基于双重注意力机制的多尺度指代目标分割方法

胡梦楠[1,2,3], 王蓉[1)*], 张文靖[1)], 张琪[1)]

[1)] (中国人民公安大学信息网络安全学院 北京 100038)
[2)] (山东警察学院治安系 济南 250200)
[3)] (山东警察学院社会治安与应急管理研究中心 济南 250200)
(20051244@ppsuc.edu.cn)

摘 要: 针对指代分割任务中视觉和语言间缺乏充分的跨模态交互、不同尺寸的目标空间和语义信息存在差异的问题, 提出了基于双重注意力机制的多尺度指代目标分割方法. 首先, 利用语言表达中不同类型的信息关键词来增强视觉和语言特征的跨模态对齐, 并使用双重注意力机制捕捉多模态特征间的依赖性, 实现模态间和模态内的交互; 其次, 利用语言特征作为引导, 从其他层次的特征中聚合与目标相关的视觉信息, 进一步增强特征表示; 然后利用双向 ConvLSTM 以自下而上和自上而下的方式逐步整合低层次的空间细节和高层次的语义信息; 最后, 利用不同膨胀因子的空洞卷积融合多尺度信息, 增加模型对不同尺度分割目标的感知能力. 此外, 在 UNC, UNC+, GRef 和 ReferIt 基准数据集上进行实验, 实验结果表明, 文中方法在 UNC, UNC+, GRef 和 ReferIt 上的 oIoU 指标分别提高了 1.81 个百分点、1.26 个百分点、0.84 个百分点和 0.32 个百分点, 广泛的消融研究也验证了所提方法中各组成部分的有效性.

关键词: 指代目标分割; 跨模态交互; 特征增强; 注意力机制; 多尺度融合
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023-00355

Multi-Scale Referring Image Segmentation Based on Dual Attention

Hu Mengnan[1,2,3], Wang Rong[1)*], Zhang Wenjing[1)], and Zhang Qi[1)]

[1)] (*School of Information and Cyber Security, People's Public Security University of China, Beijing* 100038)
[2)] (*Department of Public Order, Shandong Police College, Jinan* 250200)
[3)] (*Public Security and Emergency Management Research Center, Shandong Police College, Jinan* 250200)

Abstract: This paper proposes a multi-scale referring image segmentation method based on dual attention to solve the problem of insufficient interaction between visual and linguistic modes, as well as different structural and semantic information required by objects of different sizes. Firstly, the dual attention mechanism is used to realize the intermodal and intramodal interaction between vision and text, which enhances the ability to align visual and linguistic features accurately by using different types of information words in the expression. Secondly, using language features as guidance, useful features are selected from other levels for information exchange to further enhance feature representation. Then, dual path ConvLSTM is used to fully integrate low-level visual details and high-level semantics from bottom-up and top-down paths. Finally, multi-scale information is fused by atrous spatial pyramid pooling, increasing the perception ability of the model for different scales. Experiments on the UNC, UNC+, GRef, and ReferIt reference data sets show that the proposed method oIoU improves by 1.81 percentage points on UNC, 1.26 percentage points on UNC+, 0.84 percentage points on GRef, and 0.32 percent-

收稿日期: 2023-06-19; 修回日期: 2023-10-24. **基金项目**: 中国人民公安大学安全防范工程双一流专项(2023SYL08). **胡梦楠**(1992—), 女, 博士研究生, 讲师, CCF 会员, 主要研究方向为图像处理、计算机视觉; **王蓉**(1971—), 女, 博士, 教授, 博士生导师, CCF 会员, 论文通信作者, 主要研究方向为机器学习、图像处理; **张文靖**(1992—), 男, 博士研究生, 主要研究方向为图像处理、计算机视觉; **张琪**(1988—), 女, 博士, 讲师, CCF 会员, 主要研究方向为生物特征识别、计算机视觉.

age points on ReferIt. Extensive ablation studies have also validated the effectiveness of each component of our approach.

Key words: referring image segmentation; cross-modal interaction; feature enhancement; attention mechanism; multi-scale feature fusion

指代分割(referring image segmentation, RIS)是计算机视觉和自然语言处理交叉领域的一个新兴问题, 旨在通过理解指代表达式的内容, 对图像的特定区域或对象进行分割, 是由 Hu 等[1]在 2016 年提出. 传统的语义和实例分割需要假设一组预定义的对象类别(如, 猫、人、公共汽车等), 进而正确地分割图像中的每个语义类别或每个对象. RIS 则需要整体理解自然语言描述中复杂的语义关系, 以及图像中各种类型的视觉信息, 如对象、类别、属性、动作和位置等之间的关系. 因此, 它可以被视为交互式像素级的前景/背景分割问题, 并且其输出结果不受预定义的语义类别或对象类别的限制. 用户借助灵活且形式自由的自然语言描述, 以引导机器选择感兴趣的区域, RIS 在交互式图像编辑、语言引导的人机交互中有着广泛的应用.

然而, 由于视觉和语言之间的模态差异, 因此需要建立有效的跨模态交互机制, 实现视觉与语言特征的跨模态匹配, 是完成该任务的关键. 一般方法有 2 个阶段: 分别从文本和图像输入中提取特征, 以及通过融合多模态特征以预测分割模板. 在第 1 个阶段中, 使用循环神经网络(recurrent neural networks, RNN)[1-5]或 Transformer[6-8]对输入的语言进行编码, 使用全卷积网络(full convolutional networks, FCN)[1,5,9], Deeplab[4,6,10]或 DarkNet[3,11-12]对输入的图像进行编码, 目的是学习区分性表示. 多模态特征融合模块是现有技术关注的关键环节. 此外, 当前大多数方法多采用单尺度文本或图像信息的融合来进行分割, 很难同时覆盖不同目标大小所需要的语义和结构上下文, 导致定位错误或分割效果不佳. 因此针对视觉和语言特征缺乏足够交互、不同大小的目标所需结构和语义信息不同的问题, 本文提出了基于双重注意力机制(dual attention, DA)[13]的多尺度 RIS 方法, 通过在特征的位置和通道维度均采用自注意力机制获取全局依赖关系, 以增强特征关联, 并利用双向 ConvLSTM 和不同膨胀因子的空洞卷积, 整合低层次和高层次信息, 提高对不同尺度分割目标的感知能力.

