多粒度特征提取与损失优化的图像指代分割方法

张龙飞[1)] 孟思雨[1)] 牛春翔[1)] 毕艺瀚[1)] 王 蓉[1)] 李 晨[2)] 宋 杰[2)]

[1)](中国人民公安大学信息网络安全学院 北京 100038)

[2)](北京市公安局西城分局 北京 100000)

摘 要 图像指代分割旨在根据给定的文本描述从图像中分割特定的目标区域。现有方法虽已取得一定进展，但仍存在以下问题：一是全局上下文特征与局部细节特征尚未得到充分提取，特征表达能力不足；二是预测目标区域边界与真实目标区域边界的匹配尚未得到充分优化，特征判别性较差。针对上述问题，本文提出一种多粒度特征提取与损失优化的图像指代分割方法。首先，构建多粒度特征提取模块，通过全局分支与局部分支的交互，实现全局上下文特征与局部细节特征的深度融合，提升特征表达多样性；然后，引入动态权重分配机制，通过对多粒度特征的权重进行动态调整，减少冗余信息干扰，提升特征鲁棒性；最后，设计联合损失函数，通过优化各损失函数的比例，提升真实目标区域反向传播得到的梯度，增强特征判别性。本文在三个图像指代分割数据集 RefCOCO、RefCOCO+与 G-Ref 上开展了大量实验，与其他方法比较，整体交并比(OIoU)指标分别提升了 1.23%、1.17%与 2.09%，验证了本文方法的有效性。

关键词 图像指代分割；跨模态融合；特征增强；注意力机制；损失函数

中图法分类号 TP391

Multi-granularity feature extraction and Loss Optimization Referring Image Segmentation Method

ZHANG Long-Fei[1)] MENG Si-Yu[1)] NIU Chun-Xiang[1)] BI Yi-Han[1)] WANG Rong[1)] LI Chen[2)] SONG Jie[2)]

[1)](College of Information and Cyber Security, People's Public Security University of China, Beijing 100038)

[2)](Xicheng Branch of Beijing Municipal Public Security Bureau, Beijing 100000)

Abstract Referring image segmentation aims to accurately segment specific target object regions in an image based on a given natural language description. This task requires a fine-grained understanding of both visual content and linguistic semantics, as well as precise alignment between multimodal features. Although existing referring image segmentation methods have achieved notable progress, several critical challenges remain. First, many approaches fail to sufficiently capture and integrate global contextual information and local detailed features, resulting in limited feature representation capability and incomplete semantic understanding of complex scenes. Second, the optimization of the predicted target boundaries is often inadequate, leading to imprecise segmentation results and weak feature discrimination, particularly in scenarios involving small objects, complex backgrounds, or ambiguous textual descriptions. To address these challenges, this paper proposes a novel referring image segmentation method based on multi-granularity feature extraction and loss optimization. The proposed method aims to enhance feature representation diversity, improve boundary alignment accuracy, and

本课题得到国家自然科学基金（62076246）；中央高校基本科研业务费专项资金项目（2024JKF11）资助. 张龙飞，硕士研究生，CCF学生会员，主要研究领域为计算机视觉、指代分割等.孟思雨，硕士研究生,主要研究领域为计算机视觉、深度伪造检测等. 牛春翔，硕士研究生，主要研究领域为计算机视觉、异常行为识别等. 毕艺瀚，博士研究生,主要研究领域为计算机视觉、行人重识别等. 王蓉（**通信作者**），博士研究生，教授， CCF会员,主要研究领域为计算机视觉、人工智能与模式识别等. 李晨，硕士研究生,主要研究领域为计算机视觉、计算机语言、机器学习等. 宋杰，本科，主要研究领域为计算机视觉、计算机语言、机器学习等.

strengthen feature discrimination through a unified and effective framework. Specifically, a multi-granularity feature extraction module is designed to deeply integrate global context features and local detail features. This module consists of a global branch and a local branch, which interact with each other to capture complementary information at different semantic levels. The global branch focuses on modeling high-level semantic context and long-range dependencies, enabling the model to understand the overall scene structure and object relationships, while the local branch emphasizes fine-grained spatial details and boundary information, which are crucial for accurate object localization and segmentation. Furthermore, to mitigate the influence of redundant or irrelevant information introduced by multi-granularity features, a dynamic weight allocation mechanism is introduced. This mechanism adaptively adjusts the contribution of features at different granularities according to their relevance to the target object. By dynamically reweighting feature representations during training, the proposed method effectively suppresses noise, enhances informative features, and improves overall feature robustness. As a result, the model is better able to focus on the most discriminative visual cues that correspond to the textual description. In addition, a joint loss function is designed to further optimize the segmentation performance, particularly at object boundaries. The joint loss combines multiple complementary loss terms, each emphasizing different aspects of the segmentation task, such as region consistency and boundary accuracy. By carefully optimizing the proportion of each loss component, the proposed loss function improves gradient propagation in the real target regions during backpropagation, thereby enhancing feature discrimination and boundary refinement. This loss optimization strategy ensures that the model pays greater attention to difficult regions, such as object edges and visually similar background areas. Extensive experiments are conducted on three widely used referring image segmentation benchmarks, namely RefCOCO, RefCOCO+, and G-Ref. The experimental results demonstrate that the proposed method consistently outperforms existing state-of-the-art approaches. In particular, the Overall Intersection over Union (OIoU) metric is improved by 1.23%, 1.17%, and 2.09% on RefCOCO, RefCOCO+, and G-Ref, respectively. These improvements validate the effectiveness and robustness of the proposed multi-granularity feature extraction and loss optimization strategy, highlighting its potential for advancing referring image segmentation research.

Key words referring image segmentation; cross-model fusion; feature enhancement; attention mechanism; loss function

1 引言

图像指代分割旨在根据文本描述对图像中特定的目标区域进行定位与分割。相较于语义分割与实例分割，它更侧重于分割出描述指定的目标区域，而不是某一类别的全部区域。在实际应用中，图像指代分割能够根据目标外观的文本描述，从图像中对目标进行快速定位与精准分割，节省人力资源，具有广泛的应用前景和重要的研究价值。

传统的图像指代分割方法主要基于联合嵌入方法。Hu 等人[1]通过结合卷积神经网络[2]和长短期记忆网络[3]提取图像与文本特征，并采用全卷积神经网络[4]输出分割结果。Ye[5]等人引入注意力机制进行跨模态特征融合，使模型聚焦图像关键区域及文本重要信息。然而，这类基于联合嵌入的方法在跨模态特征对齐的精细度上仍有不足。因此，部分研究转向探索多尺度特征与文本指代描述的对齐机制。Yu[6]等人将文本描述分解为目标外观、位置以及关系模块，并计算模块的匹配分数。Huang[7]等人逐步利用文本描述中不同类型词汇构建空间图，结合图网络分割指代目标。此外，还有一些研究通过位置先验嵌入与统一注意力协同等方法显著提升小尺寸目标模型的分割效率。Zhu 等人[8]将分割任务统一建模为条件点回归问题，并对目标边界序列化为离散 Token 进行预测。Li 等人[9]将任务重新表述为掩码-图像-文本三模态深度交互问题，摒弃传统双编码-融合-解码范式，实现跨模态特征从低层到高层的对齐与强化。尽管上述方法能够有效提升图像指代分割的性能，但仍存在以下不足：(1)仅关注全局信息或仅聚焦局部纹理，特征间脱耦难以形成互补优势，导致模型遮挡、多尺度目标等场景下边缘感知能力弱化，对复杂场景的感知不足；(2)固定或线性的特征融合权重缺乏动态建模能力，难以适应不同样本的特征分布变化，导致特征间相互抑制；(3)当背景像素占比过大且目标极小或

