised Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5612–5621, 2021.
- [13] Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. Cascade Grouped Attention Network for Referring Expression Segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM ’20, pages 1274–1282, New York, NY, USA, October 2020. Association for Computing Machinery.
- [14] Gen Luo, Yiyi Zhou, Jiamu Sun, Xiaoshuai Sun, and Rongrong Ji. A Survivor in the Era of Large-Scale Pretraining: An Empirical Study of One-Stage Referring Expression Comprehension. *IEEE Transactions on Multimedia*, pages 1–12, 2023.
- [15] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-Task Collaborative Network for Joint Referring Expression Comprehension and Segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10031–10040, Seattle, WA, USA, June 2020. IEEE.
- [16] Varun K. Nagaraja, Vlad I. Morariu, and Larry S. Davis. Modeling Context Between Objects for Referring Expression Understanding. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 792–807, Cham, 2016. Springer International Publishing.
- [17] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring Expression Comprehension: A Survey of Methods and Datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2021.
- [18] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement, April 2018.
- [19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

- [20] Wei Su, Peihan Miao, Huanzhang Dou, Gaoang Wang, Liang Qiao, Zheyang Li, and Xi Li. Language Adaptive Weight Generation for Multi-Task Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10857–10866, 2023.
- [21] Mingjie Sun, Jimin Xiao, Eng Gee Lim, Si Liu, and John Y. Goulermas. Discriminative Triad Matching and Reconstruction for Weakly Referring Expression Grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4189–4195, November 2021.
- [22] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [23] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. CSPNet: A New Backbone that can Enhance Learning Capability of CNN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1571–1580, Seattle, WA, USA, June 2020. IEEE.
- [24] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving Weakly Supervised Visual Grounding by Contrastive Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14090–14100, 2021.
- [25] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. CRIS: CLIP-Driven Referring Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11686–11695, 2022.
- [26] Li Xu, Mark He Huang, Xindi Shang, Zehuan Yuan, Ying Sun, and Jun Liu. Meta Compositional Referring Expression Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19478–19487, 2023.
- [27] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. MAttNet: Modular Attention Network for Referring Expression Comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018.
- [28] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling Context in Referring Expressions. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 69–85, Cham, 2016. Springer International Publishing.
- [29] Yiyi Zhou, Rongrong Ji, Gen Luo, Xiaoshuai Sun, Jinsong Su, Xinghao Ding, Chia-Wen Lin, and Qi Tian. A Real-Time Global Inference Network for One-Stage Referring Expression Comprehension. *IEEE Transactions on Neural Networks and Learning Systems*, 34(1):134–143, January 2023.
- [30] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. SeqTR: A Simple Yet Universal Network for

Visual Grounding. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 598–615, Cham, 2022. Springer Nature Switzerland.