1　相关工作

1.1　目标 RIS 方法

对于目标 RIS 任务, 直接的想法是将视觉特征信息和语言特征信息映射到相同的向量空间进行融合, 然后采用语义分割网络对融合特征进行分割预测. 早期的方法[1,4-5,14]采用“拼接-卷积”的范式, 分别将卷积神经网络(convolutional neural networks, CNN)和 RNN 提取的视觉和语言特征进行拼接融合, 以获得多模态特征; 然后, 多模态特征通过 FCN 预测分割掩码. 根据融合特征的方式不同, 可以分为联合嵌入方法、图卷积方法和基于 Transformer 方法.

1.1.1　联合嵌入方法

Hu 等[1]直接连接和融合来自 CNN 和长短期记忆(long short-term memory, LSTM)的多模态特征来生成分割掩码. Li 等[4]融合多层次的视觉特征, 进一步细化分割掩码的局部细节. Liu 等[5]提出了一个循环的多模态交互模型 LSTM, 在多个时间步中顺序融合视觉和语言特征. Margffoy-Tuay 等[14]在 Liu 等[5]基础上, 在合成模块中生成动态过滤器, 以合并语言信息; 它还使用了一个增量模块, 在特征图上进行双线性上采样和卷积, 以获得图像细节. 但以上这些方法均没有明确表述模态内和模态间的关系.

注意力机制可以通过自适应地关注图像中的重要区域和语言表达中的信息关键词, 而产生语义丰富的视觉和文本表示, 更好地融合语言和视觉特征. Shi 等[15]利用单词注意来模拟关键字感知上下文, 以增强视觉特征. Yu 等[16]利用了跨模态自注意力机制, 重点关注句子中的信息词和图像中的关键区域, 以捕获每个图像区域和每个参考词之间的长期依赖关系. Chen 等[17]提出了一个 ConvRNN 模型, 以迭代视觉文本共嵌入(co-embedding)的步骤. Hu 等[18]设计了一个双向关系推理网络, 对语言和视觉之间的关系进行建模, 实现了多模态特征之间的串行相互引导. Feng 等[19]设计了一种使用语言的共同注意机制. Luo 等[20]提出了级联注意力网络. 以上方法使用模态间或者自注意力机制, 只

学习视觉嵌入或视觉文本共嵌入，用于上下文建模，它们仍然缺乏利用语言表达式中不同类型的信息词来准确对齐视觉和语言特征的能力.

1.1.2 图卷积方法

基于图卷积网络也被用于 RIS，能够更好地处理文本描述与图像间的交互关系[21]. Huang 等[22]根据实体和属性词感知图像中的所有实体，使用图卷积网络对关系空间区域图进行关系感知推理，突出被指代目标的特征，抑制与文本描述不相关的区域，实现目标图像与文本描述之间交互关系的对应. Hui 等[23]提出在解析依赖树的词图上使用图卷积网络来指导多模态上下文的学习，根据单词间的依赖关系选择性地抑制图结构部分边的权重，使得多模态特征更加贴近文本描述. Lin 等[24]提出以自下而上的方式，利用表达式的解析依赖树进行跨模态推理.

1.1.3 基于 Transformer 方法

由于 Transformer[25]具备强大的全局上下文建模的能力, Ding 等[26]使用视觉-语言 Transformer 解码器进行跨模态特征融合，显著地提高了分割性能. Jing 等[3]通过引入位置先验，在定位阶段使用基于注意力的 Transformer 解码器，将任务解耦为“定位然后分割”方案. Yang 等[8]提出将语言信息直接整合到视觉 Transformer 编码网络中，以避免复杂使用跨模态解码器，但 Transformer 需要在大规模数据集上进行模型预训练.

1.2 DA

DA 采用自注意力机制捕获丰富的语义信息，由 2 个并行的注意力模块组成[13]. 通过位置注意力模块(position attention module, PAM)得到特征图中任意 2 个位置的依赖关系，通过通道注意力模块(channel attention module, CAM)得到任意 2 个通道间的依赖关系，最后用卷积层对 2 个注意力模块的输出进行转换，并通过一个逐元素的求和来实现特征融合，其结构如图 1 所示.

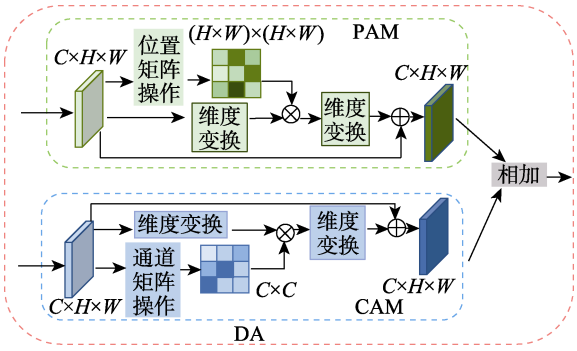


图 1 DA 结构图

1.2.1 PAM

PAM 通过利用 2 个特征点之间的关联，将更广泛的上下文信息编码为局部特征，进而相互增强各自特征表示能力，其结构如图 2 所示.