被遮挡时，模型难以有效学习目标特征，易出现漏检或误检现象。

为解决上述问题，本文提出一种多粒度特征提取与损失优化的图像指代分割方法。首先，构建多粒度特征提取模块，设计局部分支与全局分支，通过两个分支的交互，实现全局上下文特征与局部细节特征的互补表达，使模型能够更好地理解全局背景和局部细节的相互作用；然后，引入动态权重分配机制，动态调整各特征的重要性，避免冗余信息对模型训练的负面影响；最后，设计联合损失函数，通过对目标区域边界像素赋予更高的惩罚系数，对易分样本降权，强化难例梯度，减少背景区域对模型训练的干扰，提高目标区域边缘的连续性、精确性与分割的稳定性、鲁棒性。

综上所述，本文的主要贡献如下：

1）构建一种多粒度特征提取模块。设计全局分支提取全局上下文特征，设计局部分支提取局部细节特征，通过两个分支的交互，实现两种特征的深度融合，提升特征表达多样性。

2）引入一种动态权重分配机制。通过对多粒度特征的权重进行动态调整，减少冗余信息干扰，提升特征质量与鲁棒性。

3）提出一种联合损失函数。通过优化各损失函数的比例，提升真实目标区域反向传播得到的梯度，减少背景区域干扰，增强特征判别性。

2 相关工作

2.1 基于Transformer的图像指代分割方法

图像指代分割任务要求模型依据文本描述精准定位并分割图像中的目标区域，其核心挑战在于实现跨模态语义对齐与细粒度视觉定位。Transformer[10]凭借自注意力机制对长距离依赖的建模能力和跨模态交互的灵活性，逐渐成为解决该任务的主流框架。现有研究主要聚焦于探索视觉与语言特征的交互方式及目标表示形式。VLT[11]通过将图像与语言特征映射至共享多维空间，并利用自注意力机制实现跨模态全局交互。LAVT[12]将语言信息直接融入 Swin Transformer[13]的各个阶段，在构建的局部到全局特征金字塔内实现层级化语义对齐，从而提升目标定位精度与特征融合的细粒度。EFN[14]在视觉主干网络的每一层级引入双向协同机制，实现语言信号对多尺度视觉特征的自上而下调控，并在解码端引入边界增强模块以逐步恢复目标边缘细节。ReSTR[15]进一步摒弃卷积操作，采用纯 Transformer 架构分别提取视觉与语言特征，再通过自注意力进行融合，充分展现了该架构在捕获长期依赖关系及动态跨模态交互方面的优势。在目标表示形态方面，SeqTR[8]创新性地将边界框、关键点乃至像素级掩码统一表示为离散坐标序列，并基于可变形 Transformer 构建了一个跨多种视觉指代任务的统一框架，验证了序列化表示的可行性。PolyFormer[16]则更进一步，将掩码生成建模为自回归多边形顶点序列，利用多边形 Transformer 直接输出亚像素级边界轮廓。该方法避免了上采样误差，尤其适用于小目标及具有精细轮廓目标的精确分割。基于 Transformer 的上述优势，本文采用 Transformer 框架进行视觉特征提取，更好地捕捉长距离依赖增强跨模态语义对齐与细粒度视觉定位。

2.2 多粒度特征

多粒度特征提取在视觉与语言交互任务中受到广泛关注，研究者们围绕不同粒度特征的融合与利用开展了大量研究。Li 等人[17]提出将金字塔特征输入 ConvLSTM[18]，通过粗细粒度循环迭代修正分割边界，首次验证了多尺度递进细化策略在处理小目标与边界问题上的优势。Tan 等人[19]强调词级调制与句级上下文增强的重要性，通过对同一视觉特征执行不同语义粒度的选择性增强操作，构建了视觉语言双通道、词句双层协同的框架，缓解了因粒度失配导致的语义遗漏问题。Huang 等人[7]将语言信息进一步细分为实体、属性和关系三类，先用实体词进行粗定位，再利用关系词进行图推理实现细粒度分割，同时借助 ConvLSTM 对多层特征进行逐级细化，实现语义与空间尺度双层级渐进推理。Qiu 等人[20]在解码阶段引入级联金字塔结构与文本引导门控机制，在跨三个尺度循环对齐的过程中重新显化细节信息，其设计对复杂背景中的小尺寸目标更为友好。Luo 等人[21]创新性地将句子语义分解为不同粒度的节点，并在解码过程中设计多层互选与内外层动态聚合机制，显式保存细粒度粗粒度目标线索，同时引入将目标映射为类别计数的双重监督信号，弥补了现有多粒度框架在目标数量维度建模上的空白。Zhang 等人[22]提出使用可微分高斯核函数衡量像素间一致性，以此抑制冗余响应并强化难分区域的判别性，实现自下而上的层次化特征融合，在保持高分辨率细节信息的同时兼顾全局语义一致性，缓解了粒度失配带来的语义遗漏问题。现有研究虽在多粒度特征融合与语义对齐方面取得

进展，但仍存在对全局背景与局部细节的交互关系捕捉不足、不同粒度信息匹配失配导致的语义表达不完整等问题

因此，本文引入多粒度特征提取方法，以整合不同尺度、不同语义深度信息，增强模型对全局语境与局部细节关联的理解能力。

2.3 图像指代分割损失函数

损失函数的设计在图像指代分割中对模型优化方向具有关键引导作用，近年来研究者们针对边界精度、语义对齐与难例挖掘等问题提出多种创新设计。Rong 等人[23]引入文本加权的边界损失，通过计算预测掩膜与真值掩膜梯度的差异并结合文本特征加权，优化目标物体的轮廓一致性。Xie 等人[24]在损失中增加边界预测分支，同步优化分割掩膜与边界输出，相比仅使用传统交叉熵，联合训练边界损失能够获得更精细的轮廓细节。Dalaq 等人[25]提出指代感知融合损失，综合多种损失提升目标边界和困难区域的分割精度，在保证整体分割准确的同时，帮助模型关注细粒度的边界一致性。Wang 等人[26]引入文本到像素的对比学习损失，在共同嵌入空间中拉近对应语义的文本表示与像素表示距离。Zhang 等人[27]提出像素分组损失确保目标物体像素的聚合一致性，强制同一目标的像素在特征空间靠近，非目标像素远离。Nguyen 等人[28]引入短语对象对齐损失，监督每个预测的对象掩膜与输入语句中特定词组唯一对应，通过该损失强化这种匹配关系。Hai 等人[29]提出语义一致性损失，促使模型生成的视觉感知文本特征在嵌入空间上保持一致。Liu 等人[30]增加无目标类别的分类损失，预测可能的多个目标以及判断是否存在目标。现有损失函数虽在边界优化与语义对齐方面取得进展，但在目标占比较小的场景中，易受背景区域干扰，对难例样本的关注不足，导致分割漏检、误检现象以及目标边缘连续性不佳等问题。因此，本文设计联合损失函数，旨在通过强化难例学习与边界约束，减少背景干扰，提升模型在复杂场景下的分割稳定性与鲁棒性。

3 本文方法

本文方法整体架构如图 1 所示。首先，分别采用 Swin Transformer 和 BERT[31]提取视觉特征 V_i 与文本特征 L，以 V_i 为查询，L 为键值的注意力机制，为每个图像位置动态聚合最相关的文本语义，生成一幅与视觉特征图空间对齐的位置专属语言上下文图 G_i，公式如下所示。

$$G_i = Proj(Softmax(\frac{V_i^T \cdot L}{\sqrt{d}}) \cdot L^T) \tag{1}$$

将 G_i 输入多粒度特征提取模块(Multi-granularity feature extraction module，MGFE)，获得第一阶段的多粒度特征 G_i'；然后，通过动态权重分配机制(Dynamic weight allocation mechanism，DW)调整特征 G_i 与 G_i' 间的权重，得到融合语义特征 F 后，将 F 与 V_i 依次进行拼接获取各阶段的拼接特征 F'；最后，将得到的各阶段拼接特征拼接后得到 F'' 并输入轻量化解码器，通过计算和上采样操作，将预测掩码尺寸调整至与输入图像一致，并经过联合损失函数优化，得到精细化的分割掩码。