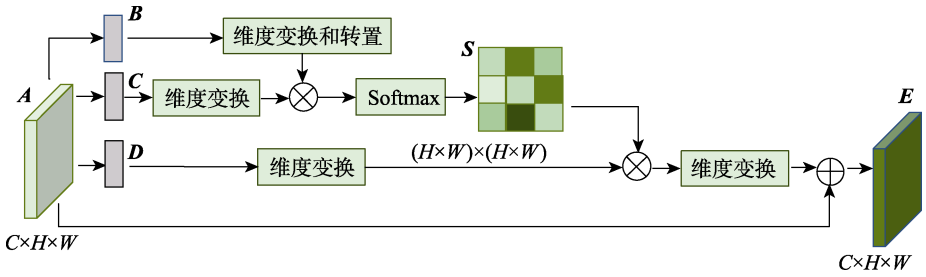


图 2 PAM 结构

首先，将原始特征 $\boldsymbol{A}\in\mathbb{R}^{C\times H\times W}$ 通过一个带有 BN 层和 ReLU 层的卷积层降维，获得特征 $\boldsymbol{B}\in\mathbb{R}^{C\times H\times W}$，$\boldsymbol{C}\in\mathbb{R}^{C\times H\times W}$ 和 $\boldsymbol{D}\in\mathbb{R}^{C\times H\times W}$；并将它们的维度分别改变为特征 $\boldsymbol{B},\boldsymbol{C},\boldsymbol{D}$ 的维度，即 $\boldsymbol{B}\in\mathbb{R}^{N\times C}$，$\boldsymbol{C}\in\mathbb{R}^{C\times N}$ 和 $\boldsymbol{D}\in\mathbb{R}^{C\times N}$. 其中，$N=H\times W$.

其次，将 $\boldsymbol{B}$ 的转置和 $\boldsymbol{C}$ 相乘得到任意 2 点特征之间的关联强度矩阵，通过 Softmax 层获得每个位置对其他未知的空间注意力映射 $\boldsymbol{S}\in\mathbb{R}^{N\times N}$，第 i 个和第 j 位置之间的相似性表示为

$$s_{ji}=\exp\left(B_i\cdot C_i\right)\Big/\sum_{i=1}^{N}\exp\left(B_i\cdot C_j\right) \tag{1}$$

其值越大，表示 2 个特征之间越相似.

最后，将 $\boldsymbol{S}$ 转置后与 $\boldsymbol{D}$ 相乘，再与原始特征 $\boldsymbol{A}$ 逐元素相加，得到最后的输出 $\boldsymbol{E}\in\mathbb{R}^{C\times H\times W}$，即

$$E_j=\alpha\sum_{i=1}^{N}\left(s_{ji}D_i\right)+A_j \tag{2}$$

其中，α 表示尺度系数，其初始化值为 0，通过逐步的学习可得到更大的权重.

1.2.2 CAM

CAM 通过整合所有通道映射之间的相关特征，显式地建模通道之间的依赖关系，增强通道下特定语义的响应能力，其结构如图 3 所示. 与 PAM 不同的是，为了保留不同通道映射之间的关系，原

始特征 $\boldsymbol{A}\in\mathbb{R}^{C\times H\times W}$ 没有使用卷积层降维，直接将其维度变换为 $\mathbb{R}^{C\times N}$ 和 $\mathbb{R}^{N\times C}$ 后，将其矩阵相乘得到 2 个通道的关联强度，再通过 Softmax 层得到通道注意力映射 $\boldsymbol{X}\in\mathbb{R}^{C\times C}$，第 i 个对第 j 个通道的影响表示为

$$x_{ji}=\exp\left(A_i\cdot A_j\right)\Big/\sum_{i=1}^{C}\exp\left(A_i\cdot A_j\right) \tag{3}$$

最后将通道注意力映射的响应值 x_{ji} 作为加权，对维度变换后的 $\boldsymbol{A}\in\mathbb{R}^{C\times N}$ 进行加权融合，得到最后的输出 $\boldsymbol{E}\in\mathbb{R}^{C\times H\times W}$，即

$$E_j=\beta\sum_{i=1}^{N}\left(x_{ji}A_i\right)+A_j \tag{4}$$

其中，β 为尺度系数，初始化值为 0，并通过逐步地学习可得到更大的权重.

为了进一步获得全局依赖关系的特征，将 2 个模块的输出结果相加并融合，获得最终的特征用于像素点的分类.

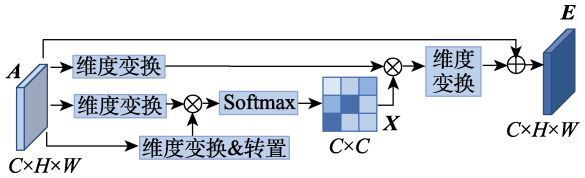


图 3　CAM 结构

2　模型网络结构

本文提出的基于 DA 机制的多尺度 RIS 方法由 4 个模块组成，分别是特征提取模块、特征融合模块、语言引导的特征增强模块和掩码分割模块，结构如图 4 所示.