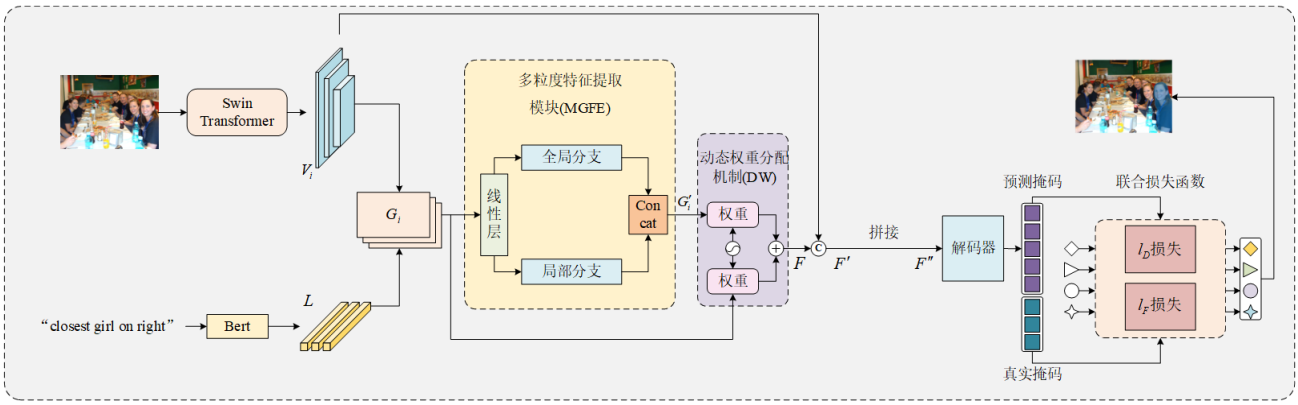

图 1 多粒度特征提取与损失优化的图像指代分割整体框架

3.1 多粒度特征提取模块

针对现有图像指代分割方法全局上下文特征与局部细节特征提取不充分、特征表达能力不足的问题，本文构建多粒度特征提取模块。如图 2 所示，该模块由两个分支构成。其中，局部分支通过深度卷积提取局部特征，结合非线性激活函数增强对细

节与目标边界区域的关注能力；全局分支通过自注意力捕获全局范围内的长距离依赖关系，获取全局特征。通过拼接局部特征与全局特征，获得多粒度特征。

具体地，融合特征 G_i 经过线性变换与全连接层处理后，生成查询 Query、键 Key 和值 Value，分别输入全局分支与局部分支。在全局分支中，通过对 K 和 V 进行平均池化聚合全局上下文，增强模型的全局感受野，接着，对 Q，K 和 V 执行注意力操作，提取全局特征并捕获图像的整体语义信息与长程依赖关系，得到全局分支的输出特征 X_{global}，公式表示如下：

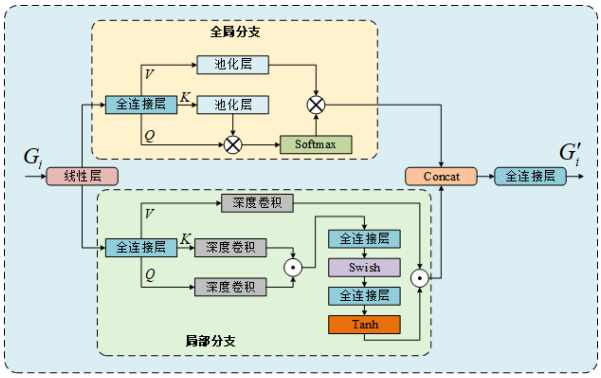

图 2　多粒度特征提取模块结构图

$$Q_x, K_x, V_x = FC(Linear(G_i)) \tag{2}$$

$$Pool(X)_{(i,j)} = \frac{1}{k \times k} \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} X_{(i+m, j+n)} \tag{3}$$

$$A_x = Softmax(\frac{Q_x \cdot \left[Pool\left(K_x\right) \right]^{\mathrm{T}}}{\sqrt{d}}) \tag{4}$$

$$X_{global} = A_x \cdot Pool\left(V_x\right) \tag{5}$$

其中，Q_x，K_x，V_x 分别为融合特征 G_i 经过线性变换与全连接层处理后，生成的查询、键和值。

在局部分支中，采用深度可分离卷积[32]（Depthwise Convolution, DWconv）分别聚合 Q、K 和 V 的局部信息，通过逐通道卷积提取各通道的局部空间特征，并利用逐点卷积融合不同通道的信息，增强模型对局部特征的上下文感知能力。将 Q 和 K 的局部特征进行逐元素相乘，经过两次全连接层、Swish 与 Tanh 激活函数处理后，生成上下文感知权重，并将该权重与 V 的局部特征进行逐元素相乘，得到局部分支的输出特征 X_{local}，公式如下所示。

$$V_z = Depth(V_x; \Theta_d) \odot Point(V_x; \Theta_p) \tag{6}$$

$$Q_i = DWconv(Q_x; \Theta_q) \tag{7}$$

$$K_i = DWconv(K_x; \Theta_k) \tag{8}$$

$$Attn_i = W_2 \cdot (Swish(W_1 \cdot (Q_i \odot K_i))) \tag{9}$$

$$Attn = \mathrm{Tanh}(\frac{Attn_i}{\sqrt{d}}) \tag{10}$$

$$X_{local} = Attn \odot V_z \tag{11}$$

其中，Θ_d，Θ_q，Θ_k 分别表示 V，Q，K 的逐通道卷积的权重参数，Θ_p 表示逐点卷积的权重，d 表示 Tokens 的通道数，最后，将局部分支与全局分支特征进行拼接，经全连接层输出多粒度特征 G_i'，公式表示如下：

$$G_i' = W_o \cdot Concat(X_{local}, X_{global}) \tag{12}$$

多粒度特征提取模块构建一个全局与局部互补的特征空间。其中，全局分支通过自注意力机制实现任意两个像素点间的交互，建模长程依赖，捕获目标与背景、其他目标之间的整体语义布局；局部分支通过深度可分离卷积强化空间细节，增强模型对目标边界和细粒度纹理的感知。通过两个分支的并行与深度融合，实现对图像多尺度信息的无损或低损编码，避免因单一特征提取方式导致的信息损失，为后续的精准分割提供更完备的特征基础。

3.2　动态权重分配

为提升特征鲁棒性，本文引入动态权重分配机制，通过调节特征 G_i 与 G_i' 间的权重，减少冗余信息干扰，如图 3 所示。首先，G_i' 通过 1×1 卷积操作与 *GELU* 激活函数，提取通道间的关键信息。然后，通过 1×1 卷积操作与 *Tanh* 激活函数进一步提取特征，生成对应权重 S_i，捕捉关键特征信息。对 G_i 执行相同操作，得到其动态权重 Y_i。最后，将权重 S_i 与 G_i' 逐元素相乘、权重 Y_i 与 G_i 逐元素相乘，再将两个结果逐元素相加，得到融合语义特征 F，公式表示如下：

$$S_i = \varphi_i(G_i') \tag{13}$$

$$Y_i = \gamma_i\left(G_i\right) \tag{14}$$

$$F = S_i \odot G_i' + Y_i \odot G_i \tag{15}$$

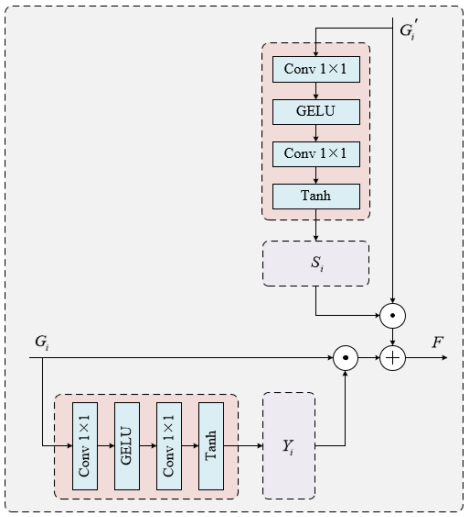

图 3 动态权重分配示意图

作为一种自适应特征选择策略，动态权重分配可被视为一个轻量级的门控网络。根据当前输入的特征图，通过注意力机制动态地生成一组空间-通道注意力权重，实现对特征通道的软选择。同时，通过降低非关键特征的权重，减少特征融合过程中的噪声与冗余信息，增强关键特征的表达，使模型能够根据具体样本的复杂度对注意力进行动态调整，提升模型的泛化能力和鲁棒性。

3.3 联合损失函数

合理的损失函数能够增强模型对目标区域的学习能力，提升分割精度。因此，本文设计联合损失函数，通过优化预测目标区域边界与真实目标区域边界的匹配，增强特征判别性。

联合损失通过动态调整易分样本和难分样本的权重，驱使模型关注存在歧义或复杂背景的目标区域，联合损失通过优化各损失函数的比例，提升真实目标区域反向传播得到的梯度，增强特征判别性，公式如下所示。

$$l_{Combine} = l_D \cdot \alpha_D + l_F \cdot \beta_F \tag{16}$$

其中，l_D 为 Dice 损失函数，l_F 为 Focal 损失函数。α_D 和 β_F 是调整两个损失函数比例的超参数。

首先，联合损失函数将模型预测输出 y_{pred} 按类别进行 *Softmax* 操作，得到每个像素属于目标区域的概率 y_{sp}，将真实区域 y_{true} 转换为 one-hot 向量，最后计算损失，公式如下所示。

$$y_{sp} = softmax(y_{pred}, dim = 1) \tag{17}$$

$$y_o = one\_hot(y_{true}, C) \tag{18}$$

$$l_D = 1 - \frac{2 \cdot \sum y_{sp} \cdot y_o}{\sum y_{sp} + \sum y_o + \grave{o}} \tag{19}$$

其中，$\sum y_{sp} \cdot y_o$ 表示预测概率与真实标签的逐像素乘积累加，等同于预测与真实的重叠区域，$\sum y_{sp} + \sum y_o$ 表示预测区域与真实区域的并集，ò 为防止分母计算中得到零值的常数。

计算每个像素预测正确的概率 p，p 越小，当前像素的预测难度越大，动态权重 *weight* 越大；计算交叉熵损失 l_{CE} 并与动态权重相乘,最后计算损失，公式如下所示。

$$p = y_{sp} \cdot y_o + (1 - y_{sp}) \cdot (1 - y_o) \tag{20}$$

$$weight = \alpha \cdot (1 - p)^{\gamma} \tag{21}$$

$$l_{CE} = y_o \cdot ln(p + \text{蝌}) + (1 - y_o) \cdot ln(1 - p + \ ) \tag{22}$$

$$l_F = -weight \cdot l_{CE} \tag{23}$$

其中，α 为平衡因子，γ 为焦点调节因子，ò 为用于防止对数计算中出现零值的常数。

联合损失函数遵循难例挖掘与梯度再平衡的优化思想。在目标区域占比极小的场景中，普通交叉熵损失容易因背景主导导致梯度消失，Focal 损失在标准交叉熵损失上增加一个调制因子，动态地降低易分样本对总损失的影响，使模型聚焦于难区分的边界区域；Dice 损失则通过集合相似度优化，促使模型关注目标区域的整体，增强目标区域的连续性与完整性。两者通过联合使用，确保在梯度反向传播过程中能够更有效地作用于真实目标区域，进而在训练过程中实现细节与整体的均衡优化，增强模型的特征判别性。

4 实验结果与分析

4.1 数据集

本文在 RefCOCO[33]、RefCOCO+[33]和 G-Ref[34]三个图像指代分割任务的公开标准数据集上对方法进行训练和评估。三个数据集的图像均来自 MSCOCO[35]数据集，并使用文本描述进行注释。RefCOCO 数据集包含 19994 张图像，50000 个注释对象和 142209 个注释表达式，描述中主要为绝对

方位词，模型需要理解并利用这些空间位置信息，其中的 TestA 测试集专门用于人物类目标，TestB 测试集专门用于物体类目标，通常包含小物体和复杂背景；RefCOCO+数据集与 RefCOCO 相似，但是禁用绝对方位词，目的是增强对外观属性的理解，使模型必须依赖于物体的属性而不是空间关系来定位目标，包含 19992 张图像，49856 个注释对象和 141564 个注释表达式，TestA、TestB 测试集目标与 RefCOCO 类似；G-Ref 数据集是一个描述更长、更复杂的数据集，有两个不同的分区，一个由 UMD 划分，另一个由 Google 划分，相较于 RefCOCO 和 RefCOCO+文本描述平均 3.5 个单词，它的文本描述平均包含 8.4 个单词，更具备上下文信息，句子长度较长，包含更多的空间关系和多个物体之间的描述。

本文采用整体交并比、平均交并比和精度 P@X 等常用指标，整体交并比是计算所有像素的整体分割精度，将全部类别视为一个整体，对所有测试图片求预测结果与真实值的交集区域以及并集区域，全称 overall IoU，在论文中简写为 OIoU；平均交并比是对测试集中每一个样本分别计算预测掩码与真实掩码之间的 IoU，然后对所有样本的 IoU 取平均，全称 mean IoU，在论文中简写为 MIoU；P@X 计算 IoU 高于阈值 X 的测试样本占比，重点衡量模型的定位能力。

4.2 实验设置

本实验模型在 Python 3.10.0，Pytorch 2.0.0 环境下进行，在 Ubuntu 22.04 操作系统上运行代码，在 NVIDIA-4090GPU 上进行训练和测试。视觉特征提取网络使用的是在 Imagenet1K[36] 上预训练的 Swin Transformer，文本特征提取网络选用的是 BERT，Swin Transformer 选用 Base 规格进行对比实验，未进行微调。训练过程中，使用 AdamW 优化器。其中，自适应学习率数值稳定性常数设置为 1×10^{-8}，控制梯度均值的指数衰减率设置为 0.9，控制梯度平方均值的指数衰减率设置为 0.999。初始学习率设置为 5×10^{-5}，权重衰减系数设置为 0.01，batch size 设置为 8，共进行 40 个 epoch。输入图像经过裁剪等预处理操作，调整图像大小为 480×480，经归一化处理后输入到网络模型中，无其他数据增强策略。

在图像编码器的选取方面，Swin Transformer 是一个通用视觉骨干网络，采用有监督训练的训练方式，通过分层结构和移位窗口自注意力机制，构建多尺度特征图；DINO[37]是一个自监督视觉模型，采用自监督训练的训练方式，使用知识蒸馏的方法，使模型从图像中自主学习鲁棒的视觉特征。在文本编码器的选取方面，BERT 是一个纯文本的自然语言处理模型，采用双向 Transformer 编码器，通过掩码语言模型和下一句预测任务进行预训练。CLIP[38]是一个专门设计用于理解图像和文本之间关系的多模态模型，采用双编码器结构，使用对比学习在共享的嵌入空间中对齐图像和文本表示。

本文侧重于具体的图像指代分割任务，需要一个强大的特征提取骨干网络；同时，需要深度理解语言语义、上下文和句间关系。基于上述考量，本文分别选取 Swin Transformer 与 BERT 作为图像编码器与文本编码器，对视觉特征与文本特征进行提取。