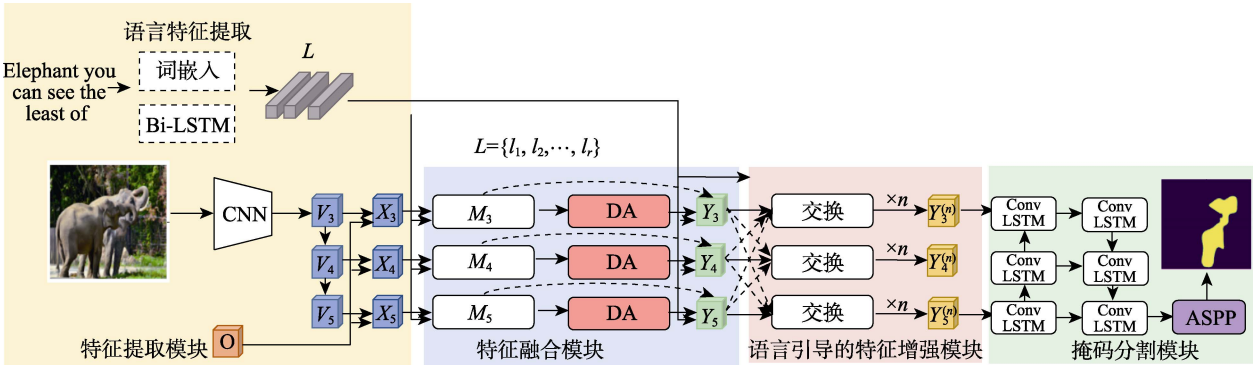


图 4　本文方法结构

在特征提取模块中，使用 Deeplab-101 作为视觉特征提取主干网络，并将 ResNet 的第 3~5 层的输出用于多级特征融合；Bi-LSTM 用于作为语言特征提取主干网络，并将每个单词表示为单词嵌入向量. 在特征融合模块，通过 DA 捕捉空间和通道维度中的全局特征依赖关系，增强语言特征与视觉特征的对齐能力，实现视觉-语言特征融合. 在特征增强模块，利用语言特征作为引导，从其他层次选择有用的特征进行信息交流，进一步增强特征表示. 在掩码分割模块，使用从上到下、从下到上的双向的 ConvLSTM，逐步整合了低层次的视觉细节和高层次的语义，然后通过单个空洞空间金字塔池化(atrous spatial pyramid pooling, ASPP)[10]解码器，利用不同膨胀因子的空洞卷积融合多尺度信息，实现掩码的精准预测.

2.1　特征提取模块

对于图像 $I\in\mathbb{R}^{H\times W\times C_v}$，通过基于 ResNet 的 Deeplab-101 网络提取多级视觉特征 $V\in\mathbb{R}^{H\times W\times C_v}$，并调整为相同的大小. 其中，$H$，$W$，$C_v$ 分别表示特征高度、宽度和通道尺寸；$\{V_3,V_4,V_5\}$ 分别对应 ResNet 的第 3~5 层的输出. 然后使用 1×1 卷积与 8 维的空间坐标特征 $O\in\mathbb{R}^{H\times W\times 8}$ 进行融合，得到 3 级视觉特征 $\{X_3,X_4,X_5\}$.

对于由 T 个单词组成的指代表达式 $\{w_t\}_{t=1}^{T}$，把词嵌入 $\{e_t\}_{t=1}^{T}$ 添加到 Bi-LSTM 中，用于提取语言特征 $L=\{l_1,l_2,\cdots,l_T\}$. 其中，T 是表达式的长度；$l_i\in\mathbb{R}^{C_l}(i\in\{1,2,\cdots,T\})$ 表示第 i 个单词的特征.

2.2　特征融合模块

特征融合部分首先将单词分为实体词、属性词、关系词和非必要词，通过一个 4 维向量 $\boldsymbol{p}_t=\left[\boldsymbol{p}_t^{\text{ent}},\boldsymbol{p}_t^{\text{attr}},\boldsymbol{p}_t^{\text{rel}},\boldsymbol{p}_t^{\text{un}}\right]\in\mathbb{R}^4$ 表示每一个单词 t 属于 4 种类型词的概率[22]，采用计算表达式所有单

词的加权，得到实体词的全局语义上下文 $\boldsymbol{q}\in\mathbb{R}^{C_l}$，并将所有必要单词的特征组合到向量 $\boldsymbol{s}\in\mathbb{R}^{C_l}$ 中. 最后采用双线性融合方法——MUTAN[27]，将 $\boldsymbol{q}$ 与每个空间区域的视觉特征融合，得到多模态特征 $\boldsymbol{M}\in\mathbb{R}^{H\times W\times C_m}$. 具体计算公式为

$$\boldsymbol{p}_t=\mathrm{Softmax}\left(\boldsymbol{W}_2\sigma\left(\boldsymbol{W}_1 l_t+\boldsymbol{b}_1\right)+\boldsymbol{b}_2\right) \tag{5}$$

$$\boldsymbol{q}=\sum_{t=1}^{T}\left(\boldsymbol{p}_{\mathrm{t}}^{\mathrm{ent}}+\boldsymbol{p}_{\mathrm{t}}^{\mathrm{attr}}\right)l_t \tag{6}$$

$$\boldsymbol{s}=\sum_{t=0}^{T}\left(\boldsymbol{p}_{\mathrm{t}}^{\mathrm{ent}}+\boldsymbol{p}_{\mathrm{t}}^{\mathrm{attr}}+\boldsymbol{p}_{\mathrm{t}}^{\mathrm{rel}}\right)l_t \tag{7}$$

$$M_i=\left(\boldsymbol{q}\boldsymbol{W}_{3i}\right)\odot\left(X\boldsymbol{W}_{4i}\right) \tag{8}$$

$$\boldsymbol{M}=\sum_{i=1}^{r}M_i \tag{9}$$

其中，$\boldsymbol{W}_1\in\mathbb{R}^{C_n\times C_l}$，$\boldsymbol{W}_2\in\mathbb{R}^{4\times C_n}$，$\boldsymbol{W}_{3i}\in\mathbb{R}^{C_l\times C_m}$，$\boldsymbol{W}_{4i}\in\mathbb{R}^{C_v\times C_m}$，$\boldsymbol{b}_1\in\mathbb{R}^{C_n}$ 和 $\boldsymbol{b}_2\in\mathbb{R}^4$ 为可学习参数；$\sigma(\cdot)$ 为 Sigmoid 激活函数；r 为超参数；$\odot$ 为元素乘积.