表 1 各数据集上不同方法的 OIoU 实验对比结果（%）

对比模型	视觉主干网络	RefCOCO			RefCOCO+			G-Ref		
		Val	TestA	TestB	Val	TestA	TestB	Val(U)	Test(U)	Val(G)
EFN[14]	ResNet-101	62.76	65.69	59.67	51.50	55.24	43.01	-	-	51.93
BUSNet[40]	ResNet-101	63.27	66.41	61.39	51.76	56.87	44.13	-	-	50.56
CGAN[41]	DarkNet-53	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69	46.54
VLT[11]	DarkNet-53	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65	52.02
LTS[42]	DarkNet-53	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25	-
SeqTR[8]	DarkNet-53	67.26	69.79	64.12	54.14	58.93	48.19	-	55.67	55.64
ReSTR[15]	ViT-B-16	67.22	69.30	64.45	55.78	60.44	48.27	54.48	-	-
MaIL[9]	ViT-B-16	69.38	71.31	66.76	62.23	65.92	52.06	62.45	62.87	61.81

PCAN[43]	ResNet-50	69.51	71.64	64.18	58.25	63.68	48.89	59.98	60.80	57.49
CRIS[26]	ResNet-101	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36	-
ETRIS[44]	ViT-B-16	70.51	73.51	66.63	60.10	66.89	50.17	59.82	59.91	57.88
CNNFormer[45]	ResNet-50+ SwinT-B	71.93	76.13	67.74	61.78	67.78	51.77	60.84	62.20	57.38
BarLeRIa[46]	CLIP-B	72.40	75.90	68.30	64.00	69.80	55.50	62.40	63.80	61.60
LAVT[12]	SwinT-B	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	60.50
MagNet[47]	SwinT-B	74.24	77.24	70.15	65.16	70.32	57.14	64.36	65.03	62.13
RISCLIP[48]	CLIP-B	73.57	76.46	69.76	64.53	69.61	55.09	63.10	64.09	-
本文	SwinT-B	73.96	77.09	70.26	63.31	70.12	54.96	63.33	64.28	62.10

4.3 对比实验

为证明所提方法的有效性，本文在 RefCOCO、RefCOCO＋和 G-Ref 三个数据集上进行实验并与现有方法进行性能比较，结果见表 1。实验表明，所提方法在三个公开的数据集上的分割精度优于现有主流方法，且相较于基线模型，本文方法的 OIoU 指标有显著提升，表明模型能更精准利用位置信息定位目标，在 TestA 上达到 77.09%，表明本文方法在人物目标分割任务中具有更强的跨模态对齐能力，尤其在复杂姿态与遮挡场景下表现更优，在 TestB 上达到 70.26%，验证其在物体细粒度语义理解上的优势，得益于局部全局建模策略。

本文方法在 RefCOCO+上进一步证明其在处理文本描述中隐含属性时的鲁棒性。RefCOCO+的物体描述更依赖抽象空间关系，物体类别的属性组合更复杂，因此模型在长距离依赖建模上更优。但在 TestB 上的效果略低，可能是由于 TestB 专门评估物体类实例，这些目标往往尺度更小、遮挡更多，IoU 对边界抖动敏感，同时 RefCOCO+描述更多为物品属性，缺少空间词，模型需依赖精细视觉属性而非位置线索。

本文方法在 G-Ref 数据集上均优于基线模型，在 Val(U)上达到 63.33%，表明模型对长文本的语义解析能力更强。

虽然近年的 SOTA 方法 MagNet 在对比结果上优于本文方法，但相较于其他模型，首先该模型在训练数据上新增真实分割掩码，数据依赖性高且如果标注有噪声或不够精确，会损害模型性能。其次该方法新增一种掩码编码器，调参成本高昂且模型复杂度高。本文方法以较低的数据依赖与简单高效的推理部署，能够在保持较高性能的同时，极大地提升模型的实用性和易用性。在同样的实验环境下进行，模型的参数量及计算复杂度对比如表 2 所示。

表 2 模型复杂度对比结果

模型	Params(M)	GFLOPs
本文方法	237.65	199.16
MagNet	286.38	242.14

4.4 消融实验

4.4.1 模块间消融实验

为验证本文所提出的多粒度特征提取模块（MGFE）、动态权重分配机制（DW）及联合损失函数 $l_{Combine}$ 的有效性，本文在 RefCOCO、RefCOCO+和 G-Ref 验证集上进行消融实验，实验结果详见表 3、表 4、表 5。

首先单独评估模块对模型性能的影响，启用 MGFE 时，模型在 RefCOCO 验证集的 MIoU 达到 75.30%；RefCOCO+验证集的 OIoU 达到 62.74%。启用联合损失函数时，模型在 RefCOCO 验证集的 OIoU 达到 72.89%；RefCOCO+验证集的 MIoU 达到 66.40%。结果表明，多粒度特征提取模块局部分支对目标边缘和细粒度纹理进行增强，全局分支捕捉长程依赖，补偿基线模型仅依赖单一特征尺度的问题，显著增强对目标的整体表征能力。对于预测概率较低的边界难分像素，联合损失函数的相对权重被放大，在梯度反向传播过程中，模型参数更新的方向被重新校准，使其优先优化这些难以区分的边界区域，并且在边界上的任何预测偏差都会直接引起损失值的显著升高。

同时启用多粒度特征提取模块和动态权重分配机制时，模型性能进一步提升，在 RefCOCO 上，MIoU 达到 75.88%；在 RefCOCO+和 G-Ref 上，MioU 分别达到 67.40%与 65.86%。结果表明，动态权重分配机制能将多粒度特征提取模块里正相关的特征通道进一步放大，通过动态调整融合特征的权重分配，增强模型对复杂场景的适应能力。

当启用所有模块时，模型在各数据集上的指标均取得最佳结果。在 RefCOCO 上，MIoU 提升至

76.17%；在 RefCOCO+和 G-Ref 上也呈现类似的提升趋势。结果表明，联合损失函数将难例权重提升，与多粒度特征提取模块的细粒度边缘建模方向一致，损失收敛更快，使多粒度特征提取模块产生的细节在反向传播中得到强化。

如图 4 所示，三者的结合能够从全局与局部、特征与损失等多个层面实现全方位优化，显著提升模型对复杂场景中目标的整体理解和细节分割能力，进一步验证模块之间互补的耦合效应。

图 5 展示在基线模型的基础上消融每一个模块的可视化效果，多粒度特征提取模块显著细化目标边界并抑制背景噪声，分割区域的外轮廓更加贴合物体真实形状；动态权重分配机制较准分割掩码的整体位置，提升整体定位精度；联合损失函数进一步优化目标，三者在基线之上形成互补增益，对提升分割性能至关重要。

表 3　RefCOCO 验证集消融实验结果（%）

MGFE	DW	$l_{Combine}$	P@0.5	P@0.7	P@0.9	MIoU	OIoU	Params(M)	GFLOPs
			84.46	75.28	34.30	74.46	72.73	227.74	192.30
✓			85.38	76.40	34.78	75.30	73.12	236.26	198.22
		✓	84.95	75.65	34.55	74.97	72.89	227.74	192.30
✓	✓		86.24	77.05	35.11	75.88	73.59	237.65	199.16
✓	✓	✓	**86.83**	**77.37**	**35.43**	**76.17**	**73.96**	237.65	199.16

表 4　RefCOCO+验证集消融实验结果（%）

MGFE	DW	$l_{Combine}$	P@0.5	P@0.7	P@0.9	MIoU	OIoU	Params(M)	GFLOPs
			74.44	65.58	30.23	65.81	62.14	227.74	192.30
✓			74.95	66.06	30.30	67.03	62.74	236.26	198.22
		✓	74.86	65.80	30.26	66.40	62.33	227.74	192.30
✓	✓		75.98	66.96	30.88	67.40	62.95	237.65	199.16
✓	✓	✓	**76.97**	**68.32**	**31.02**	**67.72**	**63.31**	237.65	199.16

表 5　G-Ref 验证集消融实验结果（%）

MGFE	DW	$l_{Combine}$	P@0.5	P@0.7	P@0.9	MIoU	OIoU	Params(M)	GFLOPs
			70.81	58.60	22.73	63.34	61.24	227.74	192.30
✓			71.37	59.07	23.61	64.32	62.30	236.26	198.22
		✓	71.13	58.85	22.91	63.82	61.48	227.74	192.30
✓	✓		72.39	60.27	24.23	65.86	63.05	237.65	199.16
✓	✓	✓	**73.19**	**61.05**	**24.69**	**66.23**	**63.33**	237.65	199.16

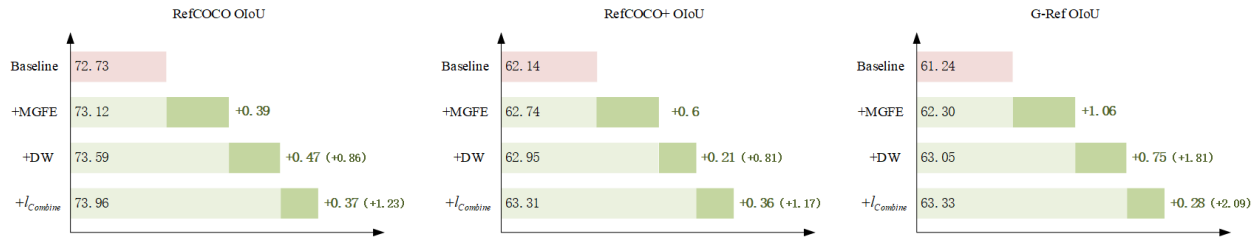

图 4 消融模块兼容性对比图

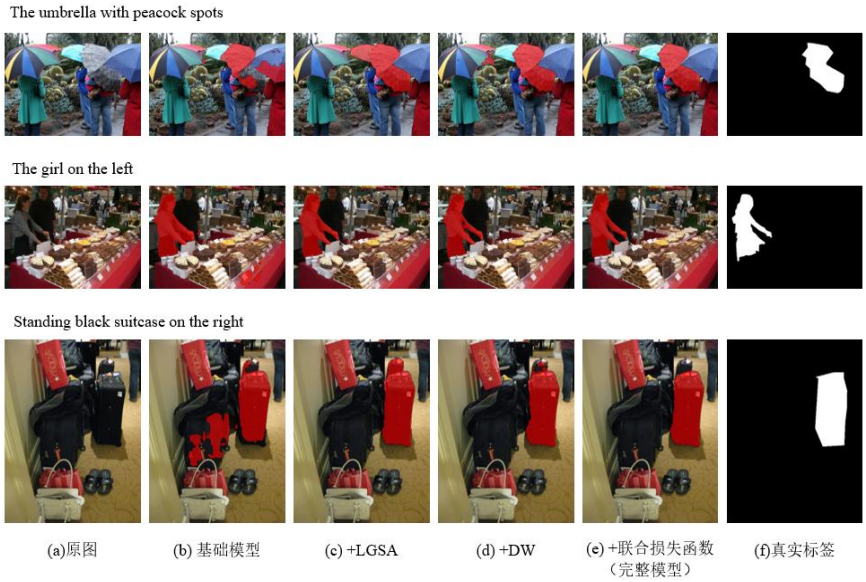

图 5 消融模块对比可视化

4.4.2 跨模态融合模块消融实验

为探究跨模态融合模块对最终结果的影响，本文在 RefCOCO 验证集上进行消融实验，结果如表 6 所示。在骨干网络中保持相同模型以及训练策略的前提下，将本文提出的多粒度特征提取模块分别替换为同类型注意力模块进行对照。EFN 基于早期特征拼接策略，其性能瓶颈可能源于模态间浅层交互导致的语义信息不齐。BRINet 通过双向递归迭代优化特征，但其指标较低，反映递归结构在长程依赖建模上的局限性。GACD 是一种基于收集和扩散的双重注意力机制，该模块聚合关键视觉信息，将收集的特征扩散到空间各位置，增强局部与全局的表示，但在精度较高指标上低于本章方法，推测其注意力权重的动态分配在极端阈值下对小尺寸目标和边界敏感区域的适应性不足。

4.4.3 权重分配消融实验

为探究特征之间权重分配对结果的提升，本文在 RefCOCO 验证集上进行消融实验，结果如表 7 所示。实验 1 是固定权重，采用恒定加权策略无法根据输入特征动态调整跨模态交互强度，导致复杂场景下的特征响应不足，而且缺乏对空间位置敏感性的建模，限制局部与全局特征的协同优化，不足以对输入特征进行有效区分或动态调整。实验 2 是 1×1 的卷积自注意力，通过单卷积核生成注意力图，相较固定权重有一定程度的动态调整能力，但卷积仅建模像素级关联，缺乏跨区域的长程依赖捕捉能力，最终导致关键位置的关注不足、精细化分割能力不够，从而表现仍低于本章方法。实验 3 是 SE 模块，通过学习来获取每个通道的重要性，提升有用特征并抑制无用特征，但只能对单个特征图进行内部校准，无法处理两个不同来源的特征。而动态权重分配协调两个特征源，通过生成两套权重来评估并融合它们，实现“1+1>2”的效果，并且通过卷积生成空间权重图，保留了完整的空间维度。

4.4.4 损失函数消融实验

为探究不同损失函数对结果产生的优化效果，本文在 RefCOCO 验证集上进行消融实验，结果如表 8 所示。首先交叉熵损失作为最常用的多分类损失函数，能够提供较为稳定的基线性能，在 MIoU 等指标上具有较为平稳的表现，然而，由于其对类

间不平衡与目标边界的关注度相对不足，在高阈值下的精细化定位和整体交并比方面仍存在提升空间。多类 Dice 损失函数对类别不平衡更具鲁棒性，在各项指标上均有一定幅度的提升，在目标区域的整体覆盖和融合度上更为突出，但它并不针对容易混淆的难分类像素进行特殊关注，可能导致边缘细节或小尺寸目标部分的漏检。边界 Dice 损失函数在 Dice 损失函数基础上进一步融入边界信息，意图兼顾目标整体与边缘的定位，与本章方法相比，其对边界的优化并未带来显著的全面增益，说明针对细节边界的强化策略在全局与局部平衡上仍有不足。

4.4.5 联合损失函数比例因子消融实验

为评估联合损失函数之间的比例因子对模型分割性能的影响，本文在 RefCOCO 验证集上进行消融实验，结果如表 9 所示。当比例因子为 0.33 时，模型在各项指标上均达到最佳性能，适当增加 l_F 的比例有助于提高模型对困难样本和细粒度目标的关注能力。相较之下比例因子为 3.0 时，其各项指标均低于 0.33 的配置，而比例因子进一步增大时，模型性能则明显下降，过度强调 l_D 可能导致模型对全局区域的覆盖能力虽然有所保障，但在处理类别不平衡以及难以区分的小尺寸目标时，模型的鲁棒性和精细分割能力不足。适当提高 l_F 的比重能够更好地引导模型聚焦于那些易于被背景干扰的难分类样本，从而在保持全局分割准确度的同时，显著提升边缘细节的刻画和定位精度。

表 6 跨模态融合模块消融实验结果（%）

融合模块	P@0.5	P@0.7	P@0.9	MIoU	OIoU
EFN[14]	83.50	73.92	33.82	73.55	71.84
BRINet[39]	82.26	72.81	33.31	72.42	70.19
GACD	83.22	74.09	32.71	73.16	71.20
LAVT[12]	84.46	75.28	34.30	74.46	72.73
本章方法	**86.83**	**77.37**	**35.43**	**76.17**	**73.96**

表 7 权重分配消融实验结果（%）

权重	P@0.5	P@0.7	P@0.9	MIoU	OIoU
实验 1	85.53	75.96	34.28	74.57	72.89
实验 2	84.60	75.31	34.14	74.34	72.49
实验 3	85.66	76.09	34.55	75.12	73.23
本章方法	**86.83**	**77.37**	**35.43**	**76.17**	**73.96**