为了将语义引导的低级特征引入高级跨模态特征，增强语言特征与视觉特征的对齐能力，将多模态特征 $\boldsymbol{M}\in\mathbb{R}^{H\times W\times C_m}$ 通过 DA 捕捉空间和通道维度中的全局特征依赖关系. 其中，PAM 学习特征的空间相互依赖性，CAM 模拟通道相互依赖性. 将语义引导的低级特征引入高级跨模态特征，得到基于注意力机制的多模态特征 $\bar{\boldsymbol{M}}^{\mathrm{DA}}\in\mathbb{R}^{H\times W\times C_m}$，并将其与多模态特征 $\boldsymbol{M}$、语言特征 L 沿通道维度拼接一起后，进行 1×1 卷积，得到输出特征 $\boldsymbol{Y}\in\mathbb{R}^{H\times W\times C_m}$.

2.3 语言引导的特征增强模块

为了能够充分地利用各个层次有用的语义特征，进一步增强特征表示，将特征融合模块的 3 层特征 Y_3，Y_4，Y_5 和单词特征 $\left[l_1,l_2,\cdots,l_T\right]$ 作为输入. 通过加权全局池化提取 $Y_i^{(k-1)}$ 的全局向量 $\boldsymbol{g}_i^{(k-1)}\in\mathbb{R}^{C_m}$ 后，用一个全连接层融合 $\boldsymbol{s}$ 和 $\boldsymbol{g}_i^{(k-1)}$ 得到包含多模态上下文 $Y_i^{(k-1)}$ 的上下文向量 $\boldsymbol{c}_i^{(k-1)}$，最后从其他 2 个级别的特征中选择与第 i 级相关的信息，形成 k 轮的第 i 级特征. 经过 n 轮特征交换以后，得到输出 $Y_3^{(n)},Y_4^{(n)},Y_5^{(n)}$. 具体过程为

$$\boldsymbol{g}_i^{(k-1)}=\boldsymbol{\Lambda}_i^{(k-1)}Y_i^{(k-1)} \tag{10}$$

$$\boldsymbol{\Lambda}_i^{(k-1)}=\left(\boldsymbol{s}\boldsymbol{W}_8\right)\left(Y_i^{(k-1)}\boldsymbol{W}_9\right)^T \tag{11}$$

$$Y_i^{(k)}=\begin{cases}Y_i^{(k-1)}+\sum\limits_{j\in\{3,4,5\}\setminus\{i\}}\sigma\left(\boldsymbol{c}_i^{(k-1)}\right)\odot Y_j^{(k-1)}, & k\geqslant 1\\ Y_i, & k=0\end{cases} \tag{12}$$

其中，$\boldsymbol{\Lambda}_i^{(k-1)}\in\mathbb{R}^{HW}$ 为权重矩阵；$\boldsymbol{W}_8\in\mathbb{R}^{C_l\times C_h}$ 和 $\boldsymbol{W}_9\in\mathbb{R}^{C_m\times C_h}$ 为变换矩阵；$\sigma(\cdot)$ 为 Sigmoid 激活函数.

2.4 掩码分割模块

不同层次特征的整合，可以显著地提高 RIS 的性能[4]. 因此在掩码分割模块，通过 ConvLSTM 自下而上和自上而下的方式融合特征，增强模块的输出 $\left\{Y_5^{(n)},Y_4^{(n)},Y_3^{(n)},Y_4^{(n)},Y_5^{(n)}\right\}$. 第 1 次使用从下而上的路径，过程如

$$\left(H_i,C_i\right)=\mathrm{ConvLSTM}\left(Y_i^{(n)},H_{i+1},C_{i+1}\right),i=3,4 \tag{13}$$

顺序集成低级别特征，用空间细节来补充高级特征，以细化掩码的局部部分，提高对浅层特征的利用. 但当低级特征集成越来越多时，对于整体轮廓至关重要的高级特征会被逐渐淡化，因此使用自上而下的路径，如

$$\left(H_i,C_i\right)=\mathrm{ConvLSTM}\left(Y_i^{(n)},H_{i-1},C_{i-1}\right),i=4,5 \tag{14}$$

以补充更多的语义多模态信息. 其中，H_i 表示每个时间步长的隐藏状态；C_i 表示每个时间步长的单元状态，并将式(14)最后一个时间步长的隐藏状态 H_{fin} 作为 ASPP 解码器的输入.

在不改变特征图大小的前提下，ASPP 可以通过采用不同采样率的空洞卷积进行采样，以增大网络的感受野，从而增强网络获取多尺度上下文的能力. 结构如图 5 所示.