表 8 损失函数消融实验结果（%）

损失函数	P@0.5	P@0.7	P@0.9	MIoU	OIoU
l_{CE}	84.46	75.28	34.30	74.46	72.73
$l_{MultiClass\_Dice}$	85.98	76.73	34.93	75.44	73.07
$l_{Dice\_Boundary}$	84.52	75.18	34.18	74.25	72.17
本章方法	**86.83**	**77.37**	**35.43**	**76.17**	**73.96**

表 9 损失函数比例因子消融实验结果（%）

比例因子	P@0.5	P@0.7	P@0.9	MIoU	OIoU
0.25	84.60	75.31	34.14	74.34	72.37
0.33	**86.83**	**77.37**	**35.43**	**76.17**	**73.96**
1.0	84.96	75.77	34.13	74.65	72.45
3.0	85.31	76.31	34.75	75.11	73.09
4.0	83.56	74.63	34.05	73.90	71.76

4.5 分割效果可视化

为验证本文方法的性能，对验证集中部分样本的分割效果进行可视化，图 6 所示，同类目标定位，模型准确区分多个候选实例，掩码位置与方位描述一致，证明其对空间指示词的解析和全局几何理解能力；遮挡细粒度人物分割，场景中目标被人群部分遮挡，预测掩码完整勾勒主体轮廓且边界平滑，表明网络可在复杂交互下提取细粒度人体形态并抑制背景误激活；多尺度非人物实例，模型在小物体、非中心区域及斜向拍摄场景下仍能精确分割并保持干净背景。

图 7 所示，第一行当文本描述简洁明且包含明确的方位信息时，即使图像中存在相似目标的干扰，模型仍能够精确地分割出指代对象，在特征图早期阶段，模型对目标区域已产生一定关注，但仍包含部分背景噪声，进入更高层特征后，注意力集中在目标区域，背景干扰明显减弱。第二行在图像中各对象间差异较小、指代目标与背景之间辨识难度较高的情形下，模型依然能够保持较高的分割精度，低层特征图对整个人物轮廓响应较为分散，随着层次加深模型逐渐聚焦在蓝衣女子身上。第三行中间特征图已经能够识别到摩托车的轮廓，但对部分背景也出现识别，通过更深层特征和文本条件的进一步约束，最终分割将目标区域集中在右侧摩托车上。第四行的场景存在部分遮挡，难度更高，在早期特征图中，蓝色车辆和遮挡人物有重叠激活区域，说明模型对复杂遮挡尚存在一定的误解，最终分割结果大致能够区分被遮挡的车辆，但在车辆与人物交界处仍有少量错误划分，与真实标注相比，说明对遮挡区域的精确划分仍具备提升空间。

图 8 所示，第一例中模型未能准确区分“左后方红色公交车”，而是将前景的红色公交车靠前的部分也包含在分割区域中，说明当两个目标外观相似、颜色相近且部分重叠时，前景的红色公交车遮挡了部分后方目标，使得后车的轮廓特征在视觉上不明显，造成分割边界模糊，无法识别深度层次和位置关系。第二例中模型分割出的笔记本区域不准确，有明显的形状误差，说明图像在弱光、反光条件下，模型将肤色与屏幕反光误认为同一类导致边界检测不清晰。这表明模型边界感知仍存在不足，对低对比度图像的鲁棒性有待提升。

5 结论

本文提出一种多粒度特征提取与损失优化的图像指代分割方法。首先，构建一种多粒度特征提取模块，通过全局分支与局部分支的交互，实现全局上下文特征与局部细节特征的深度融合，提升特征表达多样性；引入一种动态权重分配机制，通过对多粒度特征的权重进行动态调整，减少冗余信息干扰，提升特征鲁棒性；最后，设计一种联合损失函数，通过优化各损失函数的比例，提升真实目标区域反向传播得到的梯度，增强特征判别性。在三个公开数据集上的大量实验与定量定性分析表明，本文方法能够根据文本描述对图像中特定目标区域实现较为精准的定位与分割，具有一定有效性。

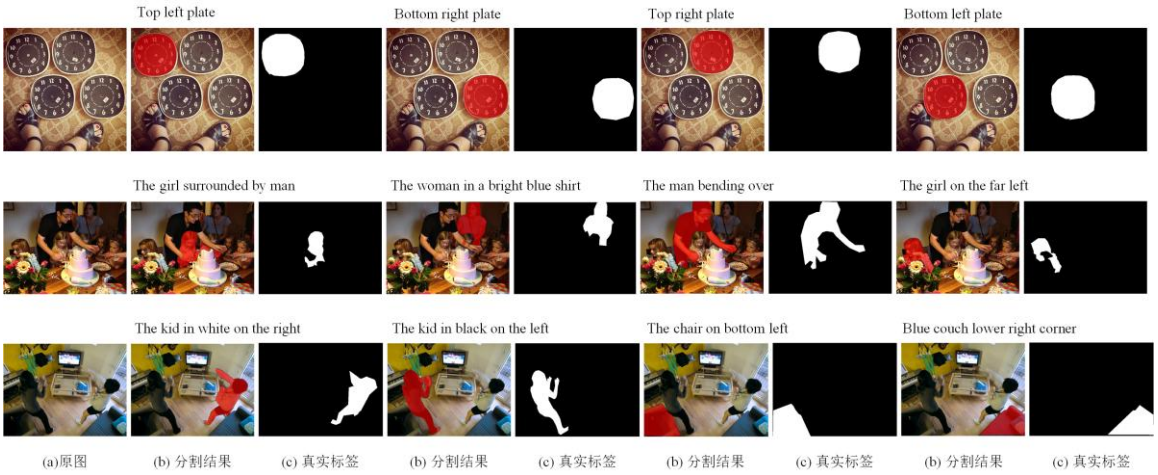

图6 分割结果可视化

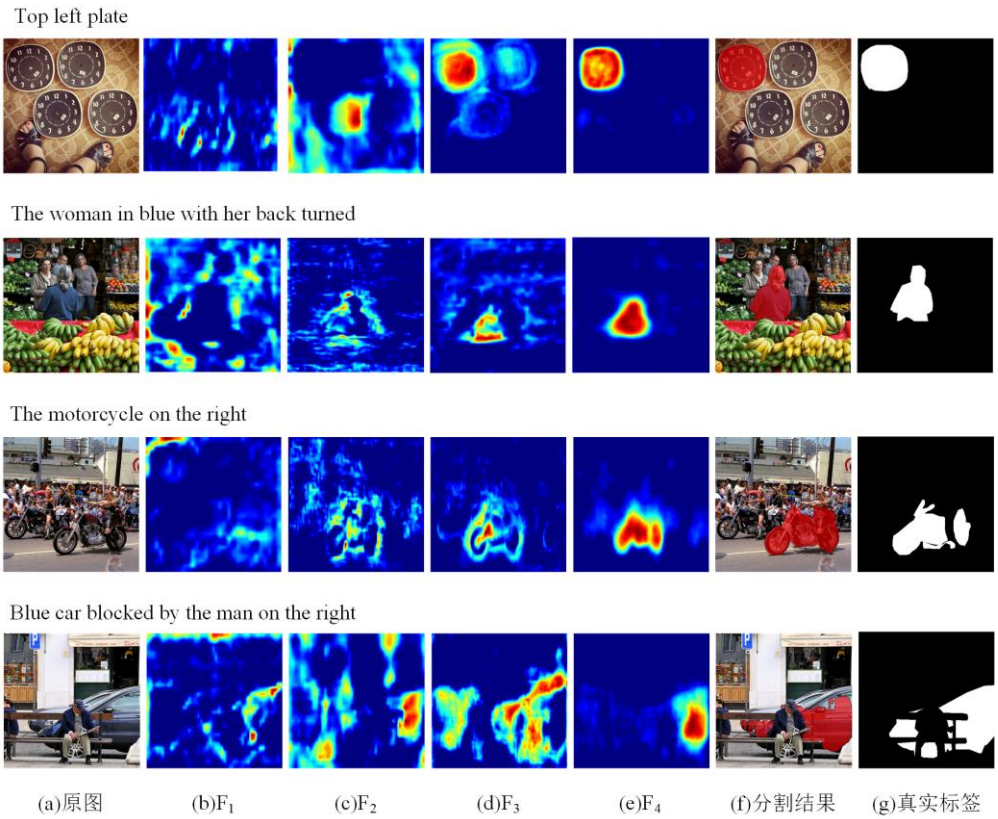

图7 分割阶段可视化

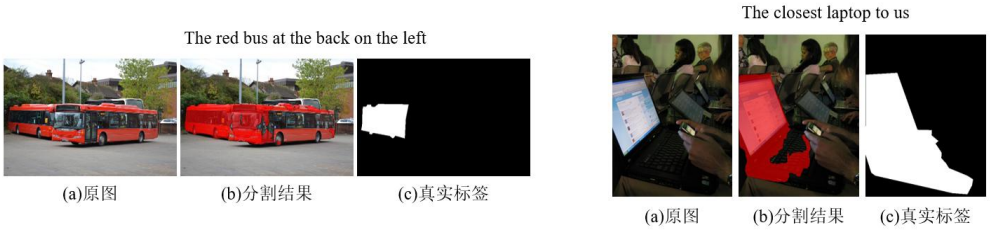

图8　分割失败结果

参 考 文 献

[1] Hu R, Rohrbach M, Darrell T. Segmentation from natural language expressions// Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 108-124.

[2] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks//Advances in Neural Information Processing Systems. Lake Tahoe, Nevada, USA, 2012: 1097-1105.

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

[4] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3431-3440.

[5] Ye L, Rochan M, Liu Z, et al. Cross-modal self-attention network for referring image segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 10502-10511.

[6] Yu L, Lin Z, Shen X, et al. Mattnet: Modular attention network for referring expression comprehension//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1307-1315.

[7] Huang S, Hui T, Liu S, et al. Referring image segmentation via cross-modal progressive comprehension//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 10488-10497.

[8] Zhu C, Zhou Y, Shen Y, et al. Seqtr: A simple yet universal network for visual grounding// Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 598-615.

[9] Li Z, Wang M, Mei J, et al. Mail: A unified mask-image-language trimodal network for referring image segmentation. arXiv preprint arXiv:2111.10747, 2021.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 5998-6008.

[11] Ding H, Liu C, Wang S, et al. Vision-language transformer and query generation for referring segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 16321-16330.

[12] Yang Z, Wang J, Tang Y, et al. Lavt: Language-aware vision transformer for referring image segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 18155-18165.

[13] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 10012-10022.

[14] Feng G, Hu Z, Zhang L, et al. Encoder fusion network with co-attention embedding for referring image segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 15506-15515.

[15] Kim N, Kim D, Lan C, et al. Restr: Convolution-free referring image segmentation using transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 18145-18154.

[16] Liu J, Ding H, Cai Z, et al. Polyformer: Referring image segmentation as sequential polygon generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 18653-18663.

[17] Li R, Li K, Kuo Y C, et al. Referring image segmentation via recurrent refinement networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 5745-5753.

[18] Shi X, Chen Z, Wang H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting//Advances in Neural Information Processing Systems. Montreal, Canada, 2015, 28: 802-810.

[19] Tan Z, Hui T, Chen J, et al. Multi-granularity multimodal feature interaction for referring image segmentation//Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision. Nanjing, China, 2020: 27-39.

[20] Qiu S, Zhang S, Ruan T. Text-Guided Refinement for Referring Image Segmentation. Applied Sciences, 2025, 15(9): 5047.

[21] Luo Z, Wu Y, Cheng T, et al. CoHD: A Counting-Aware Hierarchical Decoding Framework for Generalized Referring Expression Segmentation. arXiv preprint arXiv:2405.15658, 2024.

[22] Zhang W, Cheng Z, Chen J, et al. Hierarchical collaboration for referring image segmentation. Neurocomputing, 2025, 613: 128632.

[23] Rong F, Lan M, Zhang Q, et al. Customized SAM 2 for Referring Remote Sensing Image Segmentation. arXiv preprint

arXiv:2503.07266, 2025.

[24] Xie J, Liu J, Wang G, et al. SATR: Semantics-Aware Triadic Refinement network for referring image segmentation. Knowledge-Based Systems, 2024, 284: 111243.

[25] Dalaq A, Behzad M. Deformable Attentive Visual Enhancement for Referring Segmentation Using Vision-Language Model. arXiv preprint arXiv:2505.19242, 2025.

[26] Wang Z, Lu Y, Li Q, et al. Cris: Clip-driven referring image segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 11686-11695.

[27] Zhang Z, Zhu Y, Liu J, et al. Coupalign: Coupling word-pixel with sentence-mask alignments for referring image segmentation//Advances in Neural Information Processing Systems. New Orleans, USA, 2022, 35: 14729-14742.

[28] Nguyen E R, Le H, Samaras D, et al. Instance-aware generalized referring expression segmentation. arXiv preprint arXiv:2411.15087, 2024.

[29] Nguyen-Truong H, Nguyen E R, Vu T A, et al. Vision-aware text features in referring image segmentation: From object understanding to context understanding//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2025: 4988-4998.

[30] Liu C, Ding H, Jiang X. Gres: Generalized referring expression segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 23592-23601.

[31] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA, 2019: 4171-4186.

[32] Chollet F. Xception: Deep learning with depthwise separable convolutions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1251-1258.

[33] Yu L, Poirson P, Yang S, et al. Modeling context in referring expressions//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 69-85.

[34] Mao J, Huang J, Toshev A, et al. Generation and comprehension of unambiguous object descriptions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 11-20.

[35] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 740-755.

[36] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 248-255.

[37] Zhang H, Li F, Liu S, et al. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605, 2022.

[38] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision// Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2021: 8748-8763.

[39] Hu Z, Feng G, Sun J, et al. Bi-directional relationship inferring network for referring image segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 4424-4433.

[40] Yang S, Xia M, Li G, et al. Bottom-up shift and reasoning for referring image segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 11266-11275.

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

[42] Jing Y, Kong T, Wang W, et al. Locate then segment: A strong pipeline for referring image segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 9858-9867.

[43] Chen B, Hu Z, Ji Z, et al. Position-aware contrastive alignment for referring image segmentation. arXiv preprint arXiv:2212.13419, 2022.

[44] Xu Z, Chen Z, Zhang Y, et al. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 17503-17512.

[45] Yao K, Feng G, Gao X, et al. CNNFormer: A CNN-Transformer Hybrid Model for Referring Image Segmentation// Proceedings of the International Conference on Intelligent Computing. Singapore, 2025: 357-368.

[46] Wang Y, Li J, Zhang X, Shi B, Li C, Dai W, Xiong H, Tian Q. Barleria: An efficient tuning framework for referring image segmentation//Proceedings of the International Conference on Learning Representations. Kigali, Rwanda, 2024.

[47] Chng Y X, Zheng H, Han Y, et al. Mask grounding for referring image segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2024: 26573-26583.

[48] Kim S, Kang M, Kim D, Park J, Kwak S. Extending CLIP's image-text alignment to referring image segmentation//Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL). Seattle, USA, 2024: 4611-4628.

ZHANG Long-Fei, M.S. His main research interests include computer vision and referring segmentation.

MENG Si-Yu, M.S. Her main research interests include computer vision and deepfake detection.

NIU Chun-Xiang, M.S. His main research interests include computer vision and abnormal behavior recognition.

BI Yi-Han, Ph.D. His main research interests include computer vision and Person re-identification.

WANG Rong, Ph.D., professor, Her main research interests include computer vision, artificial intelligence and pattern recognition.

LI Cheng, M.S., His research interests include computer vision, computer language, machine learning.

SONG Jie,Bachelor. His research interests include computer vision, computer language, machine learning.