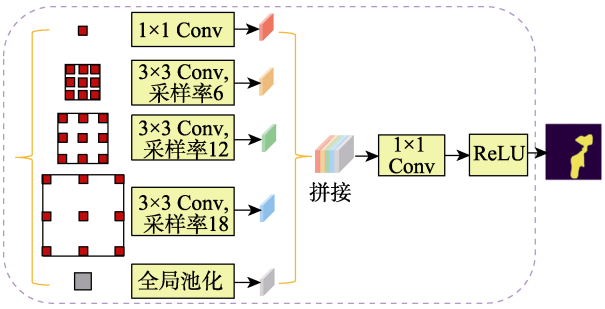


图 5 ASPP 结构图

对于输入 H_{fin}，ASPP 模块分为 3 个部分组成，分别为1×1的卷积(最上层红色)、3 个3×3的空洞卷积(中间层橘色、绿色、蓝色)、全局平均池化层(最下层灰色). 将上述不同尺度得到的特征在通道维度拼接到一起，扩大通道数；再通过1×1卷积将通道数降低到预期的数值，从而在不增大计算量的前提下扩大滤波器的感受野，增加模型对不同尺度分割目标的感知能力，得到输出

$$P_{\mathrm{mask}}=R\left(C_{1\times 1}\left(\mathrm{concat}\left(H_{\mathrm{fin}}^{1\times 1},H_{\mathrm{fin}}^{3\times 3},H_{\mathrm{fin}}^{3\times 3},H_{\mathrm{fin}}^{3\times 3},H_{\mathrm{fin}}^{\mathrm{GP}}\right)\right)\right) \tag{15}$$

同时使用标准二值交叉熵损失作为损失函数, 即

$$L_{\text{loss}} = -\frac{1}{\Omega}\sum_{m=1}^{\Omega}\big(G(m)\ln P_{\text{mask}}(m) + (1-G(m))\ln\big(1-P_{\text{mask}}(m)\big)\big) \tag{16}$$

其中, P_{mask} 为概率映射, 表示在分割掩码中每个像素是前景的可能性; $R(\cdot)$ 为 ReLU 激活函数; G 为人工标注的目标真实掩码; Ω 为图像中像素的集合; m 为集合中第 m 个像素.

3　仿真实验与结果分析

3.1　实验设置

3.1.1　数据集

本文在 4 个数据集 UNC[28], UNC+[28], GRef[29] 和 ReferIt[30] 上进行训练与测试. 其中, UNC, UNC+和 GRef 这 3 个数据集基于 MS COCO[31]构建, 分别包含 19 994, 19 992 和 26 711 幅图像, 142 209, 141 564 和 104 560 个语言表达式, 50 000, 49 856 和 54 822 个分割对象. ReferIt 数据集取自 IAPR TC-12[32]数据集, 包含 19 894 幅图像和 130 525 个语言表达式, 涉及 96 654 个分割对象.

3.1.2　实施细节

本文实验在 NVIDIA RTX2070 GPU 上使用 Tensorflow 框架进行训练和测试.

输入图像大小为 320×320 像素, 特征通道尺寸 $C_v = C_l = C_m = C_h = 1000$, 设置 ConvLSTM 单元尺寸为 500, 双线性融合的超参数 r=5, 特征交换轮数为 3, 初始学习率为 2.5×10^{-4}, 权重衰减为 5×10^{-4}. 使用 Adam 优化器进行优化, 批量大小设置为 1, 训练迭代次数达到 700 000 时停止训练, 并使用 DenseCRF[33]对分割结果进行优化.

3.1.3　评估指标

本文采用全局交并比(overall intersection-over-union, oIoU)和精度百分比(Prec@X)作为模型的评估指标[1,33]. oIoU 表示所有测试样本的预测值和真实值之间的总相交区域和总并集区域的比率; Prec@X 计算测试集中预测值 IoU 高于阈值 X 百分比, 其中 $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$.

3.2　实验与结果分析

3.2.1　性能比较

为了验证模型的有效性, 将本文与其他方法进行了比较, 结果如表 1 所示.

表 1　不同模型在 4 个数据集上的 oIoU 性能比较　　%

方法	分辨率/像素	主干网络	UNC			UNC+			GRef	ReferIt
			val	testA	testB	val	testA	testB	val	test*
LSTM-CNN[1]	320×320	ResNet101(voc-seg)							28.14	48.03
RMI[5]	320×320	ResNet101(voc-seg)	44.33	44.74	44.63	29.91	30.37	29.43	34.40	57.34
DMN[14]	320×320	DPN92(ImageNet)	49.78	54.83	45.13	38.88	44.22	32.39	36.76	52.81
CMPC+DCRF[22]	320×320	ResNet101(voc-seg)	61.36	64.53	59.64	49.56	53.44	43.23	49.05	65.53
LSCM+DCRF[23]	320×320	ResNet101(voc-seg)	61.47	64.99	59.55	49.34	53.12	43.50	48.05	66.57
BRINet+DCRF[18]	320×320	ResNet101(voc-seg)	61.35	63.37	59.57	48.57	52.87	42.12	48.04	
SANet[24]	320×320	ResNet101(voc-seg)	61.84	64.95	57.43	50.38	55.36	42.74	44.53	65.88
CGAN[20]	320×320	ResNet101(voc-seg)	59.25	62.37	53.94	46.16	51.37	38.24	46.54	
CMF+DCRF[34]	320×320	ResNet101(voc-seg)	62.03	64.70	60.10	49.62	53.85	42.22	45.65	64.72
本文	320×320	ResNet101(voc-seg)	62.44	65.06	61.19	50.22	54.56	44.34	49.16	66.08
本文+DCRF	320×320	ResNet101(voc-seg)	**63.35**	**66.05**	**61.91**	**50.97**	**55.43**	**44.76**	**49.99**	**66.09**
	416×416	ResNet101(voc-seg)	**65.16**	**67.86**	**63.03**	**52.13**	**57.01**	**45.47**	**51.13**	**66.63**

注. DCRF 表示 DenseCRF 后处理; 粗体表示最优值.

由表 1 可知, 除了 DMN[14]采用在 ImageNet 上预训练的 DPN92[35]作为视觉特征提取主干网络外, 其余方法都使用在 Pascal VOC 数据集上预训练的 Deeplab ResNet101 作为主干网络. 此外, 输入图像的分辨率均设置为 320×320 像素. 因此, 为了公平比较, 本文采用相同的实验设置.

在不使用 DenseCRF 后处理的情况下, 除了 UNC+val, UNC+testA 和 ReferIt 数据集之外, 本文模型在其他数据集上达到最优, 提升了 0.11~1.60 个百分点. 在使用 DenseCRF 后处理的情况下, 在所有数据集上均实现了最优, 与对比方法中最优的方法相比, 本文方法在 UNC val, testA, testB 上 oIoU 分别提高 1.32 个百分点、1.06 个百分点、1.81 个百分点; 在 UNC+val, testA, testB 上分别提高 0.59 个百分点、0.07 个百分点和 1.26 个百分点; 在 Gref 上提高 0.94 个百分点; 在 ReferIt 上提高 0.32 个百分点. 在更具挑战性的数据集上, 如没有位置词的 UNC+、包含更长更丰富查询表达式的 GRef、表达式内容更加丰富的 ReferIt, 本文方法能够探索表达式中隐含的结构信息, 捕获多模态特征的长期依赖

关系, 从而有效地在语言-视觉特征中进行推理.

本文还将输入图像分辨率增加为 416×416 像素, 表 1 所示结果表明, 图像分辨率影响分割效果, 分辨率越大, 分割效果越好.

3.2.2 消融实验

本文在 UNC val 数据集上进行消融实验, 验证不同模块的有效性. 输入图像分辨率为 320×320 像素, 所有消融实验采用第 2.1 节描述的特征提取过程. 基线模型通过 concat 连接 DeepLab-101 提取的视觉特征和用 LSTM 提取的语言特征, 然后使用 MUTAN 进行特征融合, 并对它们的融合特征进行预测. DA 表示在特征融合模块使用双重注意力, CE 表示使用文本引导的特征增强模块, MUL 表示使用多尺度掩码分割模块, 实验结果及可视化分别如表 2 和表 3 所示. 实验结果表明, 第 2~7 号退化模型以及第 8 号本文模型的性能均优于 1 号基线模型.

表 2 UNC val 数据集消融实验结果 %

模型	DA	CE	MUl	prec@0.5	prec@0.6	prec@0.7	prec@0.8	prec@0.9	oIoU
1				53.13	43.88	32.08	16.39	3.80	46.61
2	√			70.42	61.85	49.34	28.78	5.38	59.79
3		√		68.92	61.95	51.58	34.89	9.04	59.46
4			√	70.42	62.19	49.65	29.57	5.60	59.76
5	√	√		70.88	63.47	53.30	36.27	9.30	60.63
6		√	√	71.55	64.16	53.80	36.62	9.52	60.70
7	√		√	72.59	67.15	57.25	41.35	11.82	62.57
8	√	√	√	75.56	59.38	59.65	42.32	11.96	63.35

表 3 不同方法部分成功分割结果

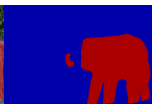
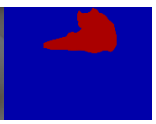
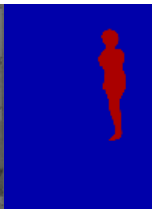
文本描述	输入图像	1 号基线模型	2 号退化模型	5 号退化模型	CMPC	本文	真实掩码
“Guy in purple”							
“Background person blue shirt”							
“Baby giraffe”							
“Bigger broccoli”							
“Man in black bent over”							
“Top part of table”							
“Tennis player in white shirt”							
“Adult in white shirt”							
“Apple on the bottom to the right of the orange in middle”							

3.2.3　可视化分析

如表 3 所示，本文方法能够更加准确地定位文本描述的对象，并且在边缘、细节方面更加清晰、准确，锯齿状现象有明显改善.

失败案例如表 4 所示. 第 1 种情形为描述对象不唯一. “Bowl behind the others can only see part”(只能看到一部分的后面的碗)图像中后面有 2 个碗，且都只能看到一部分，因此造成定位分割失败. 第 2 种情形为描述对象与周围背景过于相似、难以区分. “Baby on right that you can see most of”(右边能看到最大的小象)，“Back broccoli”(后面的西蓝花)，“Adult zebra”(成年斑马)3 幅图像中，大象与小象、2 个西蓝花、成年斑马与小斑马边界不清晰，难以准确地分割. 第 3 种情形为描述较为复杂. “Arm of the man with backpack”(背背包男人的手臂)，本文方法能够找到背背包的男人，但是并没有精准地将手臂部分分割出来；“A lady with a black swim suit holding a umbrella on the beach”(沙滩上，一个拿着雨伞、穿黑色泳衣的女人)本文方法能够找到描述目标，但是分割结果缺少下半身部分. 文本中出现“背包” “男人” “手臂” “雨伞” “泳衣” “女人” “沙滩”等多种描述词语，表明本文方法对于复杂描述的分割有待提高.

表 4　本文方法失败分割结果

文本描述	输入图像	分割结果	真实掩码
“Bowl behind the others can only see part”			
“Baby on right that you can see most of”			
“Back broccoli”			
“Adult zebra”			
“Arm of the man with backpack”			
“a lady with a black swim suit holding a umbrella on the beach”			

4　结　语

本文针对视觉和语言特征缺乏足够交互、不同大小的目标所需结构和语义信息不同的问题，提出了一种基于 DA 机制的多尺度跨模态 RIS 方法. 首先，通过对多模态特征在位置维度和通道维度采用自注意力机制，自适应地聚合上下文信息，增强特征对齐能力；其次，以语言特征为引导，从其他层次选择有用的视觉信息，进一步增强特征表示；最后，采用双向 ConvLSTM 和 ASPP，聚合低层次的空间细节、高层次的语义信息以及不同尺度信息，提高对不同尺寸分割目标的感知能力. 在 UNC，UNC+，GRef 和 ReferIt 数据集上测试本文方法，其 oIoU 指标均有提升，边缘锯齿状现象有明显改善；但对于分割目标与背景过于相似、描述对象过于复杂等情况，其分割准确度有待进一步提高.

参考文献(References):

[1] Hu R H, Rohrbach M, Darrell T. Segmentation from natural language expressions[C] //Proceedings of the 14th European Conference on Computer Vision. Heidelberg: Springer, 2016: 108-124

[2] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780

[3] Jing Y, Kong T, Wang W, *et al*. Locate then segment: a strong pipeline for referring image segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 9853-9862

[4] Li R Y, Li K C, Kuo Y C, *et al*. Referring image segmentation via recurrent refinement networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 5745-5753

[5] Liu C X, Lin Z, Shen X H, *et al*. Recurrent multimodal interaction for referring image segmentation[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 1280-1289

[6] Bellver M, Ventura C, Silberer C, *et al*. RefVOS: a closer look at referring expressions for video object segmentation[OL]. [2023-06-19]. https://arxiv.org/pdf/2010.00263v1.pdf

[7] Devlin J, Chang M W, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding[OL]. [2023-06-19]. https://arxiv.org/pdf/1810.04805.pdf

[8] Yang Z, Wang J Q, Tang Y S, *et al*. LAVT: language-aware vision transformer for referring image segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and

Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 18134-18144

[9] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3431-3440

[10] Chen L C, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation[OL]. [2023-06-19]. https://arxiv.org/pdf/1706.05587.pdf

[11] Luo G, Zhou Y Y, Sun X S, *et al*. Multi-task collaborative network for joint referring expression comprehension and segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 10031-10040

[12] Redmon J, Farhadi A. YOLOv3: an incremental improvement[OL]. [2023-06-21]. https://arxiv.org/pdf/1804.02767.pdf

[13] Fu J, Liu J, Tian H J, *et al*. Dual attention network for scene segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 3141-3149

[14] Margffoy-Tuay E, Pérez J C, Botero E, *et al*. Dynamic multimodal instance segmentation guided by natural language queries[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 656-672

[15] Shi H C, Li H L, Meng F M, *et al*. Key-word-aware network for referring expression image segmentation[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 38-54

[16] Yu L C, Lin Z, Shen X H, *et al*. MAttNet: modular attention network for referring expression comprehension[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 1307-1315

[17] Chen D J, Jia S H, Lo Y C, *et al*. See-through-text grouping for referring image segmentation[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 7453-7462

[18] Hu Z W, Feng G, Sun J Y, *et al*. Bi-directional relationship inferring network for referring image segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 4423-4432

[19] Feng G, Hu Z W, Zhang L H, *et al*. Encoder fusion network with co-attention embedding for referring image segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 15501-15510

[20] Luo G, Zhou Y Y, Ji R R, *et al*. Cascade grouped attention network for referring expression segmentation[C] //Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 1274-1282

[21] Qiu Shuang, Zhao Yao, Wei Shikui. A survey of referring image segmentation[J]. Journal of Signal Processing, 2022, 38(6): 1144-1154(in Chinese)
(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)

[22] Huang S F, Hui T R, Liu S, *et al*. Referring image segmentation via cross-modal progressive comprehension[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 10485-10494

[23] Hui T R, Liu S, Huang S F, *et al*. Linguistic structure guided context modeling for referring image segmentation[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 59-75

[24] Lin L, Yan P X, Xu X Q, *et al*. Structured attention network for referring image segmentation[J]. IEEE Transactions on Multimedia, 2022, 24: 1922-1932

[25] 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. Red Hook: Curran Associates Inc, 2017: 6000-6010

[26] Ding H H, Liu C, Wang S C, *et al*. Vision-language transformer and query generation for referring segmentation[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 16301-16310

[27] Ben-Younes H, Cadene R, Cord M, *et al*. MUTAN: multimodal tucker fusion for visual question answering[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2631-2639

[28] Yu L C, Poirson P, Yang S, *et al*. Modeling context in referring expressions[C] //Proceedings of the 14th European Conference on Computer Vision. Heidelberg: Springer, 2016: 69-85

[29] Mao J H, Huang J, Toshev A, *et al*. Generation and comprehension of unambiguous object descriptions[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 11-20

[30] Kazemzadeh S, Ordonez V, Matten M, *et al*. ReferItGame: referring to objects in photographs of natural scenes[C] //Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 787-798

[31] Lin T Y, Maire M, Belongie S, *et al*. Microsoft COCO: common objects in context[C] //Proceedings of the 13th European Conference on Computer Vision. Heidelberg: Springer, 2014: 740-755

[32] Escalante H J, Hernández C A, Gonzalez J A, *et al*. The segmented and annotated IAPR TC-12 benchmark[J]. Computer Vision and Image Understanding, 2010, 114(4): 419-428

[33] Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with Gaussian edge potentials[C] //Proceedings of the 24th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2011: 109-117

[34] Yang J H, Huang Y, Ma Z Y, *et al*. CMF: cascaded multi-model fusion for referring image segmentation[C] //Proceedings of the IEEE International Conference on Image Processing. Los Alamitos: IEEE Computer Society Press, 2021: 2289-2293

[35] Chen Y P, Li J N, Xiao H X, *et al*. Dual path networks[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2017: 4470-